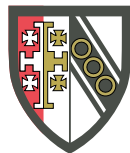




UNIVERSITY OF  
CAMBRIDGE

# Morphological competence in neural natural language processing

Paula Czarnowska



Selwyn College

This dissertation is submitted in July 2022 for the degree of Doctor of Philosophy



## DECLARATION

---

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration. I further state that no substantial part of my thesis has already been submitted or is being concurrently submitted for any degree or other qualification at the University of Cambridge or other similar institution. This dissertation does not exceed the prescribed limit of 60 000 words.

Paula Czarnowska  
20 July 2022



# ABSTRACT

---

## Morphological competence in neural natural language processing

*Paula Czarnowska*

In case-marking languages (CMLs), such as Polish or Finnish, a substantial portion of grammatical information is expressed at the word-level. The word-forms provide information about their inherent properties, like tense or mood, but also encode information about relations between the words. This is in contrast to morphologically impoverished languages, like English, which undergo little inflection. The linguistic factors associated with CMLs make them a challenge for data-driven, neural natural language processing (NLP). To successfully process a CML, neural NLP models must be *morphologically competent*; i.e., they have to capture both the meaning and function of different components of a word form and recognise the importance of morphological signals within a language.

Despite the importance of morphological competence for language processing, the neural NLP models have never been directly tested for that linguistic ability. This gap in the literature is the more important given that most neural NLP models are developed with English language in mind and later applied, without any adaptations, to other languages. It remains unclear whether the architectures and optimization techniques developed on English are able to extract all the essential information from the word-forms of CMLs and whether they can interpret this information at the clausal-level to solve NLP tasks.

In this thesis I investigate whether state-of-the-art neural models for CMLs utilise morphosyntactic information when solving a task for which this information is key: dependency parsing. To answer this question I propose a new evaluation paradigm which involves evaluating the models on various counterfactual versions of dependency corpora. Through evaluation of Polish, Russian, Finnish and Estonian dependency parsers, I reveal that the models often fail to recognise morphology as the primary indicator of syntax; instead of generalising based on the case and agreement markings, they learn to over-rely on word order and lexical semantics. Following this finding, I experiment with two methods of increasing the models' reliance on morphology: one based on the alteration of the training data and another involving an enhanced training objective. Finally, through creating synthetic CMLs by manipulating selected typological properties of Polish, I investigate whether the models have a 'preference' for the means of encoding case information and reveal that syncretism and high fusion are amongst the properties that drive the models away from relying on morphology as a signal to subject/objecthood.



## ACKNOWLEDGEMENTS

---

First and foremost, I would like to thank my supervisor, Prof Ann Copestake, for her invaluable support and guidance throughout this project. Despite her busy schedule as the Head of the Department she was always there for me, offering words of wisdom through the ups and the downs of my PhD journey. I could not have asked for a more knowledgeable and supportive mentor.

Another big thank you goes to Prof Ryan Cotterell, who co-advised me through the second half of this journey. I am grateful for his insights, mentorship and many fruitful discussions which inspired me to shift the direction of my early PhD research towards the questions posed in this thesis.

I would also like to thank the rest of the NLIP group, particularly Siân, Guy A., Guy E. and Chris, as well as my college friends, especially Monica, Max, Anna and Mark, for the PhD camaraderie which kept me sane throughout the past years.

I am also grateful for the Vice-Chancellor's & Selwyn College Scholarship, the financial support from the Department of Technology and Computer Science, and the G. D. Fahrenheit Scholarship from the City Council of Gdansk, which enabled my studies at Cambridge.

Finally, the biggest thank you goes to my family, especially my parents, Marta and Chris, for their never-ending support, encouragement and unwavering belief in me.



# CONTENTS

---

|          |                                                             |           |
|----------|-------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                         | <b>17</b> |
| 1.1      | Motivation . . . . .                                        | 17        |
| 1.2      | This thesis . . . . .                                       | 19        |
| 1.2.1    | Thesis outline . . . . .                                    | 20        |
| 1.2.2    | Key contributions . . . . .                                 | 22        |
| 1.3      | Research papers . . . . .                                   | 23        |
| <b>2</b> | <b>Linguistic background and motivation</b>                 | <b>25</b> |
| 2.1      | Dependency grammar . . . . .                                | 26        |
| 2.1.1    | Universal Dependencies . . . . .                            | 26        |
| 2.2      | Morphology . . . . .                                        | 29        |
| 2.2.1    | Two traditions of morphological description . . . . .       | 29        |
| 2.2.2    | Word, lexeme, inflected form, lemma, paradigm . . . . .     | 30        |
| 2.2.3    | Morpheme, morph, allomorph . . . . .                        | 30        |
| 2.2.4    | Affixation and other morphological patterns . . . . .       | 31        |
| 2.2.5    | Inflection vs derivation . . . . .                          | 32        |
| 2.3      | Morphosyntax . . . . .                                      | 33        |
| 2.3.1    | Government and agreement . . . . .                          | 33        |
| 2.3.2    | Inflectional features . . . . .                             | 34        |
| 2.4      | Inflectional morphology in language understanding . . . . . | 38        |
| 2.4.1    | Different signals of grammatical function . . . . .         | 39        |
| 2.4.2    | Coding strategies of Polish and Russian . . . . .           | 40        |
| 2.4.2.1  | Case . . . . .                                              | 41        |
| 2.4.2.2  | Agreement . . . . .                                         | 43        |
| 2.4.3    | Coding strategies of Finnish and Estonian . . . . .         | 44        |
| 2.4.3.1  | Case . . . . .                                              | 44        |
| 2.4.3.2  | Agreement . . . . .                                         | 47        |
| 2.4.4    | Syncretism: where morphology fails syntax . . . . .         | 47        |
| 2.4.5    | Disambiguation strategies . . . . .                         | 50        |
| 2.4.5.1  | Word order freezing . . . . .                               | 51        |

|          |                                                                                                    |           |
|----------|----------------------------------------------------------------------------------------------------|-----------|
| 2.4.5.2  | Lexical semantics . . . . .                                                                        | 51        |
| 2.5      | Human language processing and acquisition: the competition model . . . .                           | 53        |
| 2.5.1    | The competition model . . . . .                                                                    | 53        |
| 2.5.2    | Insights into Slavic and Finnic languages . . . . .                                                | 55        |
| 2.6      | Morphological typology . . . . .                                                                   | 56        |
| 2.6.1    | Traditional view of morphological typology . . . . .                                               | 56        |
| 2.6.2    | Typology of Bickel and Nichols . . . . .                                                           | 57        |
| 2.6.2.1  | Phonological fusion . . . . .                                                                      | 58        |
| 2.6.2.2  | Exponence . . . . .                                                                                | 59        |
| 2.6.2.3  | Flexivity . . . . .                                                                                | 59        |
| 2.6.2.4  | Synthesis . . . . .                                                                                | 59        |
| 2.6.2.5  | Beyond fusion, exponence, flexivity and synthesis . . . . .                                        | 60        |
| 2.7      | The unique challenge . . . . .                                                                     | 60        |
| 2.7.1    | Morphological competence – the Polish example . . . . .                                            | 61        |
| 2.7.2    | Russian, Finnish, Estonian . . . . .                                                               | 62        |
| 2.7.3    | English-centrism in NLP . . . . .                                                                  | 64        |
| 2.7.4    | Conclusion . . . . .                                                                               | 64        |
| <b>3</b> | <b>Morphological competence of neural models</b>                                                   | <b>65</b> |
| 3.1      | Laying out the terminology and the expectations . . . . .                                          | 66        |
| 3.1.1    | The modern NLP paradigm . . . . .                                                                  | 66        |
| 3.1.2    | Signals, features and their relative importance . . . . .                                          | 67        |
| 3.1.3    | Morphology vs word order vs lexical semantics . . . . .                                            | 68        |
| 3.1.4    | Relative input signal sensitivity . . . . .                                                        | 69        |
| 3.1.5    | A categorical rule for morphology? . . . . .                                                       | 70        |
| 3.2      | Misleading success on relevant NLP benchmarks . . . . .                                            | 71        |
| 3.2.1    | Promising evidence from UD parsing . . . . .                                                       | 71        |
| 3.2.2    | (But) models may not use morphosyntactic information . . . . .                                     | 73        |
| 3.2.2.1  | Shortcomings of standard NLP benchmarks . . . . .                                                  | 73        |
| 3.2.2.2  | To encode $\neq$ to use . . . . .                                                                  | 74        |
| 3.3      | Dependency parsing: a case study of benchmark shortcomings . . . . .                               | 75        |
| 3.3.1    | Experimental setup . . . . .                                                                       | 76        |
| 3.3.2    | Results . . . . .                                                                                  | 78        |
| 3.3.3    | Discussion . . . . .                                                                               | 81        |
| 3.4      | Related research efforts: linguistically motivated analysis of case-marking<br>languages . . . . . | 81        |
| 3.4.1    | Diagnostic classifiers . . . . .                                                                   | 82        |
| 3.4.2    | Agreement and acceptability prediction . . . . .                                                   | 84        |
| 3.4.3    | Experiments with competing hypothesis . . . . .                                                    | 86        |

|          |                                                                                               |            |
|----------|-----------------------------------------------------------------------------------------------|------------|
| 3.4.4    | The remaining gap in the English-centric research . . . . .                                   | 86         |
| 3.5      | Targeted evaluation in the presence of conditional constraints: a proposal .                  | 87         |
| 3.5.1    | Methodology and a road-map . . . . .                                                          | 88         |
| 3.5.2    | Final remarks . . . . .                                                                       | 89         |
| <b>I</b> | <b>Investigation of relative importance of linguistic signals</b>                             | <b>91</b>  |
| <b>4</b> | <b>Right for the wrong reasons: Reliance on word order in neural dependency parsing</b>       | <b>93</b>  |
| 4.1      | Analysis of signal validity in UD . . . . .                                                   | 95         |
| 4.1.1    | Inflectional morphology . . . . .                                                             | 95         |
| 4.1.2    | Word order . . . . .                                                                          | 97         |
| 4.1.3    | Discussion . . . . .                                                                          | 98         |
| 4.2      | Counterfactual treebanks . . . . .                                                            | 100        |
| 4.2.1    | Treebank filtering . . . . .                                                                  | 101        |
| 4.2.2    | Maintaining acceptability . . . . .                                                           | 102        |
| 4.2.3    | Reordering . . . . .                                                                          | 103        |
| 4.2.4    | Proposed evaluation . . . . .                                                                 | 104        |
| 4.3      | Experimental details . . . . .                                                                | 106        |
| 4.4      | Experiment I: Evaluation on unambiguous sentences . . . . .                                   | 108        |
| 4.4.1    | Results . . . . .                                                                             | 109        |
| 4.4.2    | Error analysis . . . . .                                                                      | 113        |
| 4.4.3    | Naturally occurring sentences . . . . .                                                       | 116        |
| 4.5      | Experiment II: Evaluation on ambiguous sentences . . . . .                                    | 117        |
| 4.5.1    | Results . . . . .                                                                             | 118        |
| 4.6      | Conclusion . . . . .                                                                          | 121        |
| <b>5</b> | <b>Accounting for lexical semantics</b>                                                       | <b>123</b> |
| 5.1      | LexMix: A framework to alter lexemes in UD treebanks for highly inflected languages . . . . . | 125        |
| 5.1.1    | Lexeme-based vs form-based perturbations . . . . .                                            | 125        |
| 5.1.2    | The LexMix framework . . . . .                                                                | 126        |
| 5.1.3    | Comparison with other work . . . . .                                                          | 129        |
| 5.2      | Experiment I: Noun lexeme permutation . . . . .                                               | 130        |
| 5.2.1    | Simulation-based experimental design . . . . .                                                | 131        |
| 5.2.1.1  | Not-so syntactic relations . . . . .                                                          | 132        |
| 5.2.1.2  | Data and models . . . . .                                                                     | 132        |
| 5.2.2    | Results . . . . .                                                                             | 133        |

|         |                                                                        |     |
|---------|------------------------------------------------------------------------|-----|
| 5.2.2.1 | LAS distributions . . . . .                                            | 133 |
| 5.2.2.2 | Performance drop noted for individual relations . . . . .              | 136 |
| 5.3     | Experiment II: Core argument rotation . . . . .                        | 137 |
| 5.3.1   | Experimental details . . . . .                                         | 137 |
| 5.3.2   | Results . . . . .                                                      | 139 |
| 5.3.2.1 | Unambiguous sentences . . . . .                                        | 139 |
| 5.3.2.2 | Ambiguous sentences . . . . .                                          | 140 |
| 5.4     | Experiment III: Morphology vs word order + lexical semantics . . . . . | 141 |
| 5.4.1   | Lemmatising baselines . . . . .                                        | 142 |
| 5.4.2   | Morphologically aware parsers . . . . .                                | 145 |
| 5.4.2.1 | LAS results . . . . .                                                  | 145 |
| 5.4.2.2 | Error analysis . . . . .                                               | 146 |
| 5.4.3   | Discussion in the light of earlier results . . . . .                   | 148 |
| 5.5     | Other transformer models . . . . .                                     | 149 |
| 5.6     | Conclusion . . . . .                                                   | 151 |

## II Particulars of models' shortcomings 153

### 6 Increasing morphological awareness of dependency parsers 155

|         |                                                              |     |
|---------|--------------------------------------------------------------|-----|
| 6.1     | Breaking word order and lexical correlations . . . . .       | 156 |
| 6.1.1   | Experimental details . . . . .                               | 157 |
| 6.1.1.1 | Training-data perturbations . . . . .                        | 157 |
| 6.1.1.2 | Evaluation methods . . . . .                                 | 158 |
| 6.1.2   | Results (unambiguous clauses) . . . . .                      | 160 |
| 6.1.2.1 | Word order perturbation only (PERTURBATION I) . . . . .      | 160 |
| 6.1.2.2 | Introducing lexical perturbation (PERTURBATION II) . . . . . | 162 |
| 6.1.2.3 | Altering more types of clauses (PERTURBATION III) . . . . .  | 163 |
| 6.1.3   | Results (ambiguous clauses) . . . . .                        | 164 |
| 6.1.4   | Discussion . . . . .                                         | 165 |
| 6.2     | Multi-task training . . . . .                                | 166 |
| 6.2.1   | Experimental details . . . . .                               | 166 |
| 6.2.2   | Results . . . . .                                            | 167 |
| 6.2.2.1 | Morphosyntactic tagging accuracy . . . . .                   | 167 |
| 6.3     | Conclusion . . . . .                                         | 170 |

### 7 The effects of information encoding: Experiments with synthetic versions of Polish 173

|     |                        |     |
|-----|------------------------|-----|
| 7.1 | Related work . . . . . | 174 |
|-----|------------------------|-----|

|          |                                                            |            |
|----------|------------------------------------------------------------|------------|
| 7.2      | Synthetic languages . . . . .                              | 175        |
| 7.3      | Experimental design . . . . .                              | 178        |
| 7.4      | Results . . . . .                                          | 178        |
| 7.5      | Conclusion . . . . .                                       | 181        |
| <b>8</b> | <b>Conclusion</b>                                          | <b>183</b> |
| 8.1      | What next? . . . . .                                       | 184        |
| 8.1.1    | Further explorations of RISS . . . . .                     | 185        |
| 8.1.2    | Towards more morphologically competent models . . . . .    | 186        |
|          | <b>References</b>                                          | <b>187</b> |
| <b>A</b> | <b>Universal Dependencies taxonomy</b>                     | <b>239</b> |
| A.1      | Dependency relations . . . . .                             | 239        |
| A.2      | Part-of-speech . . . . .                                   | 241        |
| A.3      | Morphological features . . . . .                           | 242        |
| <b>B</b> | <b>Leipzig glossing abbreviations</b>                      | <b>243</b> |
| <b>C</b> | <b>Overview of the relevant neural architectures</b>       | <b>247</b> |
| C.1      | The skip-gram model . . . . .                              | 247        |
| C.2      | Convolutional neural networks . . . . .                    | 248        |
| C.3      | Long short-term memory networks . . . . .                  | 250        |
| C.4      | Transformer models . . . . .                               | 253        |
| C.4.1    | Attention . . . . .                                        | 253        |
| C.4.2    | The original transformer (Vaswani et al., 2017) . . . . .  | 255        |
| <b>D</b> | <b>Allowing for morphology in neural NLP</b>               | <b>259</b> |
| D.1      | Composition of morphs . . . . .                            | 260        |
| D.2      | Composition of characters . . . . .                        | 261        |
| D.3      | Composition of character n-grams and word-pieces . . . . . | 264        |
| D.4      | Paradigm and tag-based approaches . . . . .                | 266        |
| D.5      | Comparison of approaches . . . . .                         | 267        |
| <b>E</b> | <b>Sentence simplification decisions</b>                   | <b>271</b> |
| E.1      | Accepted/rejected relations overview . . . . .             | 272        |
| E.1.1    | Core clausal relations . . . . .                           | 272        |
| E.1.2    | Non-core clausal relations . . . . .                       | 272        |
| E.1.3    | Nominal relations . . . . .                                | 273        |
| E.1.4    | Linking relations . . . . .                                | 274        |
| E.1.5    | MWE relations . . . . .                                    | 274        |

|          |                                                          |            |
|----------|----------------------------------------------------------|------------|
| E.1.6    | Special relations . . . . .                              | 274        |
| <b>F</b> | <b>Lexical categorisation of UD dependency relations</b> | <b>275</b> |
| F.1      | Categories of relations (based on LAS) . . . . .         | 275        |
| F.2      | Rationale . . . . .                                      | 275        |
| <b>G</b> | <b>Supplementary results</b>                             | <b>293</b> |





# INTRODUCTION

---

*Morphological complexity is not just there to increase data sparsity.*

(Bender, 2013)

## 1.1 Motivation

Recognizing how different words relate to one another in a sentence – i.e., the “who does what to whom” – is a central task in language understanding. This information can be signalled in a language through one or more of the following overt coding strategies: case marking, agreement and word order (Haspelmath et al., 2001, Siewierska and Bakker, 2008, Slobin and Bever, 1982). English – a language which dominates the field of NLP – relies primarily on the position of words in a sentence to signal grammatical function. But there are many languages that do not resemble English in its reliance on word order as the primary coding strategy. In case-marking languages (CMLs), such as Polish or Finnish, the “who does what to whom” is encoded via morphological means instead.

It naturally follows that, to successfully process a CML, a neural model must be particularly sensitive to morphology. To retrieve the core sentence meaning, it has to (i) recognise which subword units carry meaningful information, (ii) capture the function of those units, (iii) interpret this information in a broader clausal context and, last but not least, (iv) recognise the importance of this signal, ranking it above alternative cues. But despite the importance of such *morphological competence* for CML processing, the NLP models for such languages are never directly tested for that linguistic ability. This contrasts with an abundance of analogous work on English, which tests English models’ sensitivity to word order (Abdou et al., 2022, Clouatre et al., 2021, Futrell and Levy, 2019, Gupta et al., 2021, Hessel and Schofield, 2021, Papadimitriou et al., 2022, Sankar et al., 2019, Sinha et al., 2021). This gap in the literature is the more important given that

most neural NLP models are developed with English language in mind and later applied, without any adaptations, to other languages. It remains unclear whether the architectures and optimization techniques developed on English are able to extract all the essential information from the word-forms of CMLs and whether this information is used to solve NLP tasks.

Studying CML models’ morphological competence is essential to ensure that they “get the right answers for the right reasons”, but that is not the only reason why it is worthwhile. The CMLs use of morphology to express core sentence meaning also provides a unique and interesting challenge to language modeling – one which is very unlike the English word order equivalent. Unlike the English ordering of the subject, verb and object, morphological case and agreement marking can be ambiguous. When this happens, a sentence has more than one correct parse and additional strategies are required to disambiguate it. These often involve relying on word order and/or lexical semantics. This means that to accurately model the grammar of a language, the models need to recognise this conditionally-determined importance of competing linguistic signals and rely on morphology when it is unambiguous and word order/lexical semantics otherwise. This task is far from trivial, as I demonstrate in Chapter 2.

Bringing the above challenge to light (and addressing it) is the more important given the many misconceptions about morphology held within the NLP community. Perhaps the most problematic misconception is that morphology makes the modeling easier, exemplified in the following statements:

“We hypothesize that language models would be better able to learn hierarchical syntactic generalizations in morphologically complex languages (which provide frequent overt cues to syntactic structure) than in morphologically simpler languages.” (Mueller et al., 2020)

“Overt morphological case makes agreement prediction significantly easier regardless of word order.” (Ravfogel et al., 2019)

“English was by far the hardest language [...] due to its poorer morphology and higher POS ambiguity.” (Gulordava et al., 2018)

These types of conclusions are based on experimental results that always have other explanations and send a harmful message that studying models’ reliance on morphology is unnecessary and uninteresting. Throughout this thesis I demonstrate that, in contrary to the above beliefs, morphology often makes language modeling not easy, but *hard*.

## 1.2 This thesis

This thesis constitutes a comprehensive study of morphological competence of CMLs, motivated by a wish to fill the aforementioned gap in the evaluation of neural models. Specifically, I aim to answer whether the modern neural models are suited to extract the relevant information encoded in the word-forms of CMLs and interpret it globally, in a broader clausal context, as the main indicator of core sentence meaning. I carry out my investigation through linguistically motivated, targeted experiments in the scope of dependency parsing. Dependency parsing naturally requires the models to identify core relations. At the same time, it requires little-to-no world knowledge and is more elaborate than alternative syntax-targeting tasks, such as agreement prediction (discussed in more detail in Chapter 3) which makes it a more realistic scenario for how neural models are typically used in production.

To answer the above question I propose a new evaluation paradigm which allows for examination and tracking of the generalisation strategies of a neural model. Within this paradigm I evaluate neural parsers on different counterfactual versions of existing dependency corpora in which one plausible linguistic signal (word order, lexical semantics, morphology) is placed in opposition to the other two. This approach not only allows me to estimate a model’s reliance on morphology, but also to compare it to its reliance on the alternative, secondary signals to meaning on which it should develop only conditional dependence.

Throughout the thesis, I train 270 parsers for four CMLs – Polish, Russian, Finnish and Estonian – and evaluate them within the proposed paradigm. Through this experimentation I reveal that state-of-the-art neural parsers often fail to recognise morphology as the primary indicator of syntax, instead learning to over-rely on word order and lexical semantics. I found this holds regardless of the particulars of the underlying architecture, hyper-parameters, pre-training objective/data or whether the model is mono- or multilingual. In other words, the trend applies to many (if not all) neural models that form the basis of modern state-of-the-art NLP. Following this finding, in the second part of the thesis, I investigate the particulars of the models’ shortcomings; first, by testing if they can be pushed to rely on morphology through adjustments to their training data/objective; next, by investigating whether their resistance to rely on morphology as a signal to meaning can be tied to specific aspects of the morphological system of the modeled language. Through these experiments I show that the state-of-the-art neural parsers for CMLs have a far greater capacity to rely on morphology than what they settle for when trained on natural data – in many instances where the models *could* rely on morphology as the primary signal, they end up relying on word order and lexical semantics instead. I also reveal that the models do have a ‘preference’ for how case information is encoded – when trained on

synthetic languages with more explicit morphs and no case syncretism they almost always rely on morphology over the alternative cues.

Lastly, I find it important to note that while almost all of my investigations in this thesis concern the task of dependency parsing within the Universal Dependencies scheme (UD) (Nivre et al., 2016), the shortcomings I reveal are likely to apply also to other benchmarks and other NLP tasks that rely on detection of core sentence meaning. This is because the word order and lexical tendencies that the models pick up on are not merely an artifact of the data or a heuristic signal, but an inherent property of a language. With this thesis I only take one step within what I believe to be an extremely important and at the same time extremely under-explored area within NLP.

### 1.2.1 Thesis outline

Chapter 2 provides the necessary linguistic background. The main aim here is to introduce relevant terminology and point the reader towards the aspects of the studied languages which are most relevant for this thesis. In this chapter I also argue that the linguistic characteristics of Polish, Russian, Finnish and Estonian pose a challenge for both development and evaluation of neural NLP models for tasks that rely on sentence meaning. This in turn makes them a very interesting and important case for the study of neural models' morphological competence.

In Chapter 3 I bring to light an important gap in neural NLP research by demonstrating that research to date offers little insight into whether neural models are capable of relying on morphology as a signal to meaning. I do so through reviewing related work and discussing some of my preliminary experiments in the scope of dependency parsing. I conclude this chapter with a proposal for an experimental paradigm which fills the aforementioned evaluation gap and a road-map for the rest of the thesis.

Chapter 4 is the first fully experimental chapter and it is devoted to investigating the models' reliance on word order as a signal to core sentence meaning. It has two parts: (i) an analysis of the validity of word order and morphology signals in the UD *treebanks* and (ii) experiments which aim to reveal the relative importance-based ranking of those signals within a number of neural parsers trained on UD. The latter is facilitated by a new dataset, which places word order in opposition to morphology and lexical semantics. Through experiments in this chapter I reveal that models trained on Polish, Russian, Estonian and Finnish UD corpora over-rely on word order as a cue to core relations.

In Chapter 5 I turn my focus to lexical semantics and conduct a series of experiments, in which I manipulate the lexemes in UD sentences, at evaluation, to study the lexi-

cosemantic influence on models' predictions. Two of my experiments from this chapter place lexical semantics in direct opposition to morphology and word order. In the third experiment I bring together the word order and lexical alterations, putting morphology in direct opposition to both word order and lexicosemantic cues, posing it as the only linguistic signal that points to the correct interpretation. All these experiments reveal that the models frequently over-rely on lexical semantics and under-rely on morphology as a cue to core grammatical relations. They also lead to further important insights relating to the models' generalisations, which do not emerge from standard, non-targeted evaluation.

In the light of my results from the two earlier chapters, in Chapter 6 I take the first step towards building better, more morphologically competent models for CMLs and explore two straightforward approaches to improving the models' morphological competence. The first approach involves augmentation of the training data; the second, incorporation of an additional training objective. Through evaluating the new models within the evaluation paradigm proposed across Chapters 3–5, I reveal that the models' generalisation strategies can be improved through training data perturbations, suggesting that the parsers explored in the earlier chapters have a far greater capacity to rely on morphology as a cue to meaning than what they settle for.

In Chapter 7 I ask whether the resistance to rely on morphology as a signal to meaning brought to light in the earlier chapter (the models *could* rely on morphology, but tend not to) can be tied to specific aspects of the morphological system of the modeled language. In other words, do the language idiosyncrasies affect what the models end up relying on? To answer this question I propose a framework grounded in the morphological typology of Bickel and Nichols (2007), which involves creating synthetic versions of a language by manipulating selected typological properties. Through experiments on such synthetic versions of Polish, I reveal that this is indeed the case and that case syncretism and high fusion are amongst the properties that drive the models away from relying on morphology as a signal to subject/objecthood.

Finally, in Chapter 8 I conclude the thesis, reiterating the most relevant insights and outlining promising directions for future research which can be categorised into two streams: (i) further explorations into the facets of neural models' relative sensitivities to different types of input signals (a property which I discuss in Section 3.1), and (ii) improving the models' generalisations via introducing changes to their architectures and/or more elaborate training paradigms.

### 1.2.2 Key contributions

First, I bring to light a challenge within modeling of CMLs with neural networks, which has not yet been discussed or addressed within the NLP literature (Chapter 2). This challenge involves building models that can recognise the conditionally-determined importance of competing linguistic signals, relying on morphology when it is unambiguous and word order/lexical semantics otherwise. The importance of addressing this challenge goes beyond NLP for CMLs, since similar conditional constraints are bound to apply in many other scenarios, spanning different languages and tasks.

Second, I propose a novel methodology for evaluating models for CMLs with respect to key linguistic capabilities that are prerequisites for correct linguistic generalisation (Chapters 3, 4 and 5).

Third, I conduct a thorough assessment of generalisation strategies of single-objective state-of-the-art dependency parsers for Polish, Russian, Finnish and Estonian, revealing shortcomings in their generalisations which would not have come to light in standard (non-targeted) evaluation (Chapters 4, 5).

Fourth, I explore methods of improving the models' generalisation strategies, demonstrating that this can be done in a straightforward way through linguistically-motivated training data alterations (Chapter 6).

Fifth, I propose a framework for deeper study of the models' sensitivities to different types of linguistic signals, which can reveal which aspects of morphological systems are the most challenging to the models. Through preliminary experiments within this framework, I uncover properties of the Polish case system that push the models away from relying on morphology as a signal to meaning (Chapter 7).

Sixth, I release the data and code for this project to facilitate further study of morphological competence in NLP models. This includes 3 different datasets, each covering Polish, Russian, Finnish and Estonian, and code which facilitates a wide range of UD treebanks' perturbations, including those that necessitate reinflection (see Section 5.1.2) and which are not presently supported by any other existing framework of this kind.

## 1.3 Research papers

### Thesis related content

*Both papers constitute entirely my own work, with all co-authors effectively acting as supervisors.*

1. **Paula Czarnowska**, Ryan Cotterell and Ann Copestake. On the Case of Cased Languages: Word Order vs Morphology in Neural Dependency Parsing. *(in review)*
2. **Paula Czarnowska**, Yogarshi Vyas, Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. Transactions of the Association for Computational Linguistics (TACL), 2021.

### Content not directly related

1. **Paula Czarnowska**, Sebastian Ruder, Ryan Cotterell and Ann Copestake. Morphologically Aware Word-Level Translation. In Proceedings of 2020 International Conference on Computational Linguistic (COLING), 2020.
2. **Paula Czarnowska**, Sebastian Ruder, Edouard Grave, Ryan Cotterell and Ann Copestake. Don't forget the long tail! A comprehensive analysis of morphological generalization in bilingual lexicon induction. In Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
3. **Paula Czarnowska**, Guy Emerson and Ann Copestake. Words are vectors, dependencies are matrices: learning word embeddings from dependency graphs. In Proceedings of the 13th International Conference on Computational Semantics (IWCS), 2019.



---

## LINGUISTIC BACKGROUND AND MOTIVATION

---

In this chapter I set out the context for the thesis by providing the necessary linguistic background. I start with a description of *dependency grammar* (Section 2.1), *morphology* (Section 2.2) and *morphosyntax* (Section 2.3). These three sections, quite general and introductory in nature, are followed by sections which target the topic of this thesis more directly. The first two of those sections – Section 2.4 and Section 2.5 – discuss the role of inflectional morphology in human language comprehension, and in Polish, Russian, Finnish and Estonian, in particular. Both of these sections constitute an important discussion which motivates a large portion of ideas and methodologies employed in this thesis. Next, in Section 2.6, I give an overview of morphological typology, which not only provides a broader context for the whole thesis, but is especially relevant for Chapter 7 – the experiments from Chapter 7 are grounded in the typological classification of Bickel and Nichols (2007) discussed in Section 2.6.

To conclude the chapter, in Section 2.7, I tie all of the previous sections together, and discuss how the linguistic characteristics of Polish, Russian, Finnish and Estonian make them an interesting and important case for study of neural models’ morphological competence.

While many of the concepts discussed in this chapter, and in particular those from sections 2.1 – 2.3, may be familiar to the reader, I define them here to (i) introduce the terminology used throughout the thesis, (ii) clarify the definitions of ambiguous terms and (iii) point the reader to the aspects of the studied languages most relevant for this thesis.

## 2.1 Dependency grammar

Syntax of a language provides a non-linear structure that defines how words and phrases relate to one another in a sentence. This structure can take different formal shape, depending on the adopted syntactic theory. The theory most relevant for this thesis is **dependency grammar**, since all of the experiments that constitute this thesis involve the task of neural dependency parsing.

Within dependency grammar, syntactic structure takes the shape of asymmetric *dependency relations* between pairs of words in a sentence (de Marneffe and Nivre, 2019). Each of those relations represents the way in which one of the words, the **dependent**, relies on another word, the **head**, for its grammatical form and linear position (Mel’čuk, 2009). Such dependency structures can be formally defined as directed graphs, consisting of a set of vertices – words or punctuation – and a set of arcs between them, leading from the head to the dependent. The graphs are often constrained, so that each word can be a dependent of exactly one head, consequently yielding a **dependency tree**. The arcs are typically assigned labels, describing the role played by the dependent with respect to its head. See Figure 2.1 for an example.

Dependency grammar frameworks can be best described as syntactic analysis schemes, rather than coherent theoretical frameworks (de Marneffe and Nivre, 2019). Their simplicity prevents them from encoding relevant elements of syntactic structure, such as relations between coordinated items which are heads of the same dependent or dependents of the same head (de Marneffe and Nivre, 2019, Tesnière, 1959, 2015). Nevertheless, this constrained nature is also what makes them particularly attractive for NLP practitioners – it makes the parsing task simpler and the representation more conceptually accessible. Another strength of dependency structures is their insensitivity to the order of words in a sentence – an attribute which makes them particularly suitable for flexible word order languages, like those studied in this thesis. Finally, they directly encode important information about word associations which is often buried in more complex representations, such as constituent-based parses (Jurafsky and Martin, 2017, p. 310). In doing so, they highlight the *grammatical function* of each word in a sentence, in turn providing support for semantic interpretation (Andrews, 2007, Bender, 2013).

### 2.1.1 Universal Dependencies

The experiments in this thesis employ corpora annotated with the Universal Dependencies (UD) scheme. UD is a framework for cross-linguistically consistent annotation of human language, which involves both morphology and syntax schemes (de Marneffe et al., 2021, Nivre et al., 2016, 2020).

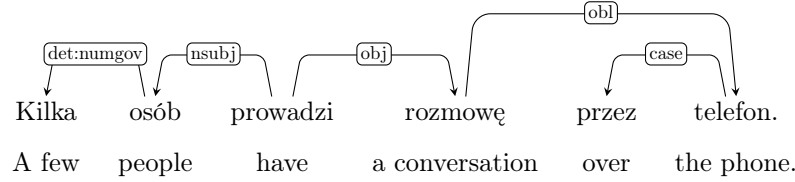


**Figure 2.1:** An example of a Polish UD dependency tree. The definition of labels from left to right: determiner which governs the number of the head, nominal subject, direct object, an oblique argument of a verb and a case marker.

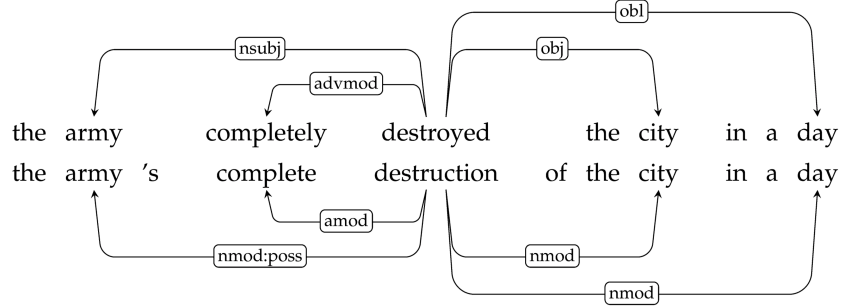


**Figure 2.2:** Partial analysis for a clause (top) and a nominal (bottom) (de Marneffe et al., 2021).

## Syntactic scheme

The syntactic dependency scheme grew out of the Stanford Dependency representations (de Marneffe and Manning, 2008, De Marneffe and Manning, 2008). It distinguishes 37 universal grammatical relations (see Appendix A.1 for the full list), which can further be augmented with additional, language-specific relation sub-types (e.g., *det:numgov* in Figure 2.1). Importantly, not all of those relations hold between a syntactic head and its dependent. Some, like *det* (determiner), *aux* (auxiliary) or *case* hold between a lexical head and its grammatical marker. Others, like *flat*, *compound* or *goeswith*, are not even syntactic and are used for lexical or other phenomena, such as compounding or annotating incorrectly edited text. The status of different relations is discussed in more detail in Chapter 5, for which it is most relevant.

UD adopts a lexicalist view of syntax, treating words as the basic units of annotation; the clitics are separated from their hosts and the contractions are undone. The annotation scheme is organised around three primary linguistic units: (i) a nominal, representing an entity (e.g., a predicate or an argument), (ii) a clause, representing an event and (iii) a modifier, which can refine nominals and clauses by describing their attributes. The relations are organized around the distinction between those units – a different type of relation is used depending on whether “the head is the head of a clause or a nominal, and whether the dependent is a clause, nominal or a modifier” (de Marneffe et al., 2021). This is illustrated in Figure 2.2, presenting partial analysis of a clause, headed by a predicate *destroyed* and a nominal, headed by the noun *destruction*.

An important aspect of the scheme is that it makes ‘higher-level’ connections between

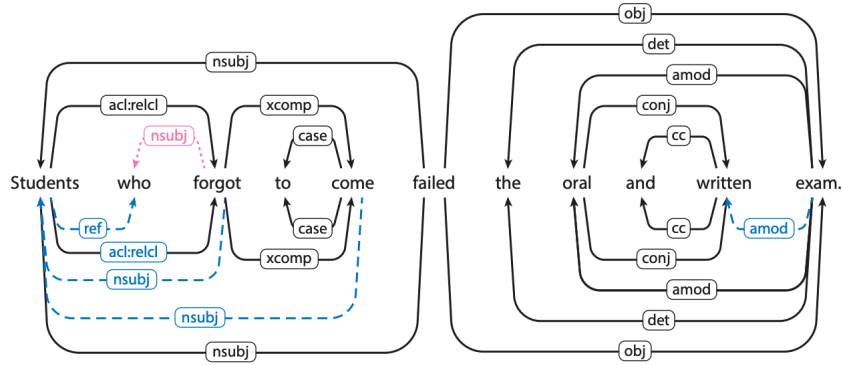


**Figure 2.3:** An example of a basic Universal Dependencies tree (top) and the enhanced dependency graph (bottom) from de Marneffe and Nivre (2019).

words, consequently representing “a sentence’s observed surface predicate-argument structure” (de Marneffe et al., 2021). In cases of a **dissociated nucleus** (Tesnière, 2015), where functions of a head are divided between a *structural center* (an auxiliary or a function word) and a *semantic center* (a lexical or a content word), UD chooses the latter as the head and connects the function words as dependents. One example of this behaviour is in handling prepositions which are treated as case-marking elements, dependent on head nouns. The nouns can then be directly connected to the head of the phrase to which the prepositional phrase was attached to, as presented in Figure 2.1. A similar treatment is also incorporated for copulas.

Beyond the basic annotation scheme, UD also defines another representation layer, called **enhanced dependencies**. This layer includes additional, deeper relations, not represented within the basic scheme. Figure 2.3 provides an illustrative example from de Marneffe and Nivre (2019), comparing the two annotations; in contrast to the basic scheme, the enhanced scheme encodes (i) the implicit subject in the control construction and (ii) the coreference link from the antecedent *students* to the relative pronoun *who*. It also directly connects the modifier *written* from the coordination *oral and written* to the modified *exam*. Note that in the basic scheme, due to the dependency tree format, the conjuncts *oral* and *written* cannot both be represented as syntactic heads of the coordinate structure. The UD follows a convention of labeling the first conjunct in the linear order as the head of all other conjuncts.

## POS and morphological annotation

UD part-of-speech (POS) tags evolved from the Google universal POS tags (Petrov et al., 2012) and distinguish 17 coarse-grained POS (see Appendix A.2 for the full tag set). These POS can be further divided into sub-classes via special set of features marked on nodes. For example, *VerbForm* feature can be used to distinguish between finite (VerbForm=Fin) and non-finite (VerbForm=Inf) forms or to mark verbal participles (VerbForm=Part).

The morphological annotation consists of a set of feature-value pairs and is based on the morphological features from the Intersect interlingua for morphosyntactic tag-sets (Zeman, 2008). Appendix A.3 lists the supported UD features, along with their possible values. Notably, in the case of dissociated nucleus, a feature belonging to the content word can be marked on a function word instead.

## Universal Dependencies treebanks

Besides being a cross-linguistic annotation scheme, UD also provides a collection of annotated treebanks. At the time of writing, the collection includes 183 treebanks for 104 languages (de Marneffe et al., 2021) (<https://universaldependencies.org/>). This includes 3 Polish, 4 Russian, 4 Finnish and 2 Estonian treebanks, discussed in more detail in the later chapters.

## 2.2 Morphology

A thesis on morphological competence would not be complete without a section with an overview of morphology. The one provided here is brief, with the primary focus on defining the most basic concepts, relevant for this thesis. Reviewing those concepts is especially important given the lack of standard terminology in the field. I omit discussions of particular morphological theories, since they are out of the scope for this thesis.

### 2.2.1 Two traditions of morphological description

Morphology can be broadly described as the study of internal structure and the formation of words. Theories of morphology can be split into two traditions, each making different assumptions about the word structure (Blevins, 2006, Blevins et al., 2019, Haspelmath and Sims, 2010). Within the first, **morpheme-based** or **subword** tradition, words are assumed to be composed of small meaningful units called **morphemes** (discussed in more detail in Section 2.2.3). Such approaches treat morphology as the ‘syntax of words’ and view morphological rules, which combine morphemes to yield words, in a similar way to syntactic rules, which combine words within a sentence (Haspelmath and Sims, 2010). The second tradition is **word-based**. It is primarily concerned with how forms of words change across different uses and constructions. The emphasis is placed on the “systematic covariation in the form and meaning of words” (Haspelmath and Sims, 2010), rather than decomposing words into smaller meaningful units. Words are viewed as the most useful unit of meaning for analysis and are not assumed to be composed of morphemes.

In this thesis I do not commit to either of those formal descriptions. I focus on empirically assessing models’ ability to retrieve and make use of information encoded

within the word form. The counterfactual methodology I propose for this purpose can be applied regardless of whether or not one assumes words are composed of morphemes. In the following subsections I discuss concepts relevant to either of the two traditions.

### 2.2.2 Word, lexeme, inflected form, lemma, paradigm

The most basic morphological notion is a **word**. There are many possible ways in which *word* can be understood and linguists have found it notoriously problematic to define (Arkadiev et al., 2018, Haspelmath and Sims, 2010, Haspelmath, 2011, Krámský, 1969). For the purpose of this thesis I use the operational definition of a *word* as a ‘contiguous sequence of letters’, which works well for the languages studied in this thesis.<sup>1</sup> However, I rarely refer to that concept. Instead, I refer to and distinguish between *inflected forms*, *lexemes* and *lemmata*.

A **lexeme** is an abstract concept that represents the core meaning shared by a set of inflected forms. RUN is an example of a verb lexeme which encodes the shared meaning of forms *run*, *runs*, *ran*, and *running*. Lexemes are abstract entities, which can be represented with a **lemma** (also called a citation form) – an inflected form lexicographers have chosen to be representative of a lexeme. For example, the lemma of RUN is the infinitive *run*. Finally, a **paradigm** is a set of all inflected forms associated with a given lexeme.

Note that I distinguish between the concepts of a *lemma* and a **stem** – the base form of a word, stripped of all inflectional markers. Consequently, I treat infinitive forms of verbs as their lemmas also in languages which, unlike English, mark infinitives with a unique affix; e.g., *biegać* is the Polish lemma of BIEGAĆ, which has a stem *biega*.

### 2.2.3 Morpheme, morph, allomorph

**Morpheme** is another basic morphological concept, but despite its prevalence in the linguistic literature, it does not have a universally agreed upon definition (Haspelmath, 2020). The term seems to be used in two distinct senses.<sup>2</sup> Some texts describe a morpheme as the smallest form in a language, where a **form** is a “recurrent pairing of content and segmental shape” (Haspelmath, 2020). This definition frequently figures in the introductions to morphology (Bender, 2013, Haspelmath and Sims, 2010, Kroeger, 2005) but it is often incompatible with actual usages of *morpheme* throughout the same text. For example, Kroeger (2005, pp. 12–14) talks about substrings *-ed*, *-dis* and *trust* in *trust-ed* and *dis-trust* as morphemes (making an assumption that morpheme *is* a form), while later (p. 288)

---

<sup>1</sup>This definition of a word is unsuitable for languages with writing systems that do not indicate word boundaries by blank spaces, such as Chinese, Japanese or Sanskrit.

<sup>2</sup>Haspelmath (2020) singles out three different senses of a morpheme. Here I place the emphasis on the difference between a morpheme as a form (sense 1 in (Haspelmath, 2020)) and a morpheme as an abstract unit (senses 2 and 3 in (Haspelmath, 2020)) and highlight only two distinct senses.

|            | CLASS I | CLASS II | CLASS III | CLASS IV |
|------------|---------|----------|-----------|----------|
| INFINITIVE | pić     | widzieć  | mieć      | wiedzieć |
| GLOSS      | ‘drink’ | ‘see’    | ‘have’    | ‘know’   |
| 1SG        | piję    | widzę    | mam       | wiem     |
| 2SG        | pijesz  | widzisz  | masz      | wiesz    |
| 3SG        | pije    | widzi    | ma        | wie      |
| 1PL        | pijemy  | widzimy  | mamy      | wiemy    |
| 2PL        | pijecie | widzicie | macie     | wiecie   |
| 3PL        | piją    | widzą    | mają      | wiedzą   |

**Table 2.1:** Present tense paradigms of four Polish verbs, each belonging to a different inflection class; an example of allomorphy.

talking about morphemes *appearing* in different forms. An alternative definition, assumed among others by Hockett (1947) and Morley (2000), treats a morpheme as an *abstract unit of meaning* which represents some minimal syntacticosemantic content. In that sense a morpheme is not a form but can be realized phonetically via forms called **morphs**. In line with this definition, a morph is the minimal linguistic form and the substrings *-ed*, *-dis* and *trust* in *trust-ed* and *dis-trust* are morphs rather than morphemes. In the remainder of this thesis, I assume the latter definition of a morpheme (an abstract concept) and employ the term *morph* when referring to minimal linguistic forms. Additionally, I follow Bickel and Nichols (2007) and use the term **formative**, when referring to markers of *inflectional information* which may or may not be forms (see Section 2.2.4).

(Abstract) morphemes may have multiple realizations, depending on e.g., phonological context or semantic properties of the base (Bender, 2013). In such instances, the different realizations of a morpheme are often called **allomorphs**. For example, the form of a verbal morpheme can depend on a verb’s inflection class (an arbitrary lexical property, see Section 2.3.2). This phenomenon is illustrated in Table 2.1, which presents Polish present tense paradigms for four verbs, each belonging to a different inflection class – the allomorphy can be observed in the suffixes for 1SG and 3PL paradigm slots. The stem of a word can also exhibit allomorphy. This is the case for the Polish stems *pije* (PIĆ), and *widzi* (WIDZIEĆ), which lose the final vowels *-e* and *-i* in 1SG and 3PL (see Table 2.1).

## 2.2.4 Affixation and other morphological patterns

Each of the word forms from Table 2.1 can be split into two separate morphs: one morph expressing the morphosyntactic category of a word and one base morph, expressing the core lexical meaning. The process of constructing new forms through concatenation of such basic units is called **affixation** and it is a type of a **morphological pattern**. At the core of affixation lies attaching affixes to stems – parts of words which convey the core lexical meaning. There are various kinds of affixes, each differently positioned within a

|              | SINGULAR | PLURAL    |  |  |
|--------------|----------|-----------|--|--|
| NOMINATIVE   | gruszka  | gruszki   |  |  |
| GENITIVE     | gruszki  | gruszek   |  |  |
| DATIVE       | gruszce  | gruszkom  |  |  |
| ACCUSATIVE   | gruszkę  | gruszki   |  |  |
| INSTRUMENTAL | gruszką  | gruszkami |  |  |
| LOCATIVE     | gruszcze | gruszkach |  |  |
| VOCATIVE     | gruszko  | gruszki   |  |  |

| SINGULAR | PLURAL |
|----------|--------|
| pear     | pears  |

**Table 2.2:** The noun paradigms of the Polish GRUSZKA and the English PEAR.

word. Suffixes follow the stem, while prefixes precede it. Other types of affixes include infixes, which occur inside the stem, and circumfixes that surround the stem, figuring on both of its sides, like *ge-* + *-en* in German *ge-fahr-en*.<sup>3</sup> Each affix can combine only with limited types of stems, determined by its **combinatory potential**. For example, the English suffix *-able* can only attach to verbs (Haspelmath and Sims, 2010).

Affixation is a type of a **concatenative** morphological pattern in which morphs are glued together to yield a new word. In contrast, **non-concatenative** patterns add no new segmentable material. An example is stem modification which can involve changes in pronunciation or in tone pattern of a word (Bender, 2013, Haspelmath and Sims, 2010).

## 2.2.5 Inflection vs derivation

The different inflected forms within a paradigm arise through the process of **morphological inflection** – the systematic alteration of a word form which adds grammatical information, e.g., number and case in Table 2.2 or number and person in Table 2.1. Importantly, inflection does not change the lexical information encoded in a word form. Rather, its primary purpose is to alter the form’s function so that it fits a specific grammatical context, e.g., as a subject or an object in a sentence.

A concept closely related to inflection is **derivation** – the process of forming word forms with novel meanings (i.e., belonging to new lexemes) through the modification of existing forms. For example, the word *unthinkable* can be constructed via adding the suffix *-able* and the prefix *un-* to the lemma *think*: *think* → *un* + *think* + *able*. In contrast to inflection, derivation changes the core, lexical meaning of the base word and it can also change its part of speech, as well as the argument structure (if the base lexeme is a verb). Both inflectional and derivational morphology pose interesting challenges for modeling natural languages, but this thesis is concerned exclusively with inflectional morphology. This is reflected in the selection of topics discussed in this chapter.

<sup>3</sup>The status of a circumfix as an affix is somewhat controversial. Haspelmath (2020) argues that, since an affix is a morph, it needs to be continuous, and so the discontinuous nature of a circumfix prevents it from being classified as an affix.

## 2.3 Morphosyntax

I now discuss morphosyntax; i.e., aspects of morphology that interface with syntax. First, I introduce agreement and government. Next, I give an overview of different features that can be expressed via inflectional morphology. The purpose of this section is two-fold; first, it provides the necessary foundations for the discussions that follow in the later sections. Second, by providing glossed examples from Polish and Finnish, it sheds more light on the morphosyntactic properties of the languages studied in this thesis.

For the linguistic glosses, I follow the Leipzig Glossing Rules (<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>) but I do not separate words into separate morphs. The abbreviations of all glosses can be found in Appendix B.

### 2.3.1 Government and agreement

Inflectional morphology is relevant to syntax because it expresses the “combinatoric potential of the inflected word forms” (Bender, 2013); i.e., which words go together within a sentence. It does so through two kinds of syntactic relations: agreement and government. In **syntactic agreement**, the morphological properties of one word (the controller) are expressed in the form of another word (the target). In noun agreement, the controller is usually the head noun and the typical targets include adjectives and determiners, as in the Polish sentences below:

- (1) a. Polish  
Widzę                      te                      czarne                      koty.  
see(IPFV).1SG.PRS this.PL.ACC.M.AN black.PL.ACC.M.AN cat(M.AN).PL.ACC  
‘I see those black (male) cats.’
- b. Polish  
Widzę                      tą                      czarną                      kotkę.  
see(IPFV).1SG.PRS this.SG.ACC.F black.SG.ACC.F cat(F).SG.ACC  
‘I see this black (female) cat.’

In the first sentence, the demonstrative form *te* (TEN) and the adjective form *czarne* (CZARNY) agree in gender, number, animacy and case with the accusative, plural form (*koty*) of an animate, male noun (KOT). Similarly, in the second sentence, forms *ta* and *czarna* agree with the accusative, singular form of a female noun (*kotkę*).<sup>4</sup> Note that in Polish, gender, number and case are expressed via a single morph. The above sentences are also a good example of how features involved in agreement may not be explicitly marked on the controller (to which they ‘belong’). In (1) this is the case for gender and animacy,

---

<sup>4</sup>In Polish animacy is only relevant for agreement with male nouns. The forms of targets agreeing with nouns of female and neuter gender do not depend on the animacy of those nouns.

which are lexical properties of nouns, but are not marked on nouns via any morphological patterns (see Section 2.3.2).

For verbs, the typical agreement pattern goes the other way around – it is the head verb that agrees with one or more of its dependents. In the following Polish sentences, the verbs agree with their subjects in person, number, gender and animacy:

- (2) a. Polish  
Jola przeczytała artykuł.  
Jola(F).SG.NOM read(PFV).3SG.PST.F article(M).SG.ACC  
'Jola read an article.'
- b. Polish  
Koty biegały.  
cat(M.AN).PL.NOM run(IPFV).3PL.PST.M.AN  
'The cats were running.'
- c. Polish  
Skończyłyśmy.  
finish(PFV).1PL.PST.F  
'We've finished (fem).'

Note that in example 2c the subject is missing from the sentence (i.e., it has been dropped) and the only information available about it is encoded in the agreement features overt on the verb – 1st person plural feminine. While it is not possible in Polish, some languages, like the Chicheŵa language, can drop both subject and object at the same time, marking all relevant features on the verb (Bender, 2013).

In **syntactic government** one word (the governor) requires other words to have particular morphological properties (instead of marking *its own* properties on other words, as in agreement). For example, most Polish verbs require their subject to appear in the nominative case and their objects in the accusative case. An example of this can be seen in examples (1), (2a) and (2b). This type of government is called **case government** and it plays an important role in communicating meaning within cased languages (see Section 2.4).

## 2.3.2 Inflectional features

Languages differ in what information they communicate through inflection but they often select from the following features: number, case, gender, animacy, person, tense, aspect and mood (see Table 2.3 for example values). Many of those features have figured in the previous sections; here they are discussed in more detail.

| NOUNS, PRONOUNS                                       | VERBS                                                         | ADJECTIVES, DEMONSTRATIVES,<br>RELATIVE PRONOUNS, ADPOSITIONS |
|-------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|
| <b>number</b><br>(singular, plural, ...)              | <b>number</b><br>(singular, plural, ...)                      | <b>number</b><br>(singular, plural, ...)                      |
| <b>person</b><br>(1st, 2nd, 3rd)                      | <b>person</b><br>(1st, 2nd, 3rd)                              | <b>person</b><br>(1st, 2nd, 3rd)                              |
| <b>gender</b><br>(masculine, feminine, ...)           | <b>gender</b><br>(masculine, feminine, ...)                   | <b>gender</b><br>(masculine, feminine, ...)                   |
| <b>animacy</b><br>(animate, inanimate,<br>human, ...) | <b>animacy</b><br>(animate, inanimate,<br>human, ...)         | <b>animacy</b><br>(animate, inanimate,<br>human, ...)         |
| <b>case</b><br>(nominative,<br>accusative, ...)       | <b>tense</b><br>(present, future, past)                       | <b>case</b><br>(nominative,<br>accusative, ...)               |
|                                                       | <b>aspect</b><br>(perfective, imperfective,<br>habitual, ...) |                                                               |
|                                                       | <b>mood</b><br>(indicative, subjunctive,<br>imperative, ...)  |                                                               |

**Table 2.3:** Common inflectional features and their example values (based on Table 5.1 from Haspelmath and Sims (2010)).

**Number** is “related to the cardinality of the set picked out by the referent” (Bender, 2013). Some possible values include singular (a single entity), plural (more than one, or some other number of entities)<sup>5</sup>, dual (two entities) and trial (three entities). Number is a property of a noun and can be marked on the noun itself and, through agreement, on the noun’s dependents and/or on the head verb.

**Case** is a property of a noun which reflects its syntactic and semantic role in a sentence. It is common for each case to have one basic, canonical role in a language. In many languages, the nominative canonically marks subjects, the accusative marks direct objects and the basic role of the genitive is to indicate possessors (Haspelmath and Sims, 2010). These canonical use cases are reflected in the following Polish and Finnish examples:

- (3) a. Polish  
 Znajoma                      Kasi                      go                      zatrudniła.  
 friend(F).SG.NOM Kasia(F).SG.GEN he.SG.ACC hire(PFV).3SG.PST.F  
 ‘Kasia’s friend hired him.’

<sup>5</sup>In some languages, including English, plural is also used for zero entities (Bender, 2013).

b. Finnish

Kasian ystävä palkkasi hänet.  
 Kasia.SG.GEN friend.SG.NOM hire.3SG.PST he.SG.ACC  
 ‘Kasia’s friend hired him.’

However, more often than not, a single case expresses more than just one basic function. For example, in Polish, apart from signalling the possessor, the genitive case can be used to mark a part of a whole (4a) or a goal (4b), among other possible meanings. Further, in some instances, a case may fill a role that is canonical for *another* case, like in (4c) where the subject figures in the genitive (*osób*), instead of the nominative (*osoba*) because the genitive is required by the modifier *kilka* (*a few*).

(4) a. Polish

Dodaj trochę mąki.  
 add.2SG.IMP some flour(F).SG.GEN  
 ‘Add some flour.’

b. Polish

Idziemy do kina.  
 go.1SG.PRS to cinema(N).SG.GEN  
 ‘We’re going to the cinema.’

c. Polish

Kilka osób prowadzi rozmowę.  
 a\_few.NOM person(F).PL.GEN lead(IPFV).1SG.PRS conversation(F).SG.ACC  
 ‘A few people are having a conversation.’

Importantly, while in all of the above examples the noun-head relationship is spatial, cases can also express relationships in other, more abstract domains, like emotion or time. In Slavic languages, for instance, the meaning of the genitive as ‘part of a whole’ applies also to periods of time (Janda and Hill, 2002):

(5) Pierwszego listopada.

first.SG.GEN November.SG.GEN  
 ‘1st of November’

**Gender** is another property of a noun, but in contrast to Case it is an *inherent lexical property* which does not change depending on the usage of a word. Gender of a noun manifests itself primarily via agreement with modifiers, predicates and other types of target (Spencer, 2002). For instance, the Polish adjective *głodny* (*hungry*) takes different forms when it modifies a feminine owl “głodna sowa”, a masculine snake “głodny wąż” or a neuter animal “głodne zwierzę”. Nouns themselves typically do not inflect for gender, although

| CASE         | FEMININE      | MASCULINE       | NEUTER              |
|--------------|---------------|-----------------|---------------------|
| NOMINATIVE   | głodna sowa   | głodny wąż      | głodne zwierzę      |
| GENITIVE     | głodnej sowy  | głodnego węża   | głodnego zwierzęcia |
| DATIVE       | głodej sowie  | głodnemu wężowi | głodnemu zwierzęciu |
| ACCUSATIVE   | głodną sowę   | głodnego węża   | głodne zwierzę      |
| INSTRUMENTAL | głodną sową   | głodnym wężem   | głodnym zwierzęciem |
| LOCATIVE     | głodnej sowie | głodnym wężu    | głodnym zwierzęciu  |
| VOCATIVE     | głodna sowo   | głodny wężu     | głodne zwierzę      |

**Table 2.4:** Polish inflection of three noun phrases, each with a head noun of a different gender.

some lexemes may have female and male versions and, in some languages, adjectives can be used as nouns denoting people, which reflect the sex of their referent through inflectional morphology; e.g., Polish adjective *chory/chora* (*ill*) and Russian equivalent *больной/больная* can be used as a noun to refer to a hospital patient (Spencer, 2002).

While nouns do not inflect for gender, their gender can have an effect on their **declension** – the changing of a noun’s form across different paradigm cells. A noun’s gender is often correlated with its **inflection class**, which determines the inflectional pattern in a noun’s paradigm (Ralli, 2002, Stump, 2015) – different inflectional allomorphs figure in the same paradigm slot across different inflection classes. This is demonstrated in Table 2.4 which displays Polish declension of three singular noun phrases with head nouns belonging to different genders. Again, note that number, case and gender are grouped together in a single morph.

**Animacy** is another lexical property of a noun and an inflectional feature of other POS. Prototypically, it expresses how alive the noun is, but grammatical animacy does not need to correspond to semantic animacy.

**Person** is a property marked on a verb which indicates whether the verb’s subject (or, in some languages, object or both subject and object) is the speaker (1st person), the addressee (2nd person) or a third party (3rd person) (Haspelmath and Sims, 2010).<sup>6</sup> See Table 2.1 for an example of how a Polish verb’s form changes with respect to person.

**Tense, aspect and mood** are properties of events commonly marked on verbs. Tense indicates the action’s “location in time” (Comrie, 1985) (past, present or future). Aspect expresses ways in which an action extends over time, e.g., whether the action has been completed (perfective) or uncompleted (imperfective). Mood expresses the speaker’s attitude towards an event; some examples include the imperative mood, used to signal

<sup>6</sup>Some languages have even further distinctions, e.g., the inclusive or exclusive distinction in non-singular forms (Bender, 2013).

commands (as in (4a)), indicative mood, reserved for objectively viewed events, and subjunctive which is used for non-realized events or emotions, opinions and wishes (among others). Since certain combinations of tense, mood and aspect are semantically implausible (e.g., a currently occurring finished action; i.e., present tense + perfective aspect), at times linguists combine the three features into a single inflectional category (Bickel and Nichols, 2013d, Haspelmath and Sims, 2010).

Tense, aspect and mood belong to the category of **inherent inflection** (Booij, 1993, 1996, Haspelmath and Sims, 2010) which is not “required by syntactic context, although it may have syntactic relevance” (Booij, 1996). Number marked on nouns also falls into this category and so do some grammatical cases like the locative, ablative or instrumental (Haspelmath and Sims, 2010). In contrast, **contextual inflection** is “dictated by syntax” (Booij, 1996). This category encompasses agreement markers on adjectives and verbs (including number and person), as well as the structural cases which are required by syntactic agreement or government such as the nominative or accusative. This thesis is concerned primarily with contextual inflection.

## 2.4 Inflectional morphology in language understanding

The goal of this section is to introduce the reader to the means in which Polish, Russian, Finnish and Estonian communicate the structure of transitive sentences. Topics discussed here are of direct relevance to the experimental portion of this thesis (Chapters 4 – 7) which assess and improve on the neural models’ ability to correctly recognize morphosyntax as the primary indicator of the core grammatical relations in Polish, Russian, Finnish and Estonian.

The section starts with an overview of the arsenal of signals that languages can employ to encode meaning. Next, it moves on to discuss the particulars of signals used in the four languages in question – Section 2.4.2 provides details of case and agreement in Polish and Russian. Section 2.4.3 does the same for Finnish and Estonian. While morphosyntax is the main mechanism of communicating argument structure in all four languages, at times morphology can be ambiguous. Sections 2.4.4 and 2.4.5 discuss (i) when this happens and (ii) what disambiguation strategies may be used in such instance. The section concludes with an overview of the competition model (Bates and MacWhinney, 1981, 1982) and of insights from the competition model studies on adult and child language processing in the relevant languages.

The above topics will all be referred back to and tied to topics from other sections of this chapter in Section 2.7 – “The unique challenge”.

### 2.4.1 Different signals of grammatical function

Recognizing how different words relate to one another in a sentence – i.e., the “who does what to whom” – is a central task in language understanding. This information can be signalled in a language through one or more of the following overt **coding strategies**: case marking, agreement and word order (Andrews, 2007, Bender, 2013, O’Shannessy, 2010, Rounds and Kanagy, 1998, Slobin and Bever, 1982). Structural case markings, when unambiguous (see Section 2.4.4), directly point out the syntactic role of a noun in a sentence, which is directly tied to that noun’s semantic role (Andrews, 2007, Bender, 2013). Verbal agreement markings are helpful in cases where the involved features have different values for the subject and the object. Finally, word order is a strong signal in a language in which it is relatively rigid, where an argument’s relative position with respect to the verb encodes that argument’s grammatical function.

Languages vary in terms of which strategies they employ and to what degree. Some languages are **configurational** (Haspelmath et al., 2001) and rely primarily on the position of words in a sentence to signal grammatical function. An example of such language is English. The majority of English declarative clauses follow the subject-verb-object (SVO) word order and so the argument preceding the verb is almost always the subject. Indeed, in most English sentences word order is an exclusive signal to grammatical function, as in (6a). At times, English does employ case and agreement markings but these are rarely available – the first is only marked on pronouns, while the latter, for 3rd person singular subjects (6b).

- (6) a. Daniela      chased      Mark.  
Daniela.SG chase.PST Mark.SG  
Daniela chased Mark.
- b. She              chases              him.  
she.SG.NOM chase.3SG.PRS he.SG.ACC  
She chases him.

In contrast, **case-marking** languages (Haspelmath et al., 2001), rely strongly on inflectional morphology, in particular case-marking, to signal subject and object relations. Many languages in this group exhibit **flexible word order**, allowing for all possible orderings of the subject (S), verb (V) and object (O) (Siewierska, 1998). This consequently prevents word order from encoding argument functions and frees it to encode discourse functions. Such flexibility is exhibited in the four languages studied in this thesis, which are all case-marking. Consider, for example the six Polish sentences from (7), which are translations of “Daniela chased Mark”. All of these sentences encode information about the same conceptual event, assigning identical semantic roles to Daniela and Mark. The only difference between them is pragmatic – it lies in the salience assigned to the participants

of the event and the event itself. For instance, (7b) puts the emphasis on Daniela as the subject, similar to the English “It was Daniela that chased Mark”. Sentence (7a) is stylistically and pragmatically neutral, since it exhibits the basic, most frequent word order in Polish – SVO.

- (7) a. Daniela                      gonila                      Marka.  
       Daniela(F).SG.NOM chase(IPFV).3SG.PST.F Mark(M).SG.ACC
- b. Marka                      gonila                      Daniela.  
       Mark(M).SG.ACC chase(IPFV).3SG.PST.F Daniela(F).SG.NOM
- c. Daniela                      Marka                      gonila.  
       Daniela(F).SG.NOM Mark(M).SG.ACC chase(IPFV).3SG.PST.F
- d. Marka                      Daniela                      gonila.  
       Mark(M).SG.ACC Daniela(F).SG.NOM chase(IPFV).3SG.PST.F
- e. Gonila                      Daniela                      Marka.  
       chase(IPFV).3SG.PST.F Daniela(F).SG.NOM Mark(M).SG.ACC
- f. Gonila                      Marka                      Daniela.  
       chase(IPFV).3SG.PST.F Mark(M).SG.ACC Daniela(F).SG.NOM

It is also important to note that the tendency of case-marking languages to exhibit flexible word order is only a statistical correlation, rather than a typological universal (Fedzechkina et al., 2017, Neeleman and Weerman, 2009, Siewierska, 1998). Rich inflectional marking does tend to go in hand with more flexibility in the word order and, similarly, rigidity of word order is often accompanied by weaker inflectional morphology. However, there are numerous languages that go against such tendencies. For example, Icelandic displays little word order variation despite having both case and agreement marking (Siewierska, 1998). Conversely, Papago and Abkhaz have no case marking but still allow for considerable word order variation (Siewierska, 1998).

## 2.4.2 Coding strategies of Polish and Russian

This section provides a description of the coding strategies of Polish and Russian. This information is essential for developing an understanding of what needs to be captured by neural syntactic parsers if they are to correctly model the grammars of those languages. I describe the most common, canonical marking of the core arguments, as well as the patterns of non-canonical marking. I focus on the nuances of argument marking in *transitive clauses*; i.e., clauses in which the head verb takes one subject argument and one object argument.

### 2.4.2.1 Case

**Subject** Both Polish and Russian exhibit *nominative-accusative alignment* (Comrie, 2013), marking subjects in intransitive clauses in the same way as subjects in transitive clauses. The subjects are marked almost exclusively with the nominative case, with only a few exceptions, which include situations where the subject is modified by a word that governs the case of a modified nominal. This happens primarily for subjects modified by quantifiers, the majority of which govern the genitive case, as in example (8).

(8) a. Polish

Kilka/Pięć kotów goni psa.  
a\_few/five.NOM cat(M).PL.GEN chase(IPFV).3SG.PRS dog(M).SG.ACC  
A few cats are chasing a dog.

b. Russian

Несколько/Пять кошек преследуют собаку.  
a\_few/five.NOM cat(F).PL.GEN chase(IPFV).3PL.PRS dog(F).SG.ACC  
A few cats are chasing a dog.

**Direct object** The canonical case for direct objects in both Polish and Russian is the accusative, but only in *affirmative* clauses. In negated clauses, Polish verbs *always* require the direct object to take the genitive case (9), while Russian verbs take either the accusative or genitive objects – the selection depends on factors such as definiteness (Kagan, 2020, pp. 189–222) (10).

(9) a. Polish

Wołam kotkę.  
call(IPFV).1SG.PRS cat(F).SG.ACC  
I'm calling the (female) cat.

b. Polish

Nie wołam kotki.  
NEG call(IPFV).1SG.PRS cat(F).SG.GEN  
I'm not calling the (female) cat.

(10) a. Russian

Я пил воду.  
I.NOM drink(IPFV).1SG.PST.M water(F).SG.ACC  
I (male) was drinking water.

b. Russian (Kagan, 2020, pp. 189–222)

Я не пил воду.  
I.NOM NEG drink(IPFV).1SG.PST.M water(F).SG.ACC  
I (male) wasn't drinking (the) water.

- c. Russian (Kagan, 2020, pp. 189–222)

Я не пил воды.  
I.NOM NEG drink(IPFV).1SG.PST.M water(F).SG.GEN  
I (male) wasn't drinking (any) water.

In some situations, the genitive can be marked on the direct object, regardless of the affirmation. One such phenomenon is the **intentional genitive** whereby certain intentional verbs take genitive objects also in affirmative clauses (Kagan, 2020, pp. 189–222). Some examples include Polish verbs *bronić* (*to defend*) (11a) and *dotykać* (*to touch*), and Russian *хотеть* (*to want*) and *заслуживать* (*to deserve*) (11b). Notably, for some verbs (or verb-object combinations) genitive object marking is obligatory, while for others it is optional (Kagan, 2020, pp. 189–222).

- (11) a. Polish

Bronię kotki/kotkę.  
defend(IPFV).1SG.PRS cat(F).SG.GEN/cat(F).SG.ACC  
I'm defending the (female) cat.

- b. Russian (Kagan, 2020, pp. 189–222)

Ты заслуживаешь медали/медаль.  
you.NOM deserve(IPFV).2SG.PRS medal(F).SG.GEN/ medal(F).SG.ACC  
You deserve a medal.

Another phenomenon whereby Polish and Russian direct objects appear in the genitive case is **partitive genitive**. Here, the genitive is used to signal an “indeterminate amount of the matter denoted by the noun” (Kagan, 2020, pp. 189–222) (12). This use of the genitive is typically optional, and the accusative marking can be used instead.

- (12) a. Kupiłam Ci jabłek.

buy(PFV).1SG.PST.F you.DAT apple(N).PL.GEN  
I (female) bought you (some) apples.

- b. Kupiłam Ci jabłka.

buy(PFV).1SG.PST.F you.DAT apple(N).PL.ACC  
I (female) bought you (the) apples.

- c. (Kagan, 2020, pp. 189–222)

Я купила тебе яблок.  
I.NOM buy(PFV).1SG.PST.F you.DAT apple(M).PL.GEN  
I (female) bought you (some) apples.

- d. (Kagan, 2020, pp. 189–222)

Я купила тебе яблоки.  
I.NOM buy(PFV).1SG.PST.F you.DAT apple(M).PL.ACC  
I (female) bought you (the) apples.

|      |   | SINGULAR       |                 |               | PLURAL           |                  |
|------|---|----------------|-----------------|---------------|------------------|------------------|
|      |   | MASCULINE      | FEMININE        | NEUTER        | VIRILE           | NONVIRILE        |
| PRS  | 1 |                | biegnę          |               |                  | biegniemy        |
|      | 2 |                | biegniesz       |               |                  | biegniecie       |
|      | 3 |                | biegnie         |               |                  | biegną           |
| PST  | 1 | biegłem        | biegłaś         | –             | biegliśmy        | biegłyśmy        |
|      | 2 | biegłeś        | biegłaś         | –             | biegliście       | biegłyście       |
|      | 3 | biegł          | biegła          | biegło        | biegli           | biegły           |
| FUT  | 1 | będę biegł     | będę biegła     | –             | będziemy biegli  | będziemy biegły  |
|      | 2 | będziesz biegł | będziesz biegła | –             | będziecie biegli | będziecie biegły |
|      | 3 | będzie biegł   | będzie biegła   | będzie biegło | będą biegli      | będą biegły      |
| COND | 1 | biegłbym       | biegłabyś       | –             | bieglibyśmy      | biegłybyśmy      |
|      | 2 | biegłbyś       | biegłabyś       | –             | bieglibyście     | biegłybyście     |
|      | 3 | biegłby        | biegłaby        | biegłoby      | biegliby         | biegłyby         |
| IMP  | 1 |                | niech biegnę    |               |                  | biegnijmy        |
|      | 2 |                | biegnij         |               |                  | biegnijcie       |
|      | 3 |                | niech biegnie   |               |                  | niech biegną     |
| INF  |   | biec/biegnąć   |                 |               |                  |                  |

**Table 2.5:** Polish paradigm of the verb BIEC (excluding participles and impersonal forms). PRS, PST, FUT, COND, IMP and INF stand for present tense, past tense, future tense, conditional mood, imperative mood and infinitive.

Finally, some verbs in Polish and Russian require direct objects in other cases, usually the instrumental or dative case. The first group for Polish includes *rzucić* (*to throw*) and *kierować*, while for Russian *управлять* (*to govern*) and *командовать* (*to command*). The verbs governing the dative include many interaction verbs (Blume, 1998, Haspelmath et al., 2001), e.g., Polish *odpowiadać* (*to answer*) or *gratulować* (*to congratulate*).

Some linguists tie the non-canonical case-marking on direct objects to lower degree of transitivity of the expressed event, as defined by Hopper and Thompson (1980) (Haspelmath et al., 2001, Kittilä, 2009). Accusative marking of patients is linked to the **prototypical transitive events** – those which involve a “controlling agent and a thoroughly affected patient” (Kittilä, 2009). Conversely, case frames involving other grammatical cases, such as the dative or instrumental, are associated with less transitive events; e.g., those in which “an indefinite direct object is only partially involved in or affected by the action” (Haspelmath et al., 2001).

#### 2.4.2.2 Agreement

In both Polish and Russian, verbs agree with the subjects in person, number and gender. While the first two features are always unambiguously marked, regardless of the tense, aspect and mood, this is not the case for gender. In Polish, for **imperfective verbs**, which correspond to ongoing or habitual actions (see *aspect* in Section 2.3.2), gender is

marked for the past tense, future tense and conditional mood (see the paradigm of BIEC (*to run*) in Table 2.5, as an example). For **perfective verbs**, which signify completed actions, gender is marked only for the past tense and conditional mood. It is not marked for future tense because perfective verbs have no present tense, and the future is marked with the present tense markings of imperfective verbs (imperfective present morphology = perfective future morphology). In Russian, gender markings appear only in the past tense, singular forms, for both imperfective and perfective verbs.

Notably, one exception of the above subject-verb agreement patterns are Polish clauses in which subjects are modified by a quantifier. This is exemplified in the Polish sentence (8a) where the verb *gonić* does not agree with the subject in terms of the number – the form *kotów* is the *plural* form of KOT (CAT) but the verb form *goni* is singular. This is a regular pattern; instead of agreeing with the subject, the verb ‘agrees with the quantifier’, which acts like a neuter, singular noun.

### 2.4.3 Coding strategies of Finnish and Estonian

This section mirrors the preceding section on coding strategies of the Slavic language pair, and has the same objectives. As before, I discuss both the canonical and non-canonical marking of the core arguments and focus exclusively on transitive clauses.

#### 2.4.3.1 Case

**Subject** Finnish and Estonian also follow the nominative-accusative pattern of alignment. In both languages, subjects are canonically marked with the nominative case. Similarly to the Slavic languages, subjects can also appear in another case – the partitive – when modified by a quantifier (or other modifier that governs the case). Notably, in certain circumstances (e.g., in existential clauses) the partitive can also mark unmodified subjects but this almost always happens for intransitive verbs (Hiietam, 2004, Ikola, 1971, Miljan and Cann, 2013, Sands and Campbell, 2001).

In Finnish, subject may also be coded with the genitive. It is the default for non-finite verb forms (Sands and Campbell, 2001); see (13) for an example. It can also appear in verbal constructions of obligation, in which Finnish equivalents of *should*, *must* and *ought to* combine with an infinitive verb (Sands and Campbell, 2001).

- (13) Finnish  
 Älkäämme antako pienen vähemmistön terrorisoida.  
 NEG.1PL.IMP allow.1PL.IMP small.SG.GEN minority.SG.GEN terrorize.INF-1  
 kaupungin hallintoa.  
 city.SG.GEN government.SG.PRT  
 Let us not allow a small minority to terrorize the city administration.

|                     | NOUN                            |                         | PRONOUN       |        |
|---------------------|---------------------------------|-------------------------|---------------|--------|
|                     | SINGULAR                        | PLURAL                  | SINGULAR      | PLURAL |
| FINNISH (FOX, YOU)  |                                 |                         |               |        |
| NOMINATIVE          | ket <u>tu</u>                   | ket <u>ut</u>           | sinä          | te     |
| ACCUSATIVE          | <del>ket<u>tu</u></del> , ketun | <del>ket<u>ut</u></del> | sin <u>ut</u> | teidät |
| GENITIVE            | ketun                           | kettujen                | hänen         | heidän |
| PARTITIVE           | kettua                          | kerruja                 | sinua         | tetä   |
| ESTONIAN (FOX, YOU) |                                 |                         |               |        |
| NOMINATIVE          | reba <b>ne</b>                  | rebase <b>d</b>         | sina          | teie   |
| GENITIVE            | rebase                          | rebaste                 | sinu          | teie   |
| PARTITIVE           | rebast                          | rebase <b>id</b>        | sind          | teid   |

**Table 2.6:** The Finnish and Estonian structural case inflection for the translations of the noun FOX and the pronoun YOU.

**Direct object** According to many linguists, Finnish and Estonian do not have a single accusative case for nominals – the accusative is only distinguished for Finnish personal pronouns and Finnish *kuka* (*who*) (Bielecki, 2009, Karlsson and Chesterman, 2008, Kiparsky, 2001, Miljan, 2009, Timberlake, 1974). The reason for this rejection of the accusative is the fact that neither of those languages have an independent accusative marker, not shared with other cases.<sup>7</sup> The typical accusative role of the direct object is fulfilled by either the nominative or genitive case, for the singular nouns, and by the nominative, for the plural nouns (see Table 2.6). Importantly, the Finnic UD treebanks used throughout this thesis also make the no-accusative assumption and never label the noun objects with the accusative case. I follow this convention in this thesis and do not use the accusative case in the Finnish glosses.

In both Finnish and Estonian, the singular object figures in the nominative case only in impersonal constructions, imperative constructions or in certain infinitival clauses (Hiietam, 2004, Miljan and Cann, 2013, Sands and Campbell, 2001, Timberlake, 1975). Further, in Finnish a singular object takes the nominative case only if the sentence has no overt subject in the nominative (Sands and Campbell, 2001). For instance, in (14a) the governing verb is imperative and does not have an overt nominative subject; hence, the object figures in the nominative case. In contrast, in (14b), although the verb *ostaa* figures in the infinitive form, it has a nominative subject, and so its object takes the genitive case (14b). In Estonian, the nominative case of the object in imperative clauses is not dependent on the absence of the nominative subject (Hiietam, 2004), which can result in morphosyntactically ambiguous sentences.

(14) a. Finnish (Sands and Campbell, 2001)

<sup>7</sup>Historically, in Finnic there was an independent accusative case for certain nouns, marked with an *-m* suffix. It later became homonymous with the genitive suffix *-n*, likely because of their phonological similarity (Anttila, 1972, p. 103, Baerman et al., 2005b, Lees, 2005). While Finnish maintained the *-n* marker for the accusative role, the *-m* was lost in Estonian.

Anna            Marjan            ostaa            auto.  
 let.2SG.IMP Marja.SG.GEN buy.INF-1 car.SG.NOM  
 Let Marja buy a car.

- b. Finnish, adapted from Sands and Campbell (2001)

Pauli            käskettiin    Marjan            ostaa            auton.  
 Pauli.SG.NOM order.3SG.PST Marja.SG.GEN buy.INF-1 car.SG.GEN  
 Pauli told Marja to buy a car.

The Finnic nominative/genitive direct object marking alternates with the partitive, and the selection of case depends on the “aspectual properties of the clause” (Kiparsky, 1998, Lees, 2005, Karlsson and Chesterman, 2008, Kagan, 2020, pp. 99–146). In perfective sentences, expressing a *completed action*, with the *whole object involved*, the nominative/genitive marking is used – objects marked in such way are often referred to as **total objects**. In all other situations, direct objects are marked with the partitive case. Notably the partitive is also used in negated clauses, which naturally express uncompleted actions. The partitive of negation in Finnic languages bears a strong resemblance to the genitive of negation in the Slavic languages (in particular those languages in which such marking is obligatory, like Polish) (Kagan, 2020, pp. 189–222).

Some Finnish and Estonian verbs naturally pair with the partitive case because they express an indefinite action (Miljan, 2009, Kagan, 2020, pp. 99–146). In Finnish, such verbs include *rakastaa* (*to love*), *katua* (*to regret*), *katsoa* (*to watch*) or *jatkaa* (*to continue*). Notably, as pointed out by Kittilä (2009), the partitive object marking of those verbs can also be tied to their lower prototypical transitivity (as defined by Hopper and Thompson (1980)). Kittilä also explains how some verbs which normally require the partitive can take total objects if the transitivity of the described event is increased. Consider, for example, the sentences in (15a) and (15b). In the first, the verb *rakastaa* (*to love*) expresses an indefinite action without a salient effect on its target – hence the partitive marking. In (15b), however, the adverb *kuoliaaksi* (*to death*) highlights the action’s direct effect on the patient, which in turn licences the genitive case on the object.

- (15) a. Finnish (Kittilä, 2009)

Mies            rakasti            koiraansa.  
 man.SG.NOM love.3SG.PST dog.POSS.PRT  
 The man loved his dog.

- b. Finnish (Kittilä, 2009)

Mies            rakasti            koiran            kuoliaaksi.  
 man.SG.NOM love.3SG.PST dog.POSS.GEN to\_\_death  
 The man loved his dog to death.

Finally, for some predicates, Finnish objects may also appear in other cases, such as the illative for verbs like *uskoa* (*to believe*) and *luotta* (*to trust*) (16) (Sands and Campbell, 2001).

- (16) a. Finnish (Sands and Campbell, 2001)

Minä luotan sinuun.  
I.NOM trust.1SG.PRS you.ILL  
I trust you.

- b. Finnish (Sands and Campbell, 2001)

Hän uskoo siihen.  
3SG.NOM believe.3SG.PRS it.ILL  
He/she believes it.

### 2.4.3.2 Agreement

In Finnish and Estonian, subject-verb agreement involves two features: person and number. Importantly, the agreement applies only to subjects in the nominative case – in the presence of a partitive subject, the verb always appears in the third person singular form (Koskinen, 1999, Lindström, 2017, Kagan, 2020, p. 257). Agreement cues are also missing in Estonian negated non-imperative clauses, where the negation is expressed by a construction involving a negative auxiliary verb *ei* and a lexical verb, both of which have lost all inflectional marking (Miestamo, 2011). Finnish negated clauses, which exhibit a similar negative construction, kept the inflectional marking on the negative auxiliary which “acts as the finite element of the clause, carrying person-number inflection” (Miestamo, 2011).<sup>8</sup>

### 2.4.4 Syncretism: where morphology fails syntax

As described in the preceding subsections, morphology serves an important role in Polish, Russian, Finnish and Estonian, as a primary indicator of argument structure. Nevertheless, in some instances it can be ambiguous. Case syncretism, discussed here, is a principal contributor to such ambiguity.

**Syncretism** refers to a situation where a single form of a lexeme can be tied to multiple different morphosyntactic descriptions (different paradigm slots). In that, syncretism “lets down the syntax” (Baerman et al., 2005a), because it prevents morphology from unambiguously communicating the function of a word in a sentence. Importantly, syncretism only applies to particular paradigms, rather than the whole case system (Miljan and Cann, 2013). Historically, the merging of forms within an inflectional paradigm has two potential sources. It can be a result of either (i) a blind phonological change, e.g., a loss of an affix

---

<sup>8</sup>The negative verb construction in Estonian and Finnish originated from the Proto-Uralic negative verb construction which, much like Finnish, marked inflectional categories on the negative auxiliary and not on the lexical verb (Miestamo, 2011).

|     | a. ‘table’ | b. ‘student (M)’ | c. ‘new.PL’ | d. ‘book’ | e. ‘mother’ | f. ‘forty’ |
|-----|------------|------------------|-------------|-----------|-------------|------------|
| NOM | stol       | student          | novye       | kniga     | mat’        | sorok      |
| ACC | stol       | studenta         | novye       | knigu     | mat’        | sorok      |
| GEN | stola      | studenta         | novyx       | knigi     | materi      | soroka     |
| LOC | stole      | studente         | novyx       | knige     | materi      | soroka     |
| DAT | stolu      | studentu         | novym       | knige     | materi      | soroka     |
| INS | stolom     | studentom        | novymi      | knigoj    | materju     | soroka     |

**Figure 2.4:** Patterns of syncretism in Russian nouns (Baerman et al., 2005b).

or (ii) a more complex morphosyntactic readjustment, such as a merger of two distinct cases into a single case (Baerman et al., 2005a).

A type of syncretism most relevant in the context of this thesis is **case syncretism**, where a single inflected form is associated with more than one case function. Notably, Slavic languages “achieve peak of variety and complexity” (Baerman et al., 2005b) in this category. A prominent example of this is Russian, whose complexity was at the center of one of the most influential works on case syncretism – Jakobson’s 1936 “Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutung der russischen Kasus” (“Contribution to the general theory of case: general meanings of the Russian cases”). The work of Jakobson directly tied Russian case syncretism to case semantics, posing it as a “reflection of the underlying network of semantic values which make up the case system” (Baerman et al., 2005b). Figure 2.4 (borrowed from Baerman et al. (2005b)) demonstrates the variety of regular syncretic patterns in Russian. All of these patterns are of importance to NLP, but only patterns in columns (a) and (b), which involve the core grammatical cases, are relevant for this thesis.

**Polish and Russian** both exhibit prominent syncretism of the core nominative and accusative cases, classified as syncretism of Type 1 by Baerman et al. (2005b). Table 2.7 presents patterns of accusative declension in Polish singular nouns. The only contexts<sup>9</sup> in which singular nouns do not exhibit core case syncretism are (i) feminine gender, for nouns with nominative endings *–a* or *–i* (most frequent suffixes for feminine nouns) and (ii) animate or human animacy, which applies mostly to animals and people. Notably, while for the latter context the accusative is not syncretic with the nominative, it is syncretic with the non-core genitive case. This type of syncretism, classified as Type 2 by Baerman et al. (2005b), has likely developed as a repair strategy, to restore the distinction between the nominative and the accusative (Baerman and Brown, 2013, Baerman et al., 2005b, Wade and Gillespie, 2011, p. 68). In Slavic languages, Type 2 syncretism is mostly restricted to animate nouns and common nouns used figuratively to denote people (as per Table 2.7),

<sup>9</sup>Here, I use the word *context* in the sense of Baerman et al. (2005b) who use it to refer to a set of morphosyntactic feature-value pairs.

| FEMININE NOUNS ENDING IN... |             | NEUTER NOUNS |             | MASCULINE NOUNS    |           |
|-----------------------------|-------------|--------------|-------------|--------------------|-----------|
| -a/-i                       | CONSONANT   |              |             | INANIMATE          | ANIMATE   |
| -ę                          | =nominative | =nominative  | =nominative | =nominative        | =genitive |
| NOMINATIVE → ACCUSATIVE     |             |              |             |                    |           |
| dziewczyna →<br>dziewczynę  | dłoń → dłoń | okno → okno  | stół → stół | chłopak → chłopaka |           |

**Table 2.7:** The accusative declension of Polish singular nouns, based on (Tokarski, 2001, p. 84).

| FEMININE NOUNS ENDING IN... |             | NEUTER NOUNS    |             | MASCULINE NOUNS   |           |
|-----------------------------|-------------|-----------------|-------------|-------------------|-----------|
| -a/-я                       | -ь          |                 |             | INANIMATE         | ANIMATE   |
| -y/ю                        | =nominative | =nominative     | =nominative | =nominative       | =genitive |
| NOMINATIVE → ACCUSATIVE     |             |                 |             |                   |           |
| мáска → мáску               | мáзь → мáзь | болóто → болóто | стол → стол | учени́к → ученикá |           |

**Table 2.8:** The accusative declension of Russian singular nouns, based on (Wade and Gillespie, 2011, p. 73).

although exceptions do exist; e.g., *robot* (NOM) → *robota* (ACC) and *pajacyk* (*puppet clown*) → *pajacyka* (Tokarski, 2001, pp. 84–85). For plural Polish nouns, the accusative has the same form as the nominative in most contexts, apart from the masculine gender + human animacy (e.g., Polish translations of *men*, *boys*, *gentlemen*, etc.) for which the accusative is syncretic with the genitive (syncretism of Type 2).

The Russian accusative declension pattern for singular nouns closely resembles that of Polish (see Table 2.8). Nominative-accusative syncretism applies to (i) neuter nouns, (ii) inanimate masculine nouns and (iii) feminine nouns ending with -ь. For plurals, it applies to all inanimate nouns; for animate nouns of *all genders* the accusative form is equal to the genitive form (Wade and Gillespie, 2011).<sup>10</sup> This contrasts with the more restrictive scope of Type 2 syncretism in Polish plural nouns.

In both Polish and Russian, there is a correlation between the literal animacy and the distinction between the nominative and accusative form. This is a common pattern among the world’s languages – arguments of the highest animacy in the animacy hierarchy (ranging from inanimate nouns to personal pronouns) are the most likely to have distinct forms for the accusative case (Baerman and Brown, 2013). Such special treatment of animate objects can be linked to the phenomenon of **differential object marking**, in which overt object marking depends on the inherent properties of the argument, such as animacy, definiteness or proper/common distinction (Baerman, 2009, Haspelmath et al., 2001, Sinnemäki, 2014).

<sup>10</sup>The accusative-genitive syncretism in plurals in Russian applies also to toys fashioned in the human form and plurals of common nouns used to refer to people, among others (Wade and Gillespie, 2011).

|            | SINGULAR |         |         |        | PLURAL |        |         |        |
|------------|----------|---------|---------|--------|--------|--------|---------|--------|
|            | MARTA    | CHILD   | BOOK    | LETTER | MARTA  | CHILD  | BOOK    | LETTER |
| NOMINATIVE | Marta    | dziecko | książka | list   | Marty  | dzieci | książki | listy  |
| ACCUSATIVE | Martę    | dziecko | książkę | list   | Marty  | dzieci | książki | listy  |

**Table 2.9:** Declension of the three Polish nouns from sentences in example (17).

**Estonian** also exhibits syncretism of the core grammatical cases, but to a smaller degree than Slavic languages – the syncretism only applies to a few inflection classes.<sup>11</sup> The cases involved vary depending on the inflection class; all combinations can occur: (i) nominative and genitive, (ii) nominative and partitive, (iii) genitive and partitive and (iv) nominative, genitive and partitive (Lees, 2005, Erelt, 2007, p. 33, Sutrop and Roomet, 2015, p. 45).

**Finnish** exhibits no core case syncretism – the nominative, genitive and partitive cases have distinct forms, across all 51 KOTUS (Institute for the Languages of Finland) noun inflection classes.<sup>12</sup> Note, however, that if one assumed the existence of the accusative case in Finnish nouns (see Section 2.4.3), then such accusative would always share its form with either the nominative or the genitive; i.e. the language would exhibit nominative-accusative and nominative-genitive syncretism (Kiparsky, 2001, Miljan and Cann, 2013, Sulkala and Karjalainen, 1992). The same applies to Estonian.

### 2.4.5 Disambiguation strategies

The syncretic patterns described in the previous subsection can give rise to morphosyntactic ambiguity of subject and object in transitive sentences. This happens when neither case nor agreement markings single out one of the verb’s arguments as the subject/object. Consider the four sentences in (17), for which I provide glosses with all possible interpretations of noun forms (as if they were out of context). Sentence (17d), where (i) the nominative and the accusative are syncretic for *both* arguments of the verb (see Table 2.9) and (ii) the verb agrees with both arguments, is the only ambiguous one of the four. (17a) is trivially unambiguous, due to the lack of case syncretism for either argument. In (17b) the form of the object is ambiguous, but the subject’s form is unique to the nominative; hence, no sentence ambiguity. In sentence (17c), both arguments of the verb have forms that could be either nominative or accusative. Nevertheless, that sentence also has only one possible reading because the verb agrees exclusively with the first noun.

<sup>11</sup>[https://en.wiktionary.org/wiki/Appendix:Estonian\\_nominal\\_inflection](https://en.wiktionary.org/wiki/Appendix:Estonian_nominal_inflection)

<sup>12</sup><https://kaino.kotus.fi/sanat/nykysuomi/taivutustyyppit.php>

- (17) a. Polish  
 Marta                      pisze                      książkę.  
 Marta(F).SG.NOM write(IPFV).3SG.PRS book(F).SG.ACC  
 Marta writes a book.
- b. Marta                      pisze                      list.  
 Marta(F).SG.NOM write(IPFV).3SG.PRS letter(M.INAN).SG.~~NOM~~/ACC  
 Marta writes a letter.
- c. Dziecko                      pisze                      książki.  
 child(N).SG.NOM/~~ACC~~ write(IPFV).3SG.PRS book(M.INAN).SG.~~NOM~~/ACC  
 A child writes books.
- d. Diecko                      pisze                      list.  
 child(N).SG.NOM/ACC write(IPFV).3SG.PRS letter(M.INAN).SG.NOM/ACC  
 A child writes a letter./A letter writes a child.

In the presence of morphosyntactic ambiguity, the sentence has more than one correct parse and additional strategies are required to disambiguate it. Such strategies are discussed in the remainder of this subsection.

#### 2.4.5.1 Word order freezing

One cross-linguistically common disambiguation strategy, often dubbed **word order freezing**, is to exceptionally assume the most frequent word order in a language when “subject and object are not distinguished by morphological means” (Jakobson, 1971, p. 585). In the four languages studied in this thesis, the dominant, most frequent order is subject-verb-object (SVO) (Bloom, 1999, Ehala, 2006, Kaiser, 2000, Siewierska, 1993, Weist, 1983). Consequently, when applied to the Polish ambiguous sentence from (17d), this strategy recognizes the noun preceding the verb – *dziecko* (*child*) – as the subject, and the noun which follows the verb – *list* (*letter*) – as the object.

#### 2.4.5.2 Lexical semantics

Morphosyntactically ambiguous subject and object can also be disambiguated by relying on the lexicosemantic properties of the verb and its arguments. Indeed, the contrast on the morphosyntactic level is often believed to arise due to the lack of contrast on the semantic level (Comrie, 1981, pp. 124–137, de Swart, 2006, Siewierska and Bakker, 2008, Tily, 2010). In the words of Tily (2010) “When [the roles of agent and patient] can be easily guessed from meaning alone, neither case nor fixed word order are necessary, and many languages allow case omission or freer word order in exactly those situations”.

Some nouns make more plausible subjects than others. The prototypical transitive structure involves an animate, definite subject, “strongly responsible for the state of the

affairs” (Bornkessel-Schlesewsky and Schlewsky, 2009) and a strongly affected object, lower in animacy and definiteness (Comrie, 1981, p. 128). Dowty (1991) famously defines the prototypical subject and object roles in transitive clauses as sets of **proto-agent** properties, which include volitional involvement in the event, sentience and causation of an event, and **proto-patient** properties, such as undergoing a change of state, being stationary and being causally affected by another participant. Both sets can act as semantic determinants of subject/objecthood. In Dowty’s words:

In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of proto-agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of proto-patient entailments will be lexicalized as the direct object. (Dowty, 1991, p. 576)

Following this reasoning, the most plausible reading of the Polish sentence (17d) involves an animate, volitional, sentient subject – *dziecko* (*child*) – and an inanimate, causally affected object – *list* (*letter*).

The semantics of the verb also plays an important role in argument disambiguation (Andrews, 2007, Long and Prat, 2008, McRae et al., 1997, Siewierska and Bakker, 2008, Tabossi et al., 1994). The majority of verbs favour certain arguments to others – a tendency termed **selectional preference** (Katz and Fodor, 1963). A verb’s selectional preference defines the types of nouns that typically fulfill its subject and object roles. For example, the Polish verb *aresztować* (*to arrest*) has *policja* (*police*), *FBI* etc., as likely subjects and *kryminaliści* (*criminals*), *złodzieje* (*thieves*), etc., as likely objects. Consequently, the most plausible reading of the morphosyntactically ambiguous sentence in (18) involves *policjantki* as the subject.

- (18) Polish
- |                                                   |                       |                        |
|---------------------------------------------------|-----------------------|------------------------|
| Policjantki                                       | aresztowały           | kryminalistki.         |
| police_woman(F).PL.NOM/ACC                        | arrest(PFV).3PL.PST.F | criminal(F).PL.NOM/ACC |
| The police women arrested the (female) criminals. |                       |                        |

The same applies to the following example with inanimate nouns, where the most likely subject is *sąd* (*court*):

- (19) Polish
- |                               |                      |                          |
|-------------------------------|----------------------|--------------------------|
| Wyrok                         | wydał                | sąd.                     |
| verdict(M.INAN).SG.NOM/ACC    | issue(PFV).3SG.PST.M | court(M.INAN).SG.NOM/ACC |
| The court issued the verdict. |                      |                          |

## 2.5 Human language processing and acquisition: the competition model

Having outlined the different signals employed by languages to communicate “who does what to whom”, I now give an overview of a theory of human language processing and acquisition based on the interaction and competition of those signals – **the competition model** (Bates and MacWhinney, 1981, 1982). The main purpose of this section is to provide broader context. A lot of the competition model studies focus on the same linguistic domain as this thesis – recognition of the sentential agent/subject and patient/object in transitive sentences. These studies cover a wide range of languages and place the emphasis on the cross-linguistic variability in the *relative importance* of different types of linguistic signals in comprehension and language acquisition. As such, they serve as an inspiration for the evaluation approach I propose in the later chapters (Chapter 3–Chapter 5) and set out expectations for what surface cues should and should not be relied on in computational language processing, if the models are to achieve human-like competence.

Note that none of the experiments from this thesis directly implement or make use of the competition model. For this reason, the overview provided here is brief; for more detailed description see e.g., MacWhinney (2001), which forms the basis for much of the following discussion. It is also important to mention that some linguists have criticised the competition model, i.a., for vague definitions of its key components and for disallowing both intermediate levels of linguistic structure and innate structural constraints on linguistic knowledge (Gibson, 1992). Nevertheless, the body of work tied to the model gives rise to interesting insights into the differences in human language processing across different languages, which are worth discussing here.

### 2.5.1 The competition model

The competition model “quantifies the ways in which distributional properties of the input control language learning and processing” (MacWhinney, 2001). The model defines language processing as detection and interpretation of a series of competing cues to argument structure, where a cue is defined as “any linguistic property that systematically co-occurs with a given function” (Krajewski and Lieven, 2014). To interpret the meaning of a sentence, hearers consider a range of morphological, syntactic, semantic, prosodic and pragmatic cues, based on which they compute a probabilistic value for each possible interpretation and choose one with the highest likelihood.

In each language, the cues vary in their *availability* and *reliability*. The first captures how often the cue is present and has a contrastive effect. For instance, in the context of identifying a subject, an agreement cue in Polish is present in every clause (see Sec-

|     | ENGLISH                    | ITALIAN                     | GERMAN                       |
|-----|----------------------------|-----------------------------|------------------------------|
| NNV | The eraser the pig chases. | La gomma il maialino bacia. | Die Gabel küsst die Sau.     |
| VNN | Licks the cow the goat.    | Lecca la mucca la cabra.    | Leckt die Kuh die Ziege.     |
| NNV | The dog grabs the pencil.  | Il cane afferra la matita.  | Die Maus ergreift die Lampe. |

**Table 2.10:** Example competition model stimuli used in MacWhinney et al. (1984) (Table 1).<sup>13</sup>

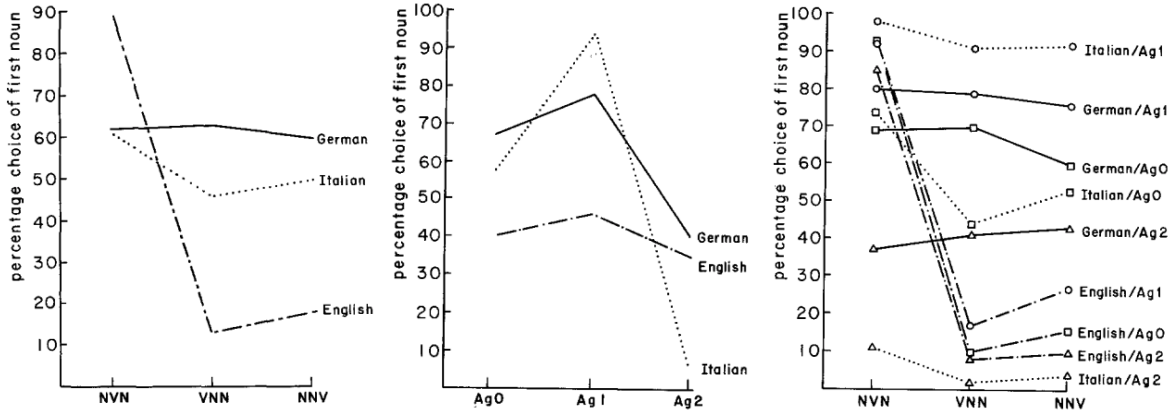


**Figure 2.5:** Agent choices in the study of MacWhinney et al. (1984), in the face of varied word order (left) agreement (middle) and both (right) for English, German and Italian. Ag0, Ag1 and Ag2 stand for ambiguous agreement, agreement with the first noun and agreement with the second noun, respectively.

tion 2.4.2.2). However, when the verb agrees with both arguments, agreement does not provide the contrast necessary to assign the subject role – i.e., it is not contrastively available. Reliability, on the other hand, captures how often the cue leads to the correct interpretation – cues that are never misleading are high in reliability. The combination of availability and reliability defines the *cue validity*, which in turn determines the *cue strength* in comprehension. The strongest cues play the most important role in comprehension and are the first to be acquired (MacWhinney, 2001, O’Shannessy, 2010). Notably, reliability plays a more important role than availability in determining validity and strength (MacWhinney, 2001, MacWhinney et al., 1984) – cues that are unreliable have the lowest validity, although they can be relied on in the absence of more valid cues.

Much empirical work on the competition model involves experiments in which human participants identify a sentential agent in simple transitive sentences, consisting of two nouns and a verb. The aim of those experiments is to (i) determine the relative ranking of cues, based on their strength in comprehension, and (ii) answer how the cue strengths interact. The sentences exhibit various combinations of cues, either competing or converging; see Table 2.10 for example stimuli from MacWhinney et al. (1984), where analysed cues involved word order, animacy, agreement and stress. The identification of the agent is interpreted as a reflection on the cue strengths in a language. Figure 2.5 shows the types of empirical data collected in such studies.

## 2.5.2 Insights into Slavic and Finnic languages

Much work on Slavic languages – primarily Polish, Russian, Czech and Serbo-Croatian – provides evidence for case marking being the most dominant cue for subject identification, outweighing word order and animacy (Kempe and MacWhinney, 1998, Lukavsky and Smolik, 2009, Slobin and Bever, 1982, Staroń and Kail, 2004, Weist, 1983). This is not surprising, given the 100% reliability of inflectional morphology cues (Kempe and MacWhinney, 1998); i.e., when morphology is unambiguous it always signals the correct interpretation. Some child language acquisition studies suggest that Slavic children start relying on this strategy from a very young age, as early as 3/4 years old (Janssen and Meir, 2019, Weist, 1983). However, those findings are inconclusive; e.g., in the study of Krajewski and Lieven (2014), Polish children of 8 years old were confused by sentences with novel verbs in which word order and case marking conflicted. At the earliest stages of language acquisition, Slavic children understand only sentences in which both word order and case marking cues are present and cooperate, signalling the same interpretation (Ibbotson and Tomasello, 2009, Krajewski and Lieven, 2014). In the words of Krajewski and Lieven, “[children’s] initial representations may resemble *gestalts* of several redundant cues” – only in the later stages of their development children start to respond to individual cues, presented independently. This special, prototypical status of redundantly marked sentences has also been shown for other languages, including English (Chan et al., 2009, Ibbotson and Tomasello, 2009, Krajewski and Lieven, 2014, MacWhinney and Bates, 1989).

There is much less body work on the competition model considering Finnic languages. However, the findings of Lemetyinen (2016) reveal that comprehension patterns for Finnish are similar to those for Slavic languages, with adults relying predominantly on case marking and youngest children relying on cue *gestalts* and gradually developing the ability to attend to isolated cues with age.

Finally, it is important to highlight that the native processing of Slavic and Finnic languages is inherently different from that of English. English speakers pay little attention to agreement and pronoun case marking, relying almost exclusively on word order as a signal to subjecthood (MacWhinney et al., 1984, Yoshimura and MacWhinney, 2010). The strong tendency to rely on word order manifests itself in English speakers’ interpretations of (ungrammatical) sentences, such as “The eraser push the dogs”, in which word order is in direct competition with animacy and morphological cues. In such sentences, English speakers universally select the first noun as the agent (MacWhinney, 2004, MacWhinney et al., 1984).

---

<sup>13</sup>The German sentences do not always match their English and Italian counterparts because MacWhinney et al. (1984) used a different noun set for German, with nouns ambiguous as to case.

## 2.6 Morphological typology

In this section, I discuss the similarities and differences among the morphological systems of Polish, Russian, Finnish and Estonian, guided by the typological classification of [Bickel and Nichols \(2007\)](#). The aim of this section is two-fold; first, it contributes to the depiction of the unique problem of modeling Slavic and Finnic languages (see Section 2.7). Second, the discussion of [Bickel and Nichols \(2007\)](#)’s typology provided here is relevant for Chapter 7 – it serves as the basis for the construction of the test artificial languages proposed there.

### 2.6.1 Traditional view of morphological typology

Morphological typology is concerned with the categorization of languages based on aspects of their morphological systems, such as their morphological complexity and the types of morphological patterns employed (see Section 2.2.4). As per [Arkadiev \(2020\)](#), the first approaches ([Schlegel, 1808](#), [von Humboldt, 1836](#), [von Schlegel, 1818](#)) classified languages into four broad categories, based on their primary means of encoding grammatical information. The languages were separated into: (i) **isolating**, with very little grammatical structure (e.g., Chinese), (ii) **analytic**, relying on independent grammatical words to signal grammatical information (e.g., English), (iii) **synthetic**, expressing most grammatical information morphologically and (iv) **polysynthetic**, with an extreme level of synthesis, allowing them to encode the meaning of a whole sentence within a single word. Within this traditional typology, synthetic languages could be further divided into **flexive** and **agglutinating**. The first express many different grammatical categories within a single formative which often fuses with the stem into forms which are hard to analyse. Within the latter, a single formative typically encodes only one category and changes shape only due to the language’s phonology (rather than the lexical identity of the stem – see Section 2.6.2.3).

But reducing the morphological diversity of a language to just a single variable has proven to be problematic. Linguists have found the discrete idealized types to be too restrictive and ill-defined, with many languages not fitting neatly into any category and some categories including languages which are very different from one another, e.g., Arabic and Latin ([Arkadiev et al., 2018](#), [Bickel and Nichols, 2013a](#), [Brown, 2010](#)). Even if the categories are viewed only as markers on a continuous scale, the one-dimensional categorization conflates many disparate typological variables which are not necessarily covariant ([Bickel and Nichols, 2007](#), [Plank, 1999](#)). Following this critique, morphological typology has shifted from categorizing languages into idealized types, towards more multidimensional approaches, which consider many variables and classify particular language constructions, rather than the whole languages ([Brown, 2010](#)). The typology of [Bickel and Nichols \(2007\)](#) is an exemplar of such elaborate morphological classification.

|                                                   | POLISH                                                   | RUSSIAN                 | FINNISH                 | ESTONIAN |
|---------------------------------------------------|----------------------------------------------------------|-------------------------|-------------------------|----------|
| FUSION, EXPONENCE, FLEXIVITY, SYNTHESIS           |                                                          |                         |                         |          |
| <b>Fusion</b> of case and tense-aspect-mood       | exclusively concatenative – formatives appended to stems |                         |                         |          |
| Case <b>exponence</b> in nouns                    | case+number expressed together in a single formative     |                         |                         |          |
| Case <b>exponence</b> in adjectives               | case+number+gender                                       |                         | case+number             |          |
| Tense-aspect-mood <b>exponence</b>                | monoexponential – a single formative exclusive for t+a+m |                         |                         |          |
| <b>Flexivity</b>                                  | flexive                                                  |                         | nonflexive              |          |
| <b>Synthesis</b> of the verb                      | –                                                        | 4-5 categories per word | 2-3 categories per word | –        |
| NOMINAL CATEGORIES                                |                                                          |                         |                         |          |
| Number of genders                                 | 3                                                        | 3                       | None                    |          |
| Number of animacy cat.                            | 3                                                        | 3                       | None                    |          |
| Number of person cat.                             | 3                                                        | 3                       | 3                       | 3        |
| Number of number cat.                             | 2                                                        | 2                       | 2                       | 2        |
| Number of cases*                                  | 7                                                        | 6                       | 14                      | 14       |
| OTHER                                             |                                                          |                         |                         |          |
| Suffixing vs prefixing                            | strongly suffixing                                       |                         |                         |          |
| <b>Locus</b> of direct object marking             | dependent marking                                        |                         |                         |          |
| Verbal person marking                             | only on the A argument                                   |                         |                         |          |
| <b>Syncretism</b> in verbal person/number marking | not syncretic                                            |                         |                         |          |
| Core case <b>syncretism</b> *                     | ✓                                                        | ✓                       | ✗                       | ✓        |

**Table 2.11:** Overview of selected morphosyntactic properties of Polish, Russian, Finnish and Estonian. Values of variables marked with \* reflect the assumption that there is no accusative case for nominals in Finnish and Estonian (see Section 2.4.3).

## 2.6.2 Typology of Bickel and Nichols

Three prominent variables involved in the Bickel and Nichols (2007) typology are:

1. **Fusion**; do affixes phonologically fuse with a stem?
2. **Exponence**; how much grammatical information is there in a formative?
3. **Flexivity**; do morphemes exhibit high degree of allomorphy?

#### 4. **Synthesis**; how many formatives are there in a word?

The first three variables are concerned with individual formatives – markers of inflectional features (see Section 2.2.3) – rather than the morphology of a whole language. Synthesis, on the other hand, is involved with the categorization of entire word forms. In the remainder of this subsection I discuss those four variables in more detail. As I do so, I also provide their values for Polish, Russian, Estonian and Finnish, which I extract from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), unless stated otherwise. Table 2.11 provides a quick insight into the differences between the four languages.

##### 2.6.2.1 Phonological fusion

Fusion refers to a degree to which affixes phonologically fuse with the stem; i.e., it is concerned with the types of morphological processes employed to form distinct words. The scale ranges from isolating to concatenative to nonlinear (Bickel and Nichols, 2013a). **Isolating** formatives are phonologically free (unbound), as the Fijian past tense marker *aa* (Bickel and Nichols, 2013a, Dixon, 1988):

(20) Boumaa Fijian (Dixon, 1988)

Au **aa** solia a=niu vei ira.  
1SG PST give.TR ART=coconut to 3PL  
I gave a coconut to them.

The **concatenative** formatives are phonologically bound to the stem and their combination with the stem often results in phonological adjustments. Finally, **nonlinear** formatives rely on non-concatenative morphological patterns (see Section 2.2.4) and realize grammatical information through the modification of the host stem. As a consequence, it is often not possible for the morphological meaning to be associated with a segmentable part of the word.

Importantly, a single language can rely on different morphological processes to mark different grammatical categories; e.g., the patterns of case marking may differ from the patterns of tense, aspect and mood (Bickel and Nichols, 2013a). Hence, while some languages can be classified as, for example, exclusively concatenative, other languages do not fall neatly into any of the categories. WALS typologizes languages based on two individual formatives: case and tense-aspect-mood (tense-like).<sup>14</sup> Both of those formatives are *concatenative* in all four languages studied in this thesis.

---

<sup>14</sup>Note that Bickel and Nichols treat tense, aspect and mood as one category. For the details of sampling those formatives see (Bickel and Nichols, 2013d).

### 2.6.2.2 Exponence

Exponence describes the semantic density of formatives; i.e., how many morphosyntactic features are realized within a single formative. The scale of exponence ranges from **cumulative** (or **polyexponential**) formatives, with many features cumulated into a single formative, to **separative** (or **monoexponential**) formatives that encode a single category. Bickel and Nichols (2013b) observe that *separative* formatives are the general default within the world's languages, but cumulative formatives are common among the Indo-European languages. For Polish and Russian, both noun and adjective formatives are cumulative – the first simultaneously encode both case and number features, the latter cumulate three features: case, number and gender. Finnish and Estonian similarly encode case and number in a single formative in both nouns and adjectives. All four languages encode tense, aspect and mood via an independent formative; i.e., it is not cumulated with agreement, voice or negation.

### 2.6.2.3 Flexivity

Flexivity is concerned with the allomorphy of the formatives and stems (see Section 2.2.3). **Flexive** formatives come in sets of allomorph variants. Importantly, the form variations are exclusively item-based – it is the lexical context that selects specific forms, e.g., the inflection class. In contrast, **nonflexive** formatives are invariant across the lexicon, although they may show variation, due to morphophonology or phonology (Bickel and Nichols, 2007). Importantly, the properties of fusion and flexivity are orthogonal with respect to one another; all possible combinations of their values are attested (Bickel and Nichols, 2007). The most common combination is *flexive concatenative*. According to the traditional typological classification (see Section 2.6.1), languages with such morphology would be classified as **fusional/flexive**. *Nonflexive concatenative* is yet another common combination; languages which exhibit this type of morphology constitute the prototypical **agglutinating** languages.

Both noun and verb formatives in Polish and Russian belong to the flexive concatenative class (see e.g., Table 2.1 and Table 2.4). Notably, verbs have more uniform suffixes than nouns, but have more morphologically heterogeneous stems (stems for nouns are relatively stable) (Bickel and Nichols, 2007, Timberlake, 2004). Formatives in Finnish and Estonian, on the other hand, are largely nonflexive – morphemes vary in shape primarily due to phonology or morphophonology (Bickel and Nichols, 2007, Ehala, 2009, Lemetyinen, 2016, Ruzsics et al., 2021).

### 2.6.2.4 Synthesis

Synthesis is a measure of semantic density. It closely resembles exponence, but in contrast

to exponence it is a category that applies at the word-level, rather than at the level of formatives. It is concerned with how much morphological and lexical information is bound together within *a word form*. The more formatives and lexical roots expressed within a word form, the more **synthetic** it is. Conversely, the less formatives and lexical roots the more **analytic** the form. The most synthetic words are **polysynthetic**. In the words of Bickel and Nichols (2007), polysynthesis “brings together not only formatives but also incorporated stems and lexical affixes into a single grammatical word”. For instance, a single word form in the Papua New Guinean Kewa language can convey the meaning of a whole English sentence (Bickel and Nichols, 2013c). Notably, the number of formatives per word form is correlated with the degree of flexivity within a language. In languages with flexive formatives, synthetic words typically include two or three formatives (Bickel and Nichols, 2007); e.g., Polish *wypije* (“I will drink and finish drinking”) contains two formatives which express five categories: tense, aspect, person and number. On the other hand, in languages with nonflexive concatenative morphology, synthetic words can have as many as ten formatives, as in the Turkish *tanıştırılmadıklarındandır* (“It is because they cannot be introduced to each other”) (Bickel and Nichols, 2007).

WALS measures synthesis of the verb by counting the inflectional categories in a maximally inflected verb form (Bickel and Nichols, 2013c). Tense, aspect and mood are treated as a single category. Similarly, agreement is counted as one category per role (e.g., subject agreement), irrespective of how many features are involved, and so are the “semantically related categories cumulated into one single inflectional slot or morpheme” (Bickel and Nichols, 2013c). Both Russian and Finnish have relatively synthetic verbal morphology, according to this metric, with Russian encoding maximally 4-5 categories per verb and Finnish 2-3 categories. WALS lacks explicit entries for Polish and Estonian. Synthesis information for nouns and adjectives is also missing. Importantly, English is situated towards the analytic end of the scale, which means that most world’s languages are more synthetic than English.

#### 2.6.2.5 Beyond fusion, exponence, flexivity and synthesis

The typology of Bickel and Nichols (2007) allows also for more fine-grained classification, e.g., based on the positioning of affixes within a word and types of syncretism. See Table 2.11 for values of such features for Polish, Russian, Finnish and Estonian.

## 2.7 The unique challenge

To summarise some of the information from the previous sections; Polish, Russian, Finnish and Estonian all have complex inflectional morphology and morphosyntax. This complexity stems from (i) the extent of information expressed through morphology, (ii) the means of

expressing it, and (iii) the variation in the marking of certain information. Among various types of information these languages express morphologically is the core sentence meaning – information about the subject and the object (see Section 2.4). Word order in those languages is flexible and does not encode the core grammatical relations. However, at times when subject and object cannot be distinguished through morphological means (and only then) the positioning of the two candidate nouns with respect to the verb can be used as a *disambiguation strategy*, along with the lexicosemantic information (Section 2.4.5).

All of the above characteristics pose a unique challenge for both development and evaluation of neural NLP models for tasks that rely on sentence meaning. I now illustrate this challenge with the Polish language, as an example.

### 2.7.1 Morphological competence – the Polish example

To detect “who does what to whom” in a Polish transitive clause models must be able to:

1. Recognize the verb and the candidate nouns in the clause.
2. Detect grammatical gender of the nouns, either based on their form or a mechanism resembling a lexical lookup.
3. Perform morphological analysis of the candidate nouns to detect their grammatical case.
4. Perform morphological analysis of the verb to detect the values of the person, number and gender (subject-verb agreement features).
5. Recognize case and agreement marking as the primary cue to grammatical function.
6. Recognize instances of non-canonical case marking (see Section 2.4.2).
7. Detect morphosyntactic ambiguity.
8. Make use of disambiguating strategies (i.e., rely on word order and lexical semantics) but *only* in the presence of morphosyntactic ambiguity.

**Table 2.12:** Steps towards recognising the core grammatical roles in Polish.

All of the above capabilities are what constitutes **morphological competence** in the context of this thesis. In the following paragraphs I define the three key difficulties that need to be overcome by a neural model to develop this linguistic skill.

**Morphological analysis** The first difficulty lies in morphological analysis, which requires the model to recognize just under 70 different noun suffixes (7 cases x 2 number classes x 5 gender+animacy classes) and differentiating between more than 75 verbal paradigm slots (3 person categories x 5 tense-mood categories x 5 gender+animacy classes + infinitive and impersonal forms). But it is not just the sheer number of paradigm slots that contributes to the overall challenge. The characteristics of Polish morphology also pose a problem; Polish exhibits high degree of fusion, and it is not always easy to identify individual morphs. It also has high degree of exponence, syncretism and allomorphy (see

Section 2.6), all of which results in difficult to learn many-to-many mappings of form to function (see Figure 2.6).

**Recognizing the relative importance of cues to meaning** The second difficulty is tied to recognizing case and agreement as the only reliable indicators of argument structure. This is not trivial, since the majority of transitive clauses in Polish follow the SVO order (Siewierska, 1993); i.e., there is a strong correlation between the syntactic role of a verb’s argument and its position in a clause. When acquiring Polish, children at first are only able to understand sentences in which word order and case markings cooperate, signalling the same interpretation (see Section 2.5). Understanding sentences with ambiguous morphology and those with non-canonical word order comes only later in development.

Further, studies on child language acquisition have noted the discrepancy between the productive use of case marking and the ability to recognize it as a cue to sentence meaning, with the latter coming later in children’s development (Krajewski and Lieven, 2014, Wittek and Tomasello, 2005). Consequently, even if the neural NLP models capture the case system of a language to the extent of (i) productively using case marking during language generation and (ii) correctly assigning higher probability scores to grammatical over ungrammatical noun forms across various contexts (as in work I discuss in Section 3.4.2), this *does not* imply the models are able to use case marking as a cue to grammatical function. I will come back to this discussion in the following chapter (Section 3.5).

**Knowing when and how to use disambiguation strategies** The last difficulty lies in recognizing word order and lexical semantic information as signals which can be relied on, but only in the presence of morphosyntactic ambiguity. This naturally requires the ability to detect morphosyntactic ambiguity, which is not trivial (see Section 2.4.5). Similarly non-trivial is developing the *conditional* dependence on word order and lexical semantics. Further, the need to occasionally rely on word order arguably makes it more difficult for the model to recognize morphology as the dominant cue to meaning (see the previous paragraph), especially given the aforementioned strong correlation between the subjecthood/objecthood and word order in Polish.

## 2.7.2 Russian, Finnish, Estonian

Nearly identical requirements to those from Section 2.7.1 apply to Russian, Finnish and Estonian – the primary difference lies in the morphological analysis. The morphology of Russian is very similar to that of Polish. Finnish and Estonian do not have grammatical gender and exhibit less allomorphy. In contrast to the Slavic language pair, their case marking is quite transparent – case suffixes are often unique to particular cases and

| Koń-<br>cówka | Przy-<br>padek | Licz-<br>ba | Ro-<br>dzaj | Grupa<br>znaczeniowa | Zakończenie fone-<br>tyczne tematu | Zakończenie<br>mianownika<br>liczby poje-<br>dynczej | Inne uwagi                                                            |
|---------------|----------------|-------------|-------------|----------------------|------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------|
| -a            | M              | poj.        | ż           | ---                  | ---                                | -a                                                   |                                                                       |
|               | D              | poj.        | m           | ---                  | ---                                | zero                                                 | zleksykalizowane                                                      |
|               |                |             | n           | ---                  | ---                                | -e, -o                                               |                                                                       |
|               | B              | poj.        | m           | żywotne              | ---                                | zero                                                 |                                                                       |
|               | MBW            | mn.         | n           | ---                  | ---                                | -e, -o                                               |                                                                       |
| -ach          | Ms             | mn.         | ---         | ---                  | ---                                | -a, -e, -i,<br>-o, zero                              |                                                                       |
| -ami          | N              | mn.         | ---         | ---                  | ---                                | -a, -e, -i,<br>-o, zero                              |                                                                       |
| -ą            | N              | poj.        | ż           | ---                  | ---                                | -a, -i, zero                                         |                                                                       |
| -e            | MBW            | poj.        | n           | ---                  | miękkie, stwardniałe               | -e                                                   |                                                                       |
|               | C              | poj.        | ż           | ---                  | twarde (łącznie z k,<br>g, ch)     | -a                                                   |                                                                       |
|               | Ms             | poj.        | ż           | ---                  | twarde (łącznie z k,<br>g, ch)     | -a                                                   |                                                                       |
|               |                |             | m           | ---                  | twarde bez k, g, ch                | zero                                                 |                                                                       |
|               |                |             | n           | ---                  | twarde bez k, g, ch                | -o                                                   |                                                                       |
|               | W              | poj.        | m           | ---                  | twarde bez k, g, ch                | zero                                                 | także nazwy<br>mężczyzn na<br>-ec : -eze                              |
|               | MW             | mn.         | m, ż        | ---                  | miękkie, stwardniałe               | -a, -i, zero                                         |                                                                       |
|               | B              | mn.         | m           | niesobowe            | miękkie, stwardniałe               | zero                                                 |                                                                       |
|               |                |             | ż           | ---                  | miękkie, stwardniałe               | -a, -i, zero                                         |                                                                       |
| -em           | N              | poj.        | m, n        | ---                  | ---                                | -e, -o, zero                                         |                                                                       |
| -ę            | B              | poj.        | ż           | ---                  | ---                                | -a, -i                                               | MBW lp. w rze-<br>czownikach na<br>-ę : -ęcia;<br>-ę : -enia          |
| -i            | D              | poj.        | ż           | ---                  | miękkie oraz k, g                  | -a, -i, zero                                         | także mianow-<br>nik lp. zleksy-<br>kalizowany                        |
|               | C Ms           | poj.        | ż           | ---                  | miękkie                            | -a, -i, zero                                         |                                                                       |
|               | W              | poj.        | ż           | ---                  | miękkie                            | -i, zero                                             |                                                                       |
|               | MBW            | mn.         | m           | niesobowe            | k, g                               | zero                                                 | leksykalnie tak-<br>że niektóre żeń-<br>skie o mianow.<br>lp. na zero |
| -i            | MW             | mn.         | m           | osobowe              | twarde (bez k, g,<br>ale z ch)     | zero                                                 |                                                                       |
|               | D              | mn.         | ż           | ---                  | miękkie                            | zero                                                 | zleksykalizo-<br>wane także<br>na -a                                  |
|               |                |             | m, n        | ---                  | miękkie                            | zero                                                 | zleksykalizo-<br>wane                                                 |

**Figure 2.6:** An overview of usecases for selected Polish suffixes (Tokarski, 2001, pp. 132–133). The first column lists all cases which can be encoded by each suffix (M: nominative, D: genitive, B: accusative, W: vocative, Ms: locative, N: instrumental, C: dative). The third column, lists numbers (poj. – single, mn. – plural). The fourth, genders (ż: feminine, m: masculine, n: neuter).

there is no interaction with gender. On the other hand, Finnish and Estonian have more grammatical cases and more elaborate patterns of non-canonical case marking, dependent i.a., on definiteness and aspectual information (see Section 2.4.3). The latter adds to the difficulty of Step 6 from Table 2.12.

### 2.7.3 English-centrism in NLP

While the first part of the challenge outlined in Section 2.7.1 – morphological analysis – is an active area of research in neural NLP (Kann et al., 2016, McCarthy et al., 2019a, Shen et al., 2016, Sorokin et al., 2017), the models are never directly tested for their use of morphology as a signal to meaning. No NLP literature has shown whether neural models can successfully perform steps 5–8 from Table 2.12 (see Section 3.5). Note that strong performance on NLP tasks, such as question answering or syntactic/semantic parsing cannot be counted as proof that the models are capable of performing those steps – the models might be getting “the right answers for the wrong reasons” and getting away with it because of flawed evaluation. For example, in Section 3.3, I demonstrate how the UD treebanks for Polish, Russian, Finnish and Estonian can be hacked with strategies that make no use of inflectional morphology.

This gap in the literature is the more important given that most neural NLP models are developed with English language in mind and later applied, without any adaptations, to other languages. As discussed in Section 2.4, languages vary widely in their means of encoding meaning. Importantly, those differences manifest themselves in sentences as simple as a *verb + two nouns*. The heavy reliance of English on word order as a “near-categorical cue to sentence interpretation” (MacWhinney et al., 1984) (see Section 2.5.2) is an exception, rather than a norm among the world’s languages. It remains unclear whether the architectures and optimization techniques developed on English are suited to processing of languages which encode meaning through inflectional morphology.

### 2.7.4 Conclusion

This section summarised the properties of Polish, Russian, Finnish and Estonian which make them an interesting case for study of neural models’ morphological competence. It also provided an overview of the unique challenge of modeling meaning in those languages – a challenge not yet directly addressed in the neural NLP literature.

Are the modern neural network architectures suited to extract the relevant information encoded in the word-forms of those languages and interpret it globally, in a broader clausal context? In the following chapter I argue that the research to date only partially answers this question and propose a set of experiments that can shed more light on the models’ capabilities.

---

# MORPHOLOGICAL COMPETENCE OF NEURAL MODELS

---

In this chapter I demonstrate that research to date is insufficient to answer the question posited in this thesis (and reiterated in Section 2.7.4) and discuss how my work fills this important gap in neural models’ evaluation.

I start the chapter by setting up the expectations for morphological competence of the mainstream NLP neural models in the context of their probabilistic, end-to-end nature and introducing some useful terminology (Section 3.1).

Next, through evidence emerging from related work (Section 3.2) and my own preliminary experimentation in the scope of dependency parsing (Section 3.3), I reveal that, while many modern neural models have the capacity to retrieve the most relevant morphosyntactic information, it is inconclusive whether they *rely* on it as a signal to meaning. These sections are followed by an overview of the most relevant related work, in which I highlight that no research to date has shown whether or not the models learn to rely on morphological signals (Section 3.4).

The chapter ends with a proposal of an experimental paradigm which fills the aforementioned gaps in evaluation of neural models, accompanied by a road-map for the rest of the thesis (Section 3.5).

Note that in this chapter I do not go into the details of any of the mentioned neural architectures (CNNs, LSTMs, transformers). To keep it concise, I place such overview in the appendix (Appendix C) where it is accompanied by a complementing overview of techniques used to allow these models to capture morphology – i.a., character-based embeddings, BPE, WORD-PIECES and UNIGRAM LM (Appendix D).

## 3.1 Laying out the terminology and the expectations

In the previous chapter I reviewed the most relevant linguistic properties of the investigated languages and argued that their characteristics pose a unique modeling challenge (Section 2.7). In this section, I expand on that discussion, grounding it in the characteristics of the mainstream neural NLP architectures. I also introduce some new terminology to describe properties of neural models relevant for this thesis which lack established terms.

### 3.1.1 The modern NLP paradigm

In recent decade, the NLP field has seen an emergence of a new paradigm which revolves around training deep and complex neural networks on raw, unprocessed text. The models are trained in a fully *end-to-end* fashion, without insertion of any linguistic knowledge. They act as **automatic feature extractors**, automatically learning the features and representations useful for a given task. Notably, the lack of linguistic knowledge applies also to morphosyntax. Instead of directly modeling morphology, modern neural models only *allow for morphology* by granting a model access to subword information. To capture subword patterns, models can operate directly on characters (Cao and Rei, 2016, Chung et al., 2017, Józefowicz et al., 2016, Kim et al., 2016, Peters et al., 2018, Pinter et al., 2017, Sutskever et al., 2011), or on subword chunks, such as character n-grams (Bojanowski et al., 2017a), word-pieces (Devlin et al., 2019b, Schuster and Nakajima, 2012, Wu et al., 2016) or BPE word segments (Gage, 1994, Sennrich et al., 2016) (see Appendix D for a comprehensive overview of techniques used to allow for morphology in neural NLP).

The above paradigm contrasts with the earlier statistical NLP approaches (1990s – 2010s) in important ways. Before the deep learning took over, the field revolved around *feature engineering*. NLP practitioners trained simple, shallow machine learning models on features they had pre-selected as the most predictive of the task. All these models were required to do was to generalise, via learning the complex interplay between the provided features (including those relating to morphosyntax) (Johnson, 2009, Socher and Manning, 2013). In contrast, nowadays, the models’ task is no longer limited to inducing feature interdependence in the context of a given task. They also need to learn what constitutes a viable feature in the first place. Both of these tasks happen in parallel within a model, which is unlike the earlier statistical approaches in which reasoning over the features naturally happens *after* their human-driven selection. Further, due to the neural models’ black-box characteristics, there is no easy way of inspecting what features they have inferred and whether these features bear a resemblance to those derived by (computational) linguists.

### 3.1.2 Signals, features and their relative importance

As discussed above, the modern deep neural networks do not have direct access to pre-selected *features*. What they have access to instead are **signals** in the input stream. Within this thesis, I define an (input stream) signal as the most basic, regular piece of information that a model can pick up on – for example, a particular affix, or a particular position of a word within the text stream. Such signals is what gives rise to **features** – pieces of information of higher level of abstraction, which are derived through interpreting and abstracting away from sets of signals attested in the training corpus. Examples of linguistic features include grammatical case (the induction of which requires a model to recognise grammatical case as a concept, grouping together different case markings in a meaningful way), relative position to a verb in a given sentence or, a more abstract, presence/absence of morphosyntactic ambiguity. Note how, while inducing the first two requires global awareness, they are actually **local features** since they apply to a particular token – it is the token that has a particular case marking or sits  $n$  slots to the left/right with respect to a verb. In contrast, morphosyntactic ambiguity is a **global feature** – it applies not to a particular token but to a whole clause. Further, it is a **higher-level feature** since it must be derived based on a range of **lower-level features**: case marking on all candidate nouns and agreement marking on the verb. I point out such differences here, since they are likely to have an effect on a model’s sensitivity to different features (see Section 3.1.4).

Given the above definitions, each feature can be tied to a specific, *minimal* set of signals it depends on. Case marking, agreement marking, morphosyntactic ambiguity are all based almost exclusively on morphological signals.<sup>1</sup> Conversely, features like the relative positioning of a word with respect to other tokens in a clause is based exclusively on word order signals. In this thesis, I focus not on the particular features the models might derive and make use of, but rather on the type of *signals* these features are based on. In particular, I am interested in examining the **relative importance** of three types of signals – morphological, word order and lexical (see Section 2.7 for the rationale behind this selection) – within the predictive processes of neural parsers. The relative importance of a (type of) signal is naturally tied to the features it gives rise to and how they are used and *ranked* within a trained model. Given a set of signal types one wants to investigate, their relative importance within a trained model can be estimated by placing the features tied to one signal type in direct opposition to features tied to other signal types – i.e., having features tied to different types of signals lead to contrasting predictions – and checking a model’s prediction. The type of signal with **higher importance** is the one associated with features aligned with the model’s output.

---

<sup>1</sup>Verbs’ valency also plays a role in identifying ambiguity, as I discuss in Section 4.1.1.

Finally, note that the features (and types of signals they depend on) outlined in this subsection are concepts set out by linguists based on a broad analysis of worlds’ languages. Neural monolingual models might (and likely do) derive a very different set of features, which do not cleanly map onto those relying on morphological signals, those relying on word order signals etc. Nevertheless, for the models to achieve close-to-human generalisation on language tasks, I expect them to derive features relatable to those proposed by linguists and based on the same (or very similar) signal types. After all, these are the features identified by experts as particularly predictive within the parsing task.

### 3.1.3 Morphology vs word order vs lexical semantics

In the context of the question posed in this thesis, if a model has access to subword information, it can in theory learn to induce features based on morphosyntactic signals (i.a., case marking, subject-verb agreement, morphosyntactic ambiguity) and center its generalisations on such features. However, the mainstream models are typically not constrained in any way that would force or predispose them to extract or use such information. Further, any limited training data supports multiple hypothesis that can lead to correct predictions. Some of those solutions might be based on other, non-morphological types of features. As I discuss in the later sections within this chapter, text data is often full of other types of information that the models can rely on instead, including linguistic correlates of core grammatical function (word order, lexical semantics) and other signals.

And even if one set of signals/features is the most optimal in the context of some training data, due to being particularly predictive or **valid** in the sense of [Bates and MacWhinney \(1981\)](#) (see Section 2.5), it does not mean a model will rank it as the most important. This is because different features have different characteristics – e.g., they can be local/global or low-level/high-level (see above). Signals they are based on also have different characteristics and can be differently accessed within a model. For example, transformer-based models (see Appendix C.4 for an overview) are fed information about tokens’ positions directly in an input embedding. In contrast, their access to morphs is less direct – information about each morph is buried in the embeddings of one or multiple word-pieces (see Appendix D.3). Consequently, a model needs to do more work to detect the latter, making it a more complex signal. Lexicosemantic signals are similarly complex since they are also likely to span multiple word-pieces. They are also less clean-cut – while one can write a set of categorical rules for word order and morphology, this is not possible for lexical semantics. It is to be expected that such feature/signal characteristics influence what forms the basis of a model’s predictions.

### 3.1.4 Relative input signal sensitivity

The question that naturally arises here is: what types of features is a model most likely to induce and rank as the most important if the data supports a range of varied solutions? And, importantly, what types of linguistic signals are these ‘preferred’ features based on?

What the model ends up relying on can be linked to its inherent property which I call **relative input signal sensitivity**, or **RISS**, for short. This property is tied to the model’s underlying architecture and it affects what types of signals form the basis of the most relevant/highest ranked features after it is trained. For example, **high sensitivity** to morphological signals means that the underlying architecture often gives rise to models which develop solutions based on features derived from morphological signals in the text stream. Conversely, high sensitivity to word order means that the underlying architecture gives rise to models that tend to rely heavily on word order based features. Note how the word *relative* highlights the importance of sensitivity-based ranking of different types of signals within an architecture. Importantly, this framing has many parallels to the competition model, which is concerned with the *relative strength* of linguistic cues in human language comprehension (see Section 2.5). The *strength* of different cues, as defined in the competition model, can thus serve as a guide for what relative (linguistic) signal sensitivity is required for an architecture to consistently develop *human-like* generalisations.

Importantly, the concept of RISS is separate from that of the relative importance of signals attested within a trained model (see Section 3.1.2). The former captures what the model is *predisposed to rely on* due to its underlying neural architecture. The latter captures what the model *actually relies on*, after it has been trained on a particular data and task. This separation of the two concepts allows us to disentangle the effects of the particulars of the underlying architecture and the effects of the training data. The relative importance of signals within a trained model can be different to its inherent sensitivities to different signals if the training data does not support the signals the model is most sensitive to; i.e., the signal preferences tied to a particular architecture may be *overridden* if the training data does not support the ‘preferred’ solution. In other words, if a neural architecture is particularly sensitive to a certain type of signal, it does not mean that models based on such architecture always base their solutions on such signals. Following the above distinction, while the *relative importance of signals* within a trained model can be tested directly – by checking what the model relies on (see Section 3.1.2) – this is not the case for a model’s *relative input signal sensitivity*. Instead, one can gauge the latter based on the analysis of which signals are supported in the training data *and* the attested relative importance of those signals, across different training datasets/tasks.

Note how the concept of RISS is very closely related to that of **inductive bias**, which can be defined as a range of factors which guide a model to settle on one solution over another (Battaglia et al., 2018, Haussler, 1988). Indeed, the varied sensitivities to different

types of input signals exhibited within a model can be viewed as a sub-type or a facet of its inductive bias. But while the concepts are closely related, in this thesis I avoid using the latter due to its vague definition and incredibly broad use – e.g., in [White and Cotterell \(2021\)](#) it is assumed to influence whether a language model performs better on languages with a particular word order, while in [McCoy et al. \(2020a\)](#) and [Warstadt et al. \(2020\)](#) it is assumed to influence whether neural models are more likely to adopt linear or hierarchical generalisations. Further, the term encapsulates not only a model’s architecture, inputs and training objective/procedure, but also training data ([Battaglia et al., 2018](#), [Bisk et al., 2015](#), [Helmbold and Long, 2015](#)).

Given all of the above terminology, the question posed in this thesis comes down to whether state-of-the-art neural models are sufficiently sensitive to the subword signals, so that they infer relevant morphosyntactic features, recognise how those features group together in a coherent way and rank them above alternative feature sets that might lead to good performance on a benchmark, but not on the underlying task (more on that in Sections 3.2–3.4). Throughout this thesis, I gauge the models’ sensitivities to morphology, word order and lexicosemantic signals by examining the relative importance of these types of signals within those models in the context of the types of signals supported by the data. I do so by placing these signal sets in opposition to one another in the evaluation data (while maintaining the grammaticality of a language) and inspecting models’ predictions (a technique mentioned in Section 3.1.2). All of this is described in more detail in the last section of this chapter, Section 3.5, which lays out a road-map for the rest of the thesis.

### 3.1.5 A categorical rule for morphology?

Finally, note that the requirement to rely on morphology as an indicator of core sentence meaning, as described in Section 2.4, is a categorical, ‘all-or-nothing’ rule. Such rules are not naturally supported by neural networks. Consequently, we cannot expect a neural NLP model to achieve 100% reliance on morphology (when unambiguous). But while 100% reliance might not be possible to achieve, a model that generalises in a human-like way should reach *close to* 100% reliance on morphology in unambiguous clauses.

It is also important to note that in reality, when day-to-day communication is taken into consideration, even morphology is not absolute. After all, people make grammatical mistakes (especially if they are not native speakers of a given language) and, despite such mistakes, their communicative intent often remains clear, given appropriate context. However, more often than not, an error in case or agreement marking in a case-marking language<sup>2</sup> results in an *ungrammatical* sentence. It is far less likely for an error to result in

---

<sup>2</sup>Here and in the rest of the thesis, I use the term *case-marking language* to refer to a language which strongly relies on inflectional morphology, in particular case-marking, to signal subject and object relations, as defined by [Haspelmath et al. \(2001\)](#).

a grammatical sentence, with correct morphological markers, signaling the non-intended interpretation – in unambiguous clauses with both arguments explicit, this would require making the ‘right kind’ of case marking errors for both subject *and* object and potentially also the agreement marking. This means that in grammatical, unambiguous sentences morphology is very close to being absolute; and these are the types of sentences I am concerned with within this thesis. Coming back to the errors; I suspect that the rise of ungrammaticality in the face of morphological errors provides a signal to the hearer which makes them likely to increase their reliance on the disambiguation strategies, similarly to how they increase their reliance on such strategies in the face of ambiguity (see Section 2.4.5), although I do not research this in this thesis.

## 3.2 Misleading success on relevant NLP benchmarks

A natural first step towards answering whether a given neural architecture can be leveraged to perform all the steps outlined in Table 2.12 (Section 2.7) is to examine its performance on NLP tasks which necessitate recognition of subject and object relations. A prime example of such task is syntactic parsing. In this section I argue that, while neural models achieve very strong performance on syntactic parsing, as manifested in UD shared tasks (Section 3.2.1), high parsing results alone are insufficient to draw conclusions regarding their morphological competence (Section 3.2.2).

### 3.2.1 Promising evidence from UD parsing

Let us take a look at the Polish, Russian, Finnish and Estonian results from the three most recent shared tasks on parsing into Universal Dependencies (UD) (see Section 2.1.1):

- (i) CoNLL 2018 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., 2018)
- (ii) IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies (Bouma et al., 2020), and
- (iii) the 2021 follow-up to IWPT 2020 Shared Task (Bouma et al., 2021)

In Table 3.1 I present F1-based labeled attachment scores (F1 LAS)<sup>3</sup> for some of the best performing models<sup>4</sup> – HIT-SCIR (Che et al., 2018) for CoNLL, TurkuNLP (Kanerva et al., 2020) and Orange (Heinecke, 2020) for IWPT 2020, RobertNLP (Grünwald et al., 2021) and COMBO (Klimaszewski and Wróblewska, 2021) for IWPT 2021. All three shared

---

<sup>3</sup>The shared tasks use a less-standard F1 LAS instead of the more commonly used accuracy-based LAS.

<sup>4</sup>Best performing models based on F1 LAS.

|                                          | Polish       |              |              | Russian      |              | Finnish      |              | Estonian     |   |
|------------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---|
|                                          | LFG          | PDB          | PUD          | ST.RUS       | TDT          | PUD          | EDT          | EWT          |   |
| F1 Labeled Attachment Score (F1 LAS)     |              |              |              |              |              |              |              |              |   |
| HIT-SCIR                                 | <b>94.86</b> | –            | –            | 92.48        | 88.73        | 90.23        | 85.35        | –            | – |
| TurkuNLP                                 | 88.54        | 92.76        | 92.44        | 92.74        | 92.62        | 92.34        | 87.14        | 79.73        |   |
| Orange                                   | 93.83        | 91.81        | 91.43        | 93.31        | 89.65        | 91.19        | 87.24        | 81.22        |   |
| RobertNLP                                | 88.23        | <b>94.54</b> | 93.84        | <b>94.64</b> | <b>93.77</b> | <b>93.53</b> | <b>89.52</b> | 80.50        |   |
| COMBO                                    | 94.09        | 92.64        | 92.27        | 94.29        | 92.33        | 92.21        | 88.73        | <b>83.12</b> |   |
| Morphological Tagging – Macro-average F1 |              |              |              |              |              |              |              |              |   |
| TurkuNLP                                 | 72.85        | <b>94.74</b> | 93.70        | 97.01        | 96.48        | 96.75        | 96.24        | 92.77        |   |
| Orange                                   | 92.76        | 92.41        | 93.12        | 97.76        | 96.50        | 97.22        | 96.97        | 93.82        |   |
| COMBO                                    | 90.55        | 93.72        | <b>94.19</b> | <b>98.04</b> | <b>97.34</b> | <b>97.33</b> | <b>96.97</b> | <b>95.11</b> |   |

**Table 3.1:** F1 LAS scores (top) and F1 morphological tagging scores (bottom) for the best systems of the CoNLL 2018, IWPT 2020 and IWPT 2021 shared tasks.

tasks also evaluate on the prediction of UD morphological features (see Section 2.1.1 and Appendix A.3) and I present these results for TurkuNLP, RobertNLP and COMBO below the F1 LAS results in Table 3.1. For all three models, morphological tagging is a part of their multi-task training objective.<sup>5</sup>

All submissions from Table 3.1 are based on the same biLSTM biaffine architecture (Dozat and Manning, 2017), which constructs a full dependency graph, based on binary arc predictions for each pair of tokens in a sentence (see Section 3.3.1 for a more detailed description). All submissions also involve a contextualised pre-trained embedding component. HIT-SCIR makes use of ELMO embeddings (Peters et al., 2018) – embeddings extracted from a multi-layer biLSTM language model, which takes as inputs character-level embeddings built with a convolutional neural network (charCNN). RobertNLP makes use of the XLM-Roberta (XLM-R) (Conneau et al., 2020) – a multilingual transformer model (100 languages) trained with a masked language modeling (MLM) objective, taking as input UNIGRAM LM (Kudo, 2018) word segments. Finally, TurkuNLP, Orange and COMBO, employ monolingual BERT models (Devlin et al., 2019a), but default to XLM-R on languages for which such model is not available (COMBO) or performs worse than XLM-R (Orange). These contextualised embeddings are supplemented with other types of

<sup>5</sup>I omit tagging results for HIT-SCIR and RobertNLP because the first made use of the baseline, provided by the task organisers and the second does not mention how the features are predicted in the paper.

|           | non pre-train.<br>token emb. | mono.<br>word2vec | mono.<br>ELMO | mono.<br>BERT | XLM-R | charRNN | charCNN |
|-----------|------------------------------|-------------------|---------------|---------------|-------|---------|---------|
| HIT-SCIR  | ✓                            | ✓                 | ✓             |               |       | ✓       |         |
| TurkuNLP  |                              |                   |               | ✓**           |       |         |         |
| Orange*** | ✓                            |                   |               | ✓*            | ✓*    | ✓       |         |
| RobertNLP |                              |                   |               |               | ✓     |         |         |
| COMBO     |                              |                   |               | ✓*            | ✓*    |         | ✓       |

**Table 3.2:** Types of input representations used by the models from Table 3.1. *mono.* stands for monolingual. \*Klimaszewski and Wróblewska (2021) and Heinecke (2020) use monolingual BERT only for selected languages and XLM-R for all others. For both models this selection involves Finnish, for COMBO it also involves Polish. \*\*For Polish Kanerva et al. (2020) employed Slavic BERT. \*\*\*Heinecke (2020) does not explicitly mention using non pre-trained embeddings and character-based embeddings, but they base their submission on Straka (2018) who make use of those embeddings.

word representations. The full breakdown of input types can be found in Table 3.2.

All models perform very well on the parsing task – for most benchmarks the F1 LAS is close to mid 90s. This is largely regardless of the employed inputs, although the transformer-based models outperform the ELMO-based HIT-SCIR for all but one benchmark. The high parsing performance is accompanied by even higher results on morphological tagging – for Russian and Finnish the highest scores are close to 100%. These results may seem very encouraging in the context of questions posed in this thesis – after all, morphological processing plays a central role in parsing the four languages. From the morphological tagging results, we also see that the models encode the relevant inflectional features, meeting the pre-requisites for using morphology as a cue to syntax (steps 1–4 in Table 2.12). However, as I demonstrate in Section 3.2.2, there is more to these results than meets the eye.

## 3.2.2 (But) models may not use morphosyntactic information

### 3.2.2.1 Shortcomings of standard NLP benchmarks

Given the results discussed in Section 3.2.1, it might be tempting to infer that the models must be morphologically competent. However, in making such inference one misses out an important lack in the assembled evidence. Based purely on the results from Table 3.1, it is impossible to tell whether the models *rely on* morphology to make their predictions – the causal link between morphological tagging and parsing is missing.

While it is true that morphological competence is a key prerequisite for dependency parsing of case-marking languages (see Section 2.7), UD benchmarks are only *approx-*

*mations* of the underlying task. By definition, approximations are not always accurate. Benchmarks may fail to capture the full complexity of the task and offer loopholes that models might exploit. The latter is especially true for benchmarks which draw evaluation data from the very same distribution as datasets used for training. This, regrettably, is the case for most standard benchmarks in neural NLP (Linzen, 2020).<sup>6</sup> In other words, most standard benchmarks open the door for the models to over-fit to their specific characteristics. Frequencies of certain words, frequencies of specific constructions, collocation of words – all of these features (and more) can be exploited by models to make predictions, regardless of their relevance for the task. And indeed, neural models do end up utilising such features.

English models performing superbly on *datasets* for challenging NLP tasks have recently been shown to fall short of some of the most basic linguistic abilities, such as recognising the difference between *most* and *least*, capturing the scope of negation and distinguishing between the subject and the object (Ribeiro et al., 2020). As a replacement for linguistic knowledge, the models utilise spurious associations with the target labels, hacking the datasets with various heuristics (McCoy et al., 2019, Poliak et al., 2018, Warstadt et al., 2020). For instance, in English natural language inference (NLI) task – which is to predict a type of relation (neutral, contradiction, entailment) holding between two sentences – neural models have been shown to pick up on the presence/absence of particular words (Naik et al., 2018), sentence length (Gururangan et al., 2018), as well as an overlap in lexemes and phrases (McCoy et al., 2019). These worrying tendencies seem to hold regardless of the underlying architecture. All three: LSTMs, transformers and tree-based models have been shown to be susceptible to such heuristics. And similar issues have been uncovered for other tasks, including paraphrase detection (Chaves and Richter, 2021, Gupta et al., 2021, Sinha et al., 2021), sentiment classification (Gupta et al., 2021, Sinha et al., 2021), argument reasoning comprehension (Niven and Kao, 2019) and dialog systems (Sankar et al., 2019), among others. And while, to the best of my knowledge, existing work in this scope considers only English, there is no reason to suspect models trained on other languages behave differently.

### 3.2.2.2 To encode $\neq$ to use

In previous subsection I stressed that good performance on standard benchmarks often does not reflect the true competence on a task – instead of solving the underlying task, models may simply be leveraging loopholes in the data. I now discuss that the latter happens even in cases where the model encodes the linguistic information prerequisite for developing the correct, human-like strategies. In other words, when both the task-appropriate, human-like

---

<sup>6</sup>See Linzen (2020) and Kuhnle (2020) for a more detailed (critical) discussion on the dominant evaluation paradigm in NLP.

strategy and the heuristic strategy are available to a model, its solutions tend to be based on the latter.

Let’s take as an example a linguistic ability of particular relevance to this thesis – differentiating between the subject and object relations. McCoy et al. (2019) shows that English LSTM and BERT-based NLI models fail dramatically – i.e., ‘reach’ performance of 0% – on test cases which specifically test for the recognition of (i) the subject and the object and (ii) the role of such roles in entailment. Ribeiro et al. (2020) reveals a similar issue for English BERT trained on question answering, although their model fails somewhat less dramatically – ‘only’ in 60.8% of cases. For BERT, this happens despite the model encoding substantial amount of syntactic structure in both its vector representations (Chrupała and Alishahi, 2019, Hewitt and Manning, 2019, Jawahar et al., 2019, Liu et al., 2019a, Tenney et al., 2019b) and its attention weights (Clark et al., 2019, Htut et al., 2019, Lin et al., 2019). Indeed, the aforementioned failures of BERT are startling given that it can predict features of the subject/object of the main clause (Krasnowska-Kieraś and Wróblewska, 2019), predict correct verbal agreement features (Goldberg, 2019), and recreate substantial portions of the underlying parse trees (Hewitt and Manning, 2019, Tenney et al., 2019a). And all that without any fine-tuning or direct structural supervision during pre-training.

Given that the relevant information is there, why don’t the models rely on it to make predictions? The jury is still out on this question; investigating models’ predispositions, often referred to as *inductive biases* (see Section 3.1.4), is an active area of English NLP research (Lin et al., 2019, McCoy et al., 2018, 2020a, Warstadt et al., 2020), very close in spirit to the topic and ideas explored throughout this thesis (more on that in Section 3.4.3).

Coming back to the results for UD parsing, research in related domains gives reasons to question the good results from Table 3.1. The models might not rely on appropriate linguistic signals to make these predictions and high performance on morphological tagging does not mean the models use that information for parsing. Indeed, in the next section I empirically demonstrate that UD benchmarks for Polish, Russian, Finnish and Estonian are susceptible to issues very similar to those I have discussed for English in this section.

### 3.3 Dependency parsing: a case study of benchmark shortcomings

As discussed in Section 2.7, morphological competence is a key pre-requisite for human-like competence in recognising core grammatical relations in Polish, Russian, Finnish and Estonian. This involves the ability to conduct morphological analysis, to recognise the relative importance of linguistic signals and to conditionally employ disambiguation strategies. In this section, I demonstrate that UD treebanks – a standard NLP benchmark

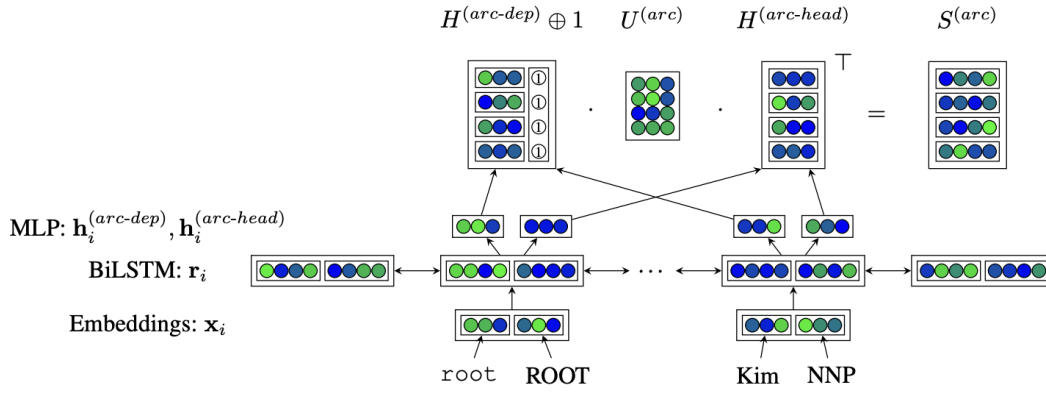


**Figure 3.1:** Architecture of Dozat and Manning (2017)’s parser. Figure 2 from Dozat and Manning (2017).

for dependency parsing – do not accurately capture this importance of morphology.

I reveal this through an experiment which gauges whether morphological competence is requisite for good performance on Polish, Russian, Finnish and Estonian UD treebanks. In this experiment, I strip the languages from all of their morphology and train **morphologically blind** baseline parsers, which I compare to models trained on unaltered language. Note how through eliminating morphology I deprive the baselines from the only signal that unambiguously encodes core grammatical roles (see Section 2.4). This allows me to gain insight into how much information about word relations is encoded via non-morphological means in the treebanks. In a follow-up experiment, I also train an alternative baseline on Polish UD data which I manually *decased*; this is done to pinpoint the importance of case-markings. The results from both experiments show that neural models can reach a surprisingly high performance on UD without access to the signal that is most predictive and reliable for the task.

Note that the method of analysis described above is reminiscent of word-perturbation techniques employed within English NLP to judge models’ sensitivity to English syntax (Abdou et al., 2022, Clouatre et al., 2021, Gupta et al., 2021, Hessel and Schofield, 2021, Papadimitriou et al., 2022, Sankar et al., 2019, Sinha et al., 2021). It is also related to Ravfogel et al.’s (2018) experiment on Basque, in which they remove case information from noun suffixes while experimenting on the agreement prediction task (Linzen et al., 2016a); although I find it important to point out that Ravfogel et al.’s motivation is opposite to mine, as the authors view case marking as a shortcut signal that makes the task easier.

### 3.3.1 Experimental setup

**Base model** As a base model I use the biaffine parser of Dozat and Manning (2017), which is the basis for most modern state-of-the-art parsers (see Section 3.2.1). Dozat and Manning’s model is a first-order graph-based parser (McDonald et al., 2005), trained

to find the highest scoring parse tree for a given input sentence. To make the search tractable, the model employs independence assumptions and decomposes each tree’s score into the sum of scores of its dependency arcs. The arc scores are computed based on the outputs of a multi-layer biLSTM (see Appendix C.3), which processes a sequence of token embedding + POS embedding concatenations. The biLSTM model acts as an automatic feature extractor, retrieving all information necessary to identify the arcs and the labels. For languages considered in this thesis, this includes all relevant subword features.

To obtain the arc scores, the outputs of the biLSTM are transformed (in parallel) by two different, dimensionality reducing Rectified Linear Unit (ReLU)<sup>7</sup> layers. This yields two representations per each token – one is the token’s representation as a head and the other as a dependent. Those representations are used to compute an arc score for each pair of tokens in a sentence, via a biaffine classifier (see Figure 3.1 for an illustration). Once the heads are predicted (by taking the highest scoring head at train time and with a minimum spanning tree algorithm at test time), another biaffine classifier is used to predict the labels, based on the biLSTM outputs. Just like for the arc prediction, the biLSTM vectors are first fed through ReLU layers before being used as inputs in the label classifier.

To get a more complete picture, I experiment with two different types of inputs: (i) fine-tuned BERT embeddings and (ii) embeddings constructed with a character-level convolutional neural network (CNN) (LeCun et al., 1989, 1990), followed by a highway network (Srivastava et al., 2015a), as in Kim et al. (2016). At the time of writing, the first, transformer-based type of input is a standard in neural NLP, forming the basis of most state-of-the-art results. The second type of input does not involve a pre-trained component, but gives the parser direct access to words’ characters and led to good results in previous work (Józefowicz et al., 2016, Peters et al., 2018). Note that beyond these inputs, which give the models access to subword information, the models are not optimised for morphological processing in any way.

I use a separate, monolingual BERT for each language – Polish Polbert (Kłeczek, 2020), Russian RuBERT (Kuratov and Arkhipov, 2019), Finnish FinBERT (Virtanen et al., 2019) and Estonian EstBERT (Tanvir et al., 2021). For each BERT I use its base, cased version. Lastly, in contrast to Dozat and Manning (2017), I do not use POS embeddings as an additional input to the model, since including *gold* POS in preliminary experimentation led to inconsistent and only minor improvements.

**Baselines** The baselines mirror the architecture of the base model. The most conservative baseline lemmatises all words in the input sentence before any further processing. This is done both at train and at test time. For this model, which I refer to as *baseline L*

---

<sup>7</sup> $f(x) = \max(0, x)$

(for *lemma*), I extract lemmata directly from UD annotations. For Polish, I additionally experiment with another baseline which applies a more targeted perturbation; instead of removing all morphology, it removes only the *case information* for the POS that inflect for case (see example (21)). To enable this model, which I call *baseline D* (for *decase*), I manually decased the Polish treebank, changing all cased forms to their nominative base, while maintaining the originally marked number, gender and animacy. This resulted in gold data from which one can draw reliable conclusions.

(21) Polish

ciężkich            książek             $\implies$  ciężkie            książki  
heavy.F.PL.GEN book(F).PL.GEN  $\implies$  heavy.F.PL.NOM/ACC book(F).PL.NOM/ACC

**Data** I select the largest available Universal Dependencies treebank for each language – Polish PDB (Wróblewska, 2018), Russian SynTagRus (Droganova et al., 2018), Finnish TDT (Haverinen et al., 2014) and Estonian EDT (Muischnek et al., 2014). The PDB treebank consists of 22,152 sentences, the SynTagRus treebank of 87,336 sentences, the TDT treebank of 15,136 sentences and the EDT of 30,973 sentences.

**Experimental details** I build the code within the AllenNLP (Gardner et al., 2018) framework and base the models on the AllenNLP’s implementation of the biaffine parser<sup>8</sup>. For the biaffine architecture, I use the same hyperparameter settings as Dozat and Manning (2017). For the CNN embeddings I follow the hyperparameters of Kim et al. (2016). I optimise the models with the Adam algorithm with dense and sparse gradients (Kingma and Ba, 2015), using a learning rate of  $1 \times 10^{-5}$  for BERT and  $1 \times 10^{-3}$  for the remaining parameters. During evaluation, for decoding I employ the Edmond’s algorithm (Chu, 1965, Edmonds, 1967), which finds a minimum spanning tree in a graph where words act as nodes and scored dependency arcs as edges.

For each [language, type of input, type of pre-processing] combination I train four models, each with a different seed (the total of 72 models). This was done to draw more reliable conclusions, accounting for the fact that different initialisation can lead to different generalisations (Liška et al., 2018, McCoy et al., 2020b). In the body of the thesis I report the average scores; for standard deviation scores see Appendix G (Tables G.1 and G.2).

### 3.3.2 Results

I present the (more standard) accuracy-based labeled and unlabeled attachment scores (LAS/UAS) on the development splits of the UD treebanks in Table 3.3, for the BERT-based models, and in Table 3.4, for the CNN-based models. For each language, I report a score for all relations (*all*) and scores for the core grammatical roles in transitive clauses

<sup>8</sup> Accessible within AllenNLP version 0.9.0.

| #                                    | Polish |       |      |       | Russian |       |      |       | Finnish |       |      | Estonian |       |       |
|--------------------------------------|--------|-------|------|-------|---------|-------|------|-------|---------|-------|------|----------|-------|-------|
|                                      | all    | nsubj | obj  | iobj  | all     | nsubj | obj  | iobj  | all     | nsubj | obj  | all      | nsubj | obj   |
| #                                    | 28.9K  | 365   | 293  | 73    | 95.7K   | 818   | 762  | 56    | 15.6K   | 148   | 148  | 37.0K    | 324   | 324   |
| LAS, Dozat and Manning (2017) + BERT |        |       |      |       |         |       |      |       |         |       |      |          |       |       |
| M                                    | 90.5   | 98.5  | 95.5 | 79.5  | 92.9    | 98.9  | 98.7 | 96.9  | 91.9    | 99.8  | 98.5 | 86.9     | 96.9  | 95.1  |
| L                                    | 87.0   | 89.9  | 91.3 | 66.4  | 90.6    | 93.9  | 93.0 | 80.3  | 85.2    | 87.8  | 90.5 | 81.0     | 78.1  | 75.2  |
| D                                    | 89.2   | 95.4  | 93.2 | 66.2  |         |       |      |       |         |       |      |          |       |       |
| M-L                                  | ↓3.5   | ↓8.6  | ↓4.2 | ↓13.0 | ↓2.3    | ↓5.0  | ↓5.7 | ↓16.5 | ↓6.7    | ↓11.8 | ↓7.9 | ↓5.8     | ↓18.8 | ↓19.9 |
| M-D                                  | ↓1.1   | ↓3.1  | ↓2.3 | ↓13.2 |         |       |      |       |         |       |      |          |       |       |
| UAS, Dozat and Manning (2017) + BERT |        |       |      |       |         |       |      |       |         |       |      |          |       |       |
| M                                    | 95.5   | 99.5  | 98.7 | 94.5  | 95.1    | 99.6  | 99.9 | 98.2  | 94.1    | 99.7  | 99.5 | 90.6     | 98.0  | 98.6  |
| L                                    | 93.0   | 94.7  | 98.4 | 91.8  | 93.5    | 96.4  | 98.4 | 94.2  | 89.6    | 95.3  | 96.0 | 87.1     | 92.4  | 95.8  |
| D                                    | 94.7   | 99.1  | 99.1 | 93.2  |         |       |      |       |         |       |      |          |       |       |
| M-L                                  | ↓2.5   | ↓4.8  | ↓0.3 | ↓2.7  | ↓1.6    | ↓3.2  | ↓1.5 | ↓4.0  | ↓4.6    | ↓4.4  | ↓3.5 | ↓3.5     | ↓5.6  | ↓2.8  |
| M-D                                  | ↓0.8   | ↓0.4  | ↑0.4 | ↓1.3  |         |       |      |       |         |       |      |          |       |       |

**Table 3.3:** LAS/UAS results for morphologically aware BERT-based model (M) and BERT-based baselines trained and tested on lemmatised language (L) and decased language (D).

with explicit subject and object.<sup>9</sup> The base parsers trained on unaltered language are marked with M, for morphology, parsers that process lemmatised language are marked with L, parsers that process decased language are marked with D.

**Baseline L** As per Table 3.3, results for BERT base models (M) mirror the success of models from Section 3.2.1. Their performance is particularly high for the core relations – LAS on *nsubj* is very close to 100 for all four languages. However, morphologically-blind equivalents of these models also reach high performance, often surpassing the performance of the CNN-based model with full access to morphology (!). Slavic L baselines, in particular, lose less than 4 LAS points across *all* relations, compared to the base model. And while they lose more points on the core relations, their scores remain around the low 90s for *nsubj* and *obj*. For Finnic languages the drop registered for BERT-based baselines L is more prominent; in particular for Estonian, where the difference between models M and L reaches 18.8 LAS and 19.9 LAS for *nsubj* and *obj*, respectively. Estonian is also the only language for which the performance of the morphologically-aware CNN-based model surpasses that of morphologically-blind BERT model. This is likely linked to the more varied word orders in the Estonian treebank, as I discuss in a subsequent chapter (Section 4.1).

For CNN inputs, the performance gaps between the base models and baselines are,

<sup>9</sup>Finnish and Estonian treebanks do not distinguish between indirect and direct object, so the scores for *iobj* relation are omitted for those languages.

|                                     | Polish |       |      |       | Russian |       |       |       | Finnish |       |       | Estonian |       |       |
|-------------------------------------|--------|-------|------|-------|---------|-------|-------|-------|---------|-------|-------|----------|-------|-------|
|                                     | all    | nsubj | obj  | iobj  | all     | nsubj | obj   | iobj  | all     | nsubj | obj   | all      | nsubj | obj   |
| LAS, Dozat and Manning (2017) + CNN |        |       |      |       |         |       |       |       |         |       |       |          |       |       |
| M                                   | 81.8   | 86.5  | 81.4 | 60.8  | 87.6    | 89.4  | 89.2  | 73.6  | 77.5    | 91.2  | 81.2  | 78.8     | 83.5  | 76.8  |
| L                                   | 78.4   | 77.8  | 77.4 | 50.3  | 79.9    | 80.0  | 77.8  | 46.0  | 68.9    | 68.1  | 56.9  | 71.8     | 60.5  | 61.0  |
| D                                   | 80.1   | 80.9  | 77.3 | 47.2  |         |       |       |       |         |       |       |          |       |       |
| M-L                                 | ↓3.4   | ↓8.7  | ↓4.0 | ↓10.4 | ↓7.7    | ↓9.4  | ↓11.4 | ↓27.7 | ↓8.6    | ↓23.2 | ↓24.3 | ↓7.1     | ↓23.0 | ↓15.9 |
| M-D                                 | ↓1.8   | ↓5.5  | ↓4.1 | ↓13.5 |         |       |       |       |         |       |       |          |       |       |
| UAS, Dozat and Manning (2017) + CNN |        |       |      |       |         |       |       |       |         |       |       |          |       |       |
| M                                   | 89.9   | 93.9  | 93.8 | 89.2  | 91.2    | 94.8  | 97.2  | 93.3  | 83.2    | 96.4  | 88.5  | 84.5     | 90.7  | 90.4  |
| L                                   | 87.3   | 87.3  | 91.0 | 85.3  | 85.8    | 87.9  | 92.5  | 88.8  | 77.8    | 89.6  | 84.0  | 81.0     | 85.0  | 89.8  |
| D                                   | 88.7   | 92.0  | 92.2 | 83.6  |         |       |       |       |         |       |       |          |       |       |
| M-L                                 | ↓2.6   | ↓6.6  | ↓2.8 | ↓3.9  | ↓5.4    | ↓6.9  | ↓4.6  | ↓4.5  | ↓5.4    | ↓6.9  | ↓4.5  | ↓3.4     | ↓5.7  | ↓0.6  |
| M-D                                 | ↓1.2   | ↓1.9  | ↓1.5 | ↓5.6  |         |       |       |       |         |       |       |          |       |       |

**Table 3.4:** LAS/UAS results for morphologically aware CNN-based model (M) and CNN-based baselines trained and tested on lemmatised language (L) and decased language (D).

in most cases, larger than those observed for BERT. This result suggests that the latter is better at leveraging alternative, non-morphological signals. Presumably, this stems from the inferior lexicosemantic capabilities of the CNN-based parsers – the models lack a pre-training component which would facilitate encoding of this kind of knowledge. Further, they are worse suited to memorising lexical information, since they have to build representations of words each time from scratch, based on their characters. In contrast, BERT operates on the level of word-pieces (see Appendix D.3) which better facilitates lexical memorisation.

Importantly, for both types of input the major performance drops happen at label assignment – drop noted for LAS is much larger than that on UAS. For example, the 19.9 LAS drop observed on object relations for BERT-based Estonian baseline L, corresponds to only 2.8 drop in UAS.

**Baseline D** The baseline D is more targeted – by removing only case information it directly targets the morphological encoding of core relations. It is therefore expected that the drop observed on *all* relations should be smaller than that for baseline L – and this is indeed the case for both BERT and CNN-based parsers. But for the core relations, I see similarly small performance decrease. This is despite the fact that the key relational encoding strategy within a language has been removed. This means that the models could be entirely blind to case and they would still perform very well on UD benchmarks.

### 3.3.3 Discussion

The perturbation applied for the baseline L is very forceful as it removes *all* morphological markings. This rids the model of the signal truly predictive not only of the core grammatical roles but also other relations, such as those between nouns and their modifiers. The fact that those models achieve overall performance close to that of models with direct access to morphology points to a serious shortcoming of UD treebanks when it comes to evaluating linguistic generalisations. And while the larger performance drop registered on the core relations is a good sign in this context, this drop is nowhere near as prominent as it would be if the benchmark targeted morphological competence prerequisite for the task.

Importantly, the strong performance of the baselines does not prove that the models do not rely on morphology when trained and evaluated on unaltered language. What they do prove, however, is that good performance on UD data is not enough to infer they do rely on morphology. The morphological signal seems to be largely redundant for neural models which can successfully rely on alternative correlates of grammatical function. Consequently, it is unclear whether parsers which are state of the art for UD data are able to (i) recognise morphology as the dominant cue, (ii) detect morphosyntactic ambiguity and (iii) conditionally rely on word order and lexical semantics, as disambiguating strategies.

To the best of my knowledge, the above provides the first account of the insufficiencies of the UD treebanks in evaluating correct linguistic generalizations. While morphological competence is a prerequisite for good performance on the dependency parsing *task* in Polish, Russian, Estonian and Finnish, it is not a prerequisite for high attachment scores in the UD *benchmark*. I find it important to note here that evaluation on UD is still extremely useful to gauge overall model performance; however, to tell whether the models “get the right answers for the right reasons” it should be supplemented with other types of evaluation, which directly target models’ linguistic abilities, as I discuss in the following Section 3.4.

## 3.4 Related research efforts: linguistically motivated analysis of case-marking languages

As discussed in Section 3.2 and demonstrated in Section 3.3, high performance on standard NLP benchmarks is often a defective indicator of the quality of the learned solution. In the words of Linzen (2020) “[the current] paradigm favors models that excel in capturing the statistical patterns of particular data sets over models that generalize as a human would”. This posits a need for a more capability-focused evaluation, which specifically targets models’ ability to develop human-like linguistic generalisations. The lack of such evaluation opens up the field to two serious issues. First, one cannot rely on standard

benchmarks to uncover the quality of the learned solution. Second, directing research efforts towards optimising for good performance on inadequate benchmarks can lead to no *real* progress in the field. Those risks are already pressing their stamp on neural NLP. Adding to the worrying findings discussed in Section 3.2.2, Glavaš and Vulić (2021) have recently shown that neural models which are better at capturing syntax are not rewarded on benchmarks for language understanding tasks, which “raises[s] concerns about whether we have made real progress on language understanding”, in the authors’ words. Talman and Chatzikyriakidis (2019) raise similar concerns by showing that models trained on one NLI dataset fail to perform well on another, even if it assumes the same notion of inference.

I am not alone in advocating for more targeted evaluation; this section echos the calls of Naik et al. (2018), Kuhnle (2020), Linzen (2020), Gauthier et al. (2020), (Hu et al., 2020), Rogers et al. (2020), Chaves and Richter (2021), Sinha et al. (2021) and many others. However, all these calls have been voiced in the context of *English* NLP. Here, I expand on what has been said before, by bringing to light issues specific to other languages and stressing that the capability-focused evaluation should be uniquely targeted to the modeled language.

Likewise, I am not the first to attempt evaluation of models’ capacity to develop linguistic generalisations. Recent research emerging from the *interpretability* and *robustness* tracks of neural NLP direct much effort towards testing such linguistic capabilities of neural models. Some of this work has already been discussed in Section 3.2.2. In the remainder of this subsection, I expand on that discussion by reviewing more relevant work. While most of the work within *linguistically-oriented deep net analysis* (LODNA) – a term coined by Baroni (2021) – is exclusive to English<sup>10</sup> some studies consider other languages, including those which signal meaning through morphology. I focus the following review primarily on such studies; for a more detailed overview of English work see e.g., Linzen and Baroni (2021), Baroni (2021), Belinkov (2021) and/or Belinkov and Glass (2019).

### 3.4.1 Diagnostic classifiers

One popular method of querying a model for linguistic information is to probe its representations with a diagnostic classifier (Conneau et al., 2018, Hahn and Baroni, 2019, Köhn, 2015, Qian et al., 2016). The classifier is trained to decode selected linguistic features from a collection of token representations. Importantly, the queried representations remain frozen (i.e., not tuned) throughout classifier training.

Insights emerging from such probing studies suggest that LSTM and transformer

---

<sup>10</sup>Only 12 out of 53 studies reviewed by the author while writing this chapter analysed monolingual models for languages other than English and only 4 of those papers analysed monolingual transformer models. 27/53 papers analysed solely English, 6/53 analysed translation models (mostly from/to English), 8/53 included other languages but experimented with multilingual models, instead of individual monolingual models.

models pre-trained with a (masked) language modeling objective can encode a non-trivial amount of morphosyntax and are able to recognise the subject and the object in a sentence, even without direct supervision during training. For instance, [Krasnowska-Kieraś and Wróblewska \(2019\)](#) show that simple diagnostic classifiers trained on top of frozen *sentence* representations from pre-trained Polish BERT can predict the number of the subject in the main clause with 91.5% accuracy, and the number of the object with the accuracy of 82.6%. For the biLSTM-based multilingual LASER model ([Artetxe and Schwenk, 2019](#)) the results are equally high – 93.8% for subject and 82.5% object. In an equivalent set of experiments on Finnish BERT, [Virtanen et al. \(2019\)](#) report 83.8% for number of the subject and 78.3% for the object. In a related line of work, [Edmiston \(2020\)](#) probes individual *token* representations extracted from monolingual BERT models for five languages, including Russian. They show that a simple linear classifier trained on top of frozen Russian BERT achieves 84% accuracy on the task of predicting morphological case, 86% on predicting gender and 92%, 99% on predicting number and person, respectively. [Edmiston \(2020\)](#) also inspect BERT’s attention mechanisms and identify a small number of attention heads which “hone in on subject-verb agreement” ([Edmiston, 2020](#)). [Eichler et al. \(2019\)](#) and [Şahin et al. \(2020\)](#) also probe *token* representations on morphosyntactic feature prediction, but focus on type-level predictions for dictionary entries and evaluate static (i.e., not contextualised) embeddings and LSTM-based ELMO ([Peters et al., 2018](#)), instead of BERT.<sup>11</sup> They report very high results for the latter on a range of languages and features; e.g., on number prediction, Finnish and Russian ELMO models reach 97.7% and 96.9% accuracy, respectively. On case prediction the accuracy is 96.7% for both languages. Another task aced by Russian and Finnish ELMO, as reported by [Şahin et al. \(2020\)](#), is detecting which morphosyntactic feature is shared by a pair of words (out of context). Here, the models achieve 98.4% (Finnish) and 97.6% (Russian) accuracy. Note that this task is a pre-requisite for modeling agreement.

Note that while this thesis is not focused on evaluating pre-trained language models *per se*, the above results are very encouraging. Pre-trained LSTMs and transformers are the building blocks of more composite, task-specific models directly trained to predict language structure; and, as shown by [Merchant et al. \(2020\)](#), they preserve much of the linguistic information they encode after the fine-tuning. And while some have raised concerns regarding whether the knowledge that resurfaces in probing experiments actually comes from the queried representation and not the diagnostic classifier ([Hewitt and Liang, 2019](#), [Pimentel et al., 2020](#), [Ravichander et al., 2021](#), [Sinha et al., 2021](#), [Voita and Titov, 2020](#), [Zhang and Bowman, 2018](#)), this is only an issue if one’s research goal is to isolate the contribution of the pre-trained model – a direction orthogonal to that of this thesis.

---

<sup>11</sup>Because of the type-level nature of their probing, [Şahin et al. \(2020\)](#) only experiment on forms which are unambiguous out of context.

Even if it is the probe that does most of the heavy lifting in probing studies reviewed here, their findings still constitute a positive evidence towards neural models’ capabilities of retrieving information encoded at the subword level.

What these studies do not answer, however, is whether the models *make use* of the morphosyntactic information to solve more elaborate tasks – neither of them test for this kind of causality.<sup>12</sup> Further, the more complex tasks, such as those that involve predicting the number of the subject/object in the main clause, are susceptible to the the same kinds of problems as those discussed in the earlier sections. Such concerns have occasionally been addressed in probing work (Hall Maudslay and Cotterell, 2021, Lasri et al., 2022, Lin et al., 2019), but all of that research concerns English.

### 3.4.2 Agreement and acceptability prediction

Beyond encoding a substantial amount of grammatical information (see Section 3.4.1), LSTM and transformer-based language models have been shown to exhibit a preference towards grammatical over ungrammatical language, as signalled by the differences in probabilities they assign to grammatical vs ungrammatical sequences. This result emerges from work that treats the models as if they were psycholinguistic subjects (Futrell and Levy, 2019, Futrell et al., 2019, Wilcox et al., 2020), evaluating them on targeted sentences designed to reveal their syntactic generalisations. For English, Linzen et al. (2016a), Marvin and Linzen (2018), Goldberg (2019) and van Schijndel et al. (2019) collectively show that both LSTMs and pre-trained transformers perform well on the **agreement prediction** task, which compares probabilities assigned by the model to plural vs singular verb forms in various contexts. In related work, Warstadt and Bowman (2019) test English GPT (Radford et al., 2018) and BERT models on the acceptability classification task (Warstadt et al., 2019) and reveal that across all constructions the models reach close to human performance<sup>13</sup> and even surpass humans on the simplest test cases.

Unfortunately, the agreement/acceptability work has scarcely been adapted to other monolingual models. Gulordava et al. (2018) is one of the few works that consider other languages for the agreement prediction task. Beyond English LSTMs they experiment on LSTMs trained on Italian, Hebrew and Russian and expand the task to include other types of agreement, e.g., between an adjective and a noun. Their Russian model achieves accuracy of 96.1% on such extended agreement task – the highest of all four languages. However, their Russian data excluded the subject-verb agreement, which is of primary

---

<sup>12</sup>This deficiency of diagnostic classifier probing is widely acknowledged by the probing community and researchers are actively working towards addressing these concerns (Elazar et al., 2021, Giulianelli et al., 2018, Ravfogel et al., 2021, Tucker et al., 2021). All of that work, however, is concentrated on English and does not consider problems specific to other languages.

<sup>13</sup>The better performing BERT reached 0.582 Matthews Correlation Coefficient (MCC), human performance was 0.697 MCC.

interest for this thesis. Further, their binary choice evaluation only involves form pairs that differ exclusively in the number feature, despite the fact that gender and case are also often involved in Russian agreement. Another work worth mentioning is that of Ravfogel et al. (2019) who experiment on agreement task for *synthetic versions* of English, constructed by manipulating word order, case systems and types of agreement. Ravfogel et al. report near 100% accuracy for LSTMs trained on languages with overt morphological case, for all three: flexible, VOS and OVS word orders. This leads them to conclude that overt case marking makes agreement task “significantly easier regardless of word order”. However, this is an overstatement, given that: (i) they only consider simple nonflexive concatenative morphology, with formatives that clearly stand out within English morphophonology (e.g., *kon*, *kar*, *ker*), and (ii) they only consider flexible order with uniform distribution of orders; i.e., they do not experiment with the harder instance of a language that has a dominant order but allows for other orderings, like the four languages studied within this thesis.

When it comes to transformers, to the best of my knowledge, there is no agreement/acceptability work that concerns monolingual models. Given the absence of such work, I review a few relevant results for multilingual BERT (Devlin et al., 2019a)<sup>14</sup> – as a proxy for monolingual models’ performance. In this track, Bacon and Regier (2019) show that mBERT performs very well on the agreement prediction task for four different types of agreement in Polish and Russian – ~90% accuracy – and reasonably well on the Finnish agreement prediction – ~75% accuracy. Mueller et al. (2020) similarly evaluate mBERT on a Russian agreement dataset – a translation of Marvin and Linzen (2018)’s benchmark – but report only 75% accuracy. Note that monolingual transformers have been shown to perform better than multilingual variants for a wide range of tasks and languages (Ács et al., 2021, Cañete et al., 2020, Chan et al., 2020, de Vries et al., 2019, Martin et al., 2020, Rust et al., 2021, Virtanen et al., 2019) so it is likely that monolingual models would perform much better on the agreement prediction task. It is also important to note that the capabilities and generalisations of multilingual models trained on various languages from different families are likely to be different to those of monolingual models.

Given the scarcity of existing work on case-marking languages, little can be concluded from this research track. And even if this was not the case, evaluation data used for agreement and acceptability tasks is also prone to contain artifacts (see previous sections). Consequently, this limits the types of conclusions that can be drawn from such research. This concern has been partially addressed by Gulordava et al. (2018) who proposed to disentangle the models’ performance from reliance on semantic and collocational information, but it has rarely been addressed in other studies. In the context of the above results, it is also important to note the discrepancy between the ability to generate

---

<sup>14</sup>See <https://github.com/google-research/bert/blob/master/multilingual.md> for more details on how the multilingual BERT was trained.

correct inflections during language production and recognising morphology as a signal to meaning in children, mentioned in Section 2.7. I mention this to emphasize that models’ morphological *awareness* and their ability to predict correct inflected forms in context do not imply that the models use morphology as a signal for sentence meaning.

### 3.4.3 Experiments with competing hypothesis

Another area of LODNA closely related to this thesis, involves studies which use the *poverty of the stimulus* (Chomsky, 1987) design where linguistic cues are placed in direct opposition to spurious surface cues (Lin et al., 2019, McCoy et al., 2018, 2020a, Warstadt et al., 2020). This is often done through carefully crafted datasets in which all (or almost all) examples in the train split support both the desirable and the undesirable generalisations, while the test examples disambiguate between the two. In this line of research, McCoy et al. (2020a) show that neither RNN nor transformer-based models exhibit true hierarchical bias, relying on rules based on linear order instead. Similarly, Warstadt et al. (2020) reveal that English transformer models, trained on no disambiguating examples, often have a tendency to rely on surface features over linguistic features.

Note that this line of work bears close resemblance to the problem studied in this thesis – due to the cooperation of lexical, word order and morphology signals in large portions of a language, neural models are rarely exposed to examples in which morphology stands in opposition to the other linguistic cues. Given that the encouragement to rely primarily on morphology is weak, as manifested in the data (see e.g., Section 4.1.2), the model can do well *on average*, after adapting either of the two generalisations – one in which it relies primarily on morphology (the correct generalisation) and another in which it over-relies on the combination of word order and lexical semantics (the incorrect generalisation).<sup>15</sup> Also note that all this occurs naturally within a language and, hence, is likely to be manifested in many different corpora. This is unlike the English studies of Warstadt et al. (2019), Lin et al. (2019) and McCoy et al. (2020a), who had to specifically construct datasets to simulate their poverty of the stimulus scenarios.

### 3.4.4 The remaining gap in the English-centric research

One serious issue with the non-English work discussed above is that it is sparse. Papers that consider case-marking languages are infrequent and inconclusive. But there is also another issue, which stems from the fact that much work on case-marking languages often closely mirrors what has already been done for English. Many of the datasets used for probing are translations of English datasets (Krasnowska-Kieraś and Wróblewska, 2019,

---

<sup>15</sup>Note that the incorrect generalisation may take various shapes and it does not necessarily exclude reliance on morphology. What is common for all of its instances, however, is the under-reliance on morphology which leads to incorrect parses in situations where the cues do not cooperate.

Virtanen et al., 2019) and, in agreement prediction track, non-English work either considers only the number feature when evaluating languages where more features are involved in agreement (Gulordava et al., 2018, Mueller et al., 2020) or applies evaluation which makes the task much easier for languages with larger verbal paradigms (Bacon and Regier, 2019).<sup>16</sup>

What emerges is a clear gap in the English-centric research. While existing research provides some evidence that the models are morphologically aware and the modern neural architectures are capable of *extracting* the relevant information, there is no research that sheds light on whether the models *use* morphology as a cue to subject/objecthood. Indeed, there is no research that even discusses the type of challenge discussed in Section 2.7, let alone attempts to address it.<sup>17</sup>

### 3.5 Targeted evaluation in the presence of conditional constraints: a proposal

To summarise previous sections, morphological tagging results (Section 3.2.1) and results emerging from LODNA work (Section 3.4) provide evidence that LSTM and transformer-based models have the capacity to retrieve much of the inflectional information that points out syntactic relations in case-marking languages. Coming back to the discussion from Section 2.7 (*‘The unique challenge’*), it appears that the models are largely capable of performing the first four steps towards recognising core grammatical roles outlined in Table 2.12. However, as attested in Section 3.2, Section 3.3 and Section 3.4, research to date provides no evidence for or against those models’ ability to perform the harder subsequent steps, which involve interpreting morphological signal in a broader clausal context. In this thesis I aim to fill this important gap by investigating whether neural models possess the following three abilities:

- (i) ability to recognise case marking and verbal agreement as primary signals to subject and object relations, ranking those signals as more important than word order and lexical semantics,
- (ii) ability to detect morphosyntactic ambiguity, and
- (iii) ability to conditionally rely on word order and lexical semantics in ambiguous clauses.

---

<sup>16</sup>In Bacon and Regier (2019) evaluation involves averaging across all possible forms in the verbal paradigm. This makes the task easier for languages with large paradigms; i.e., with many forms that clearly do not fit the context or are very rare. This is likely to result in the probability for the incorrect forms to be much lower, regardless of how well the model captures the agreement.

<sup>17</sup>The closest is perhaps the work of Tsarfaty et al. (2020) but they place the emphasis on difficulties associated with morphological analysis and disambiguation, rather than recognising the relative importance of competing signals.

The above selection is grounded in the linguistic characteristics of the languages investigated in this thesis. Accordingly, instead of inspecting correlations of target labels with *heuristic signals* unrelated to the task, as previous work has done for English (Gururangan et al., 2018, Naik et al., 2018, Poliak et al., 2018), I inspect reliance on *legitimate linguistic signals*, which rank lower than morphology but grow in importance in certain conditions.

### 3.5.1 Methodology and a road-map

I carry out my investigation through linguistically motivated, targeted experiments in the scope of dependency parsing. Dependency parsing naturally requires the models to identify core relations. It also offers a particularly generous setup for evaluation of relative importance of cues within a model, since the task requires little-to-no world knowledge and the models are directly trained to predict structure. At the same time, it is more elaborate than alternative syntax targeting tasks, such as agreement prediction, as it requires models to also learn other types of relations between words. This constitutes a more realistic scenario of how neural models are typically used in production. Within dependency parsing, I place primary focus on *transitive constructions* in which both subject and object are overtly marked. In Ibbotson and Tomasello (2009)’s words, “The transitive construction [...] is the earliest in which comprehension rests crucially on being able to successfully identify which participants are playing which roles in the event (who is doing what to whom)”. This makes it an ideal testbed for my investigations.

The rest of the thesis is separated into two parts. In the first part (PART I) which covers Chapters 4 and 5, I take a top-down approach and conduct experiments which simultaneously target all three investigated abilities. Throughout Chapter 4 and Chapter 5, I propose a novel counterfactual experimental paradigm, inspired by the competition model studies on human language processing (see Section 2.5). Within this paradigm I evaluate neural parsers on different counterfactual versions of existing dependency corpora in which one plausible linguistic signal (word order, lexical semantics, morphology) is placed in opposition to the other two. This allows me to compare a model’s reliance on morphology to its reliance on the alternative, secondary signals to meaning on which they should develop *only conditional* dependence. These experiments reveal concerning tendencies of both LSTM and transformer-based models to under-rely on morphology.

In the second part of the thesis (PART II), which consists of Chapters 6 and 7, I make use of the evaluation methodology proposed in PART I to investigate the particulars of the models’ shortcomings. More specifically I examine the effects of:

- training data and the training objective, testing whether the models can be successfully nudged towards the correct linguistic generalisation without any change to their architectures (Chapter 6), and

- morphosyntactic properties of *a language*, testing how the particulars of a morphological system can affect the relative importance of cues within a model (Chapter 7).

### 3.5.2 Final remarks

While my work is closely related to LODNA work on English (see Section 3.4), and in particular to the work discussed in Section 3.4.3, it differs from it in many important respects. First, it identifies and targets an understudied, unique problem of conditional linguistic constraints which applies to many case-marking languages. Second, unlike other work, which tests models’ reliance on heuristics, I test their reliance on linguistically relevant distractors, investigating whether they recognise the conditionally-determined importance of competing linguistic signals, relying on morphology when it is unambiguous and word order/lexical semantics otherwise. I hypothesise that such conditional constraints contribute to the difficulty of modeling sentence structure with neural architectures. Finally, it is important to note that word order and lexical tendencies are not merely an artifact of the data or a heuristic signal, but an inherent property of a language. The challenge I study here applies not only to the benchmarks I experiment with, but to modeling case-marking languages in general. Consequently, model shortcomings I reveal, have wide and important consequences for neural NLP field, as a whole.



## Part I

# Investigation of relative importance of linguistic signals



# RIGHT FOR THE WRONG REASONS: RELIANCE ON WORD ORDER IN NEURAL DEPENDENCY PARSING

---

In this chapter I inspect neural dependency parsers’ reliance on word order vs morphology as a signal to sentence meaning. I experiment on four case-marking languages: two Slavic (Polish, Russian) and two Finnic (Finnish, Estonian) and focus on the relations between the core elements of a transitive clause – subject (s), object (o) and verb (v). These relations determine the primary meaning of an utterance and are, arguably, the most important to get right. They are also a common focus in related psycholinguistics literature (see e.g., Section 2.5).

This chapter can be conceptually divided into two parts: (i) an analysis of the validity of word order and morphology signals in the *UD treebanks* and (ii) experiments which aim to reveal the relative importance of those signals exhibited in models trained on UD. The first part, treebank analysis, is contained in Section 4.1. Here, I investigate how much morphological ambiguity is there in the UD data and check the strength of word order’s correlation with grammatical function. This is done to capture how attractive both of these signals might appear to the models. Through this analysis, I reveal that, in large portions of the treebanks, word order and morphology cooperate, signaling the same interpretation – clauses are rarely morphosyntactically ambiguous and most of them follow the dominant, SVO word order. The low number of non-SVO clauses means that (i) models are only *weakly* encouraged to rely on morphology during training and (ii) models are not punished for over-relying on word order during evaluation. Given that the data largely supports both the incorrect, word order-based and the correct, morphology-based generalisations, it remains to be seen which of those two strategies is chosen by neural models.

This leads me to the second part of this chapter – an empirical investigation of word order vs morphology importance exhibited in neural parsers trained on UD (Sections 4.2–4.5). Here, I build on the intuition that if a model’s generalisations resemble those of native speakers, the following two conditions will be met:

**CONDITION I** The model’s performance will be invariant to the change of order of S, O and V in *inflectionally unambiguous* sentences.

**CONDITION II** The models’ performance on non-SVO orders will be noticeably lower than that on SVO sentences in *inflectionally ambiguous* sentences.<sup>1</sup>

If both of the above are met, I can infer that the models rank morphology as a more important, more reliable signal type than word order but, importantly, can backoff to the latter in the presence of ambiguity – just like people do (see Section 2.4). Following the above intuition, in Section 4.2 I propose a counterfactual experimental paradigm in which I manipulate the order of the core elements in a clause to study its effect on parsing performance. Specifically, given a dependency treebank, I construct six counterfactual versions – each with transitive sentences transformed to exhibit a different ordering of S, V and O – and use those new treebanks to evaluate the parsers. I create two versions of such counterfactual dataset. The first contains *only unambiguous sentences* and is constructed so that both generalisation strategies give different predictions: if a model adopts the correct, morphology-based strategy its performance will be invariant to the word order change. Conversely, if it learns to erroneously rely on the most common, canonical word order, altering that order will impair its performance. The second version of the dataset contains *only ambiguous sentences*; here, if a model recognises that word order can serve as a disambiguation strategy in the presence of ambiguity, performance will likely drop on non-SVO clauses.

The description of the new dataset is followed by the experimental portion of this chapter – covering Sections 4.3–4.5. For each language, I train and evaluate four different parsers – all based on the state-of-the-art architecture of Dozat and Manning (2017), but taking different inputs (see Section 4.3 for the finer experimental details). In this chapter, I train the parsers only with the dependency parsing objective – this is to answer what linguistic signals form the basis of the models’ solutions when they are not encouraged to pay attention to morphology via other objectives. In Section 4.4, I discuss the results of evaluation which targets **CONDITION I**; this is based on *unambiguous* version of my counterfactual dataset. Next, in Section 4.5, I discuss evaluation which targets **CONDITION II** and leverages the *ambiguous* split of the counterfactual dataset.

---

<sup>1</sup>Note that this drop does not have to be very prominent, given the lexicosemantic signal pointing towards a different interpretation. However, unlike for the unambiguous sentences where there should be close to no performance differences, for ambiguous sentences a performance drop is expected.

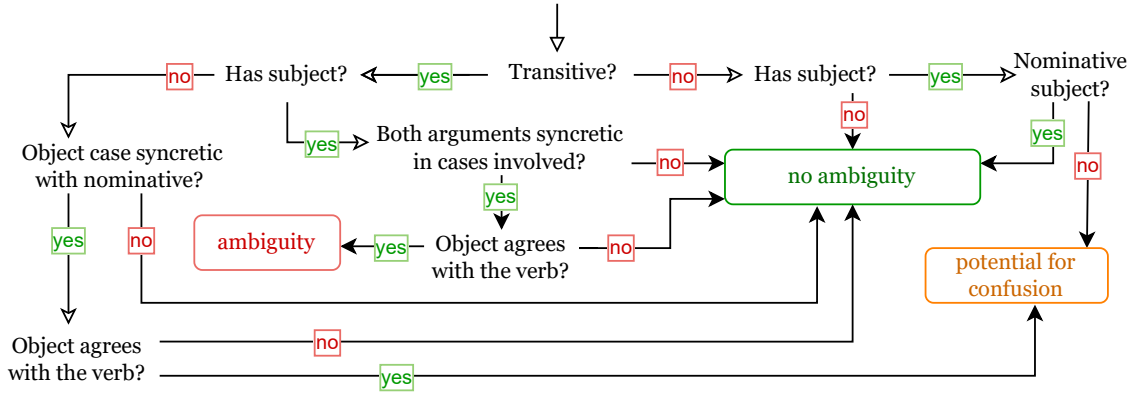


**Figure 4.1:** A flowchart demonstrating the steps towards identifying the ambiguity class. The distribution of those classes in UD treebanks is plotted in Figure 4.2.

Importantly, the main contribution of this chapter is the new evaluation methodology and the new counterfactual dataset. As I demonstrate in this chapter, these allow me to reveal a lot of interesting tendencies within the models’ generalisations, but an exhaustive investigation lies outside the scope of this thesis.

## 4.1 Analysis of signal validity in UD

In this section I examine the UD data to estimate the validity (as defined by [Bates and MacWhinney \(1981\)](#); see Section 2.5) of inflectional morphology and word order signals for core grammatical function. I do so in two subsections – each devoted to a different signal. First, in Section 4.1.1, I examine the distribution of different levels of ambiguity in inflectional morphology. This is to confirm that morphology is a widely available signal in the UD data and that the models are not actively discouraged from relying on it (as would be the case if morphological ambiguity was a common occurrence in the treebank). Next, in Section 4.1.2, I move my focus towards word order – a plausible linguistic cue on which the models should develop *only conditional* dependence. Here, I inspect the distribution of different word orders in the UD treebanks for Polish, Russian, Finnish and Estonian, to gauge the strength of word order correlation with grammatical function in the data.

For my analysis, I select the largest available treebank for each language: Polish PDB ([Wróblewska, 2018](#)), Russian SynTagRus ([Droganova et al., 2018](#)), Finnish TDT ([Haverinen et al., 2014](#)) and Estonian EDT ([Muischnek et al., 2014](#)) (see Section 3.3.1 for sentence counts).

### 4.1.1 Inflectional morphology

As discussed in Section 2.4, inflectional morphology is the main mechanism of communicating argument structure in case-marking languages. Unlike any other linguistic cue, it

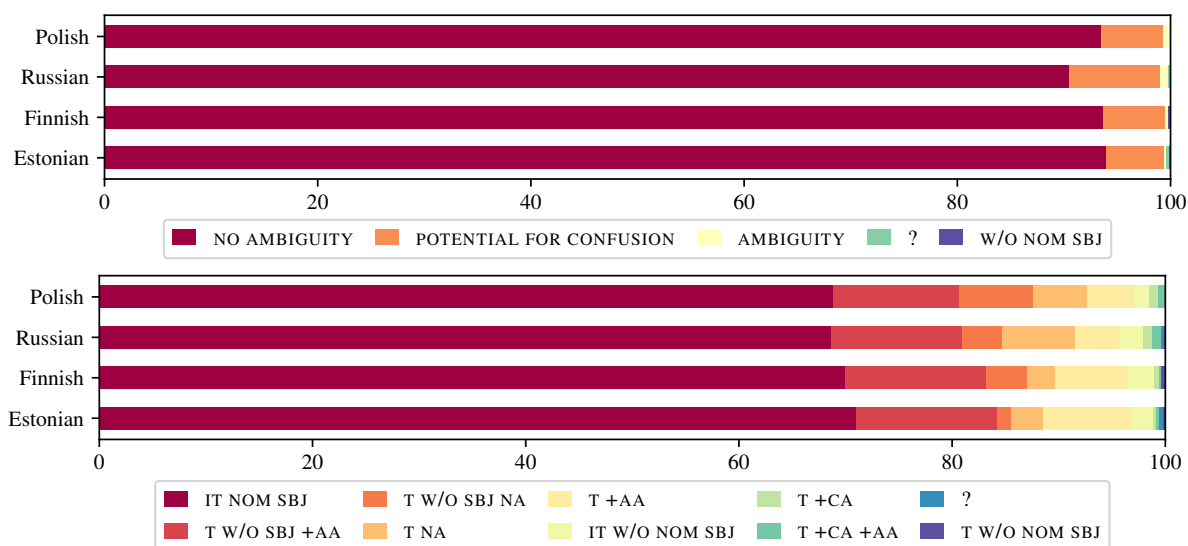


**Figure 4.2:** The distribution of inflectional ambiguity classes in UD treebanks, for coarse classes (top) and fine-grained classes (bottom). T, IT, +CA, +AA and NA stand for transitive, intransitive, case ambiguity, agreement ambiguity and no ambiguity. ? represents the *unknown* class.

has nearly 100% reliability<sup>2</sup> as a signal to core sentence meaning; i.e., when unambiguous, it always leads to the correct interpretation. The problem with morphology, however, is that it is not always available as a cue – at times morphosyntax is ambiguous (see Section 2.4.5). In this section, I ask how often this occurs in the UD treebanks for the languages in question; i.e., how often is morphology not available?

I estimate availability of morphological cues based on the distribution of clauses within different categories of ambiguity. Specifically, I define three different ambiguity classes:

- (i) **full ambiguity** – multiple interpretations of a clause are possible
- (ii) **no ambiguity** – morphosyntax unquestionably points out the core roles in a clause
- (iii) **potential for confusion** – morphology is ambiguous but a verb takes only one argument (object or subject) so the verb’s valency<sup>3</sup> is sufficient to unambiguously parse the sentence

Given a clause I identify the appropriate category by following the steps from Figure 4.1. To detect case syncretism and ambiguous agreement instances I make use of the Polish and English splits of Wiktionary.<sup>4</sup> If Wiktionary lacks lexemes needed to identify the appropriate category, I assign a clause to a supplementary *unknown* category, which I mark with a question mark in all figures.

Distributions of these three ambiguity categories, which I present in the top of Figure 4.2, prove morphology to be an unequivocally available signal, for each language considered.

<sup>2</sup>Even morphology is not absolute – some sentences contain errors.

<sup>3</sup>A verb’s valency defines the number and type of arguments it takes.

<sup>4</sup><https://en.wiktionary.org>, <https://pl.wiktionary.org>

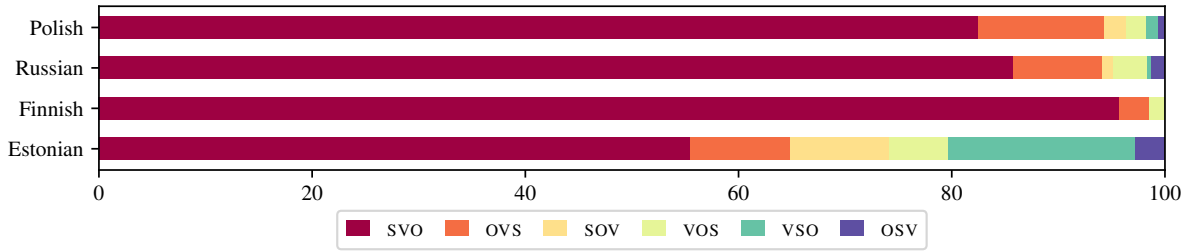


**Figure 4.3:** The distribution of ambiguous clauses within different word orders.

|         | SVO | OVS | SOV | OSV | VSO | VOS |          | SVO | OVS | SOV | OSV | VSO | VOS |
|---------|-----|-----|-----|-----|-----|-----|----------|-----|-----|-----|-----|-----|-----|
| Polish  | 6.7 | 8.1 | 1.8 | 1.4 | 2.0 | 1.1 | Finnish  | 2.0 | 0.9 | 0.0 | 25  | 0.0 | 0.0 |
| Russian | 6.5 | 4.5 | 0.9 | 6.2 | 2.2 | 1.1 | Estonian | 1.9 | 2.0 | 1.8 | 2.5 | 2.1 | 1.0 |

**Table 4.1:** % of ambiguous clauses given the word order. Note that 25% for Finnish OSV is not a reliable estimate given that TDT treebank has 3 OSV clauses in total.

In all four languages, morphology is unambiguous in more than 90% of clauses and truly ambiguous in less than 2% of clauses. The remaining clauses belong to the *potential for confusion* class, where verb valency information suffices to identify the correct role assignment. To get a more complete picture of the types of signals present in the data, I also consider a more fine-grained split between ambiguity categories, which provides information about the type of evidence for the lack/presence of ambiguity. The distribution of these categories, which correspond to the black-headed arrows from the flow-chart in Figure 4.1, is displayed in the bottom of Figure 4.2. It reveals that the treebanks are dominated by intransitive verbs with a canonically marked, nominative subject (IT NOM SBJ category) and transitive verbs with canonically marked object (T w/o SBJ +AA, T +AA, T NA). These categories account for 90%+ of each treebank. To get more clarity, I also check which orders are most likely to exhibit ambiguity (Table 4.1) and which word orders are most frequent, given ambiguity (Figure 4.3). It appears that ambiguity is most likely to appear in clauses in which the verb separates the two arguments; i.e., the SVO and OVS clauses. These word orders also dominate the ambiguous clauses present in the data; most prominently for Polish, Russian and Finnish treebanks, which very rarely expose the models to ambiguity in non-SVO/OVS clauses (<6%).

### 4.1.2 Word order

While all four languages allow for each of the six orderings of S, V and O, the majority of their clauses follow the canonical SVO order (Bloom, 1999, Ehala, 2006, Kaiser, 2000, Siewierska, 1993, Weist, 1983) – a default order for declarative sentences. The dominance of SVO is exemplified in Figure 4.4, which presents the distribution of word orders in the UD data.

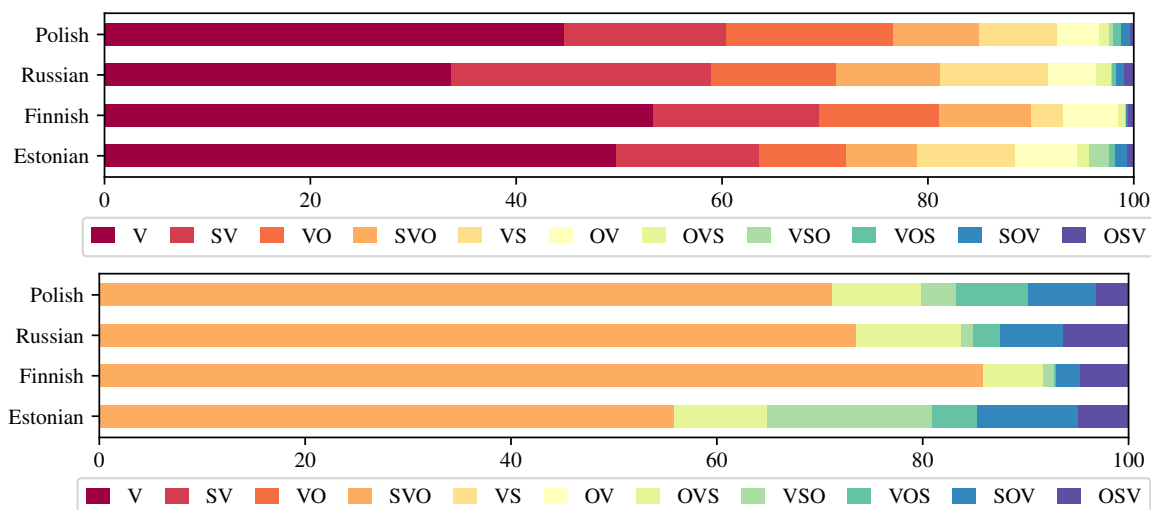


**Figure 4.4:** The distribution over the orders of the core clausal elements in the UD corpora for all clauses (top) and transitive clauses with explicit subject and object (bottom).

For all treebanks 78%+ of verbs either lack explicit core arguments or have arguments ordered according to the dominant SVO (SV, VO, SVO). This statistic directly translates to the reliability of word order as a cue to argument structure (i.e., the proportion of times it leads to the correct interpretation, as per Section 2.5). When I only consider transitive verbs for which both subject and object arguments are explicit<sup>5</sup>, SVO accounts for ~70% of possible orders in Slavic languages. For Finnic languages this proportion is 86% for the Finnish and 56% for the Estonian treebank. Notably, the latter stands out as the corpus with the most varied word orders. This word order variability is a likely contributor to the poor performance of the Estonian baselines from Section 3.3.

Note how the above word order imbalance applies to all splits of the treebanks – the models are trained and evaluated mainly on clauses adhering to the SVO ordering. In the following Section 4.1.3 I discuss how this can be problematic, even given the strong presence of morphological signals I revealed in the previous subsection.

### 4.1.3 Discussion

**Morphology vs word order** As evidenced in Section 4.1.1, most clauses in the UD treebanks support the role of morphology as a strong signal to meaning. On top of that, the treebanks contain clauses that do not adhere to the dominant SVO order, which should actively discourage the models from relying on word order. Admittedly, these clauses are infrequent, as pictured in Figure 4.4, but taken together with the validity of morphology as a cue, this should push the models towards the correct linguistic generalisation. This is especially so, given that the models are rarely exposed to evidence that does not support morphology as the most optimal signal to core relations – morphosyntactically ambiguous

<sup>5</sup>All four languages are null-subject languages in which a clause may lack an explicit subject.

| COUNTERFACTUAL TREEBANKS                                                                                    |                                                                            |                                                                                                              |                                                                            |
|-------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| svo                                                                                                         | Chłopak trzyma czerwone wiadro.<br>Kilku mężczyzn płynie po wodzie łodzią. | osv                                                                                                          | Czerwone wiadro chłopak trzyma.<br>Łodzią kilku mężczyzn po wodzie płynie. |
| ovs                                                                                                         | Czerwone wiadro trzyma chłopak.<br>Łodzią płynie po wodzie kilku mężczyzn. | vso                                                                                                          | Trzyma chłopak czerwone wiadro<br>Płynie kilku mężczyzn po wodzie łodzią.  |
| sov                                                                                                         | Chłopak czerwone wiadro trzyma.<br>Kilku mężczyzn łodzią po wodzie płynie. | vos                                                                                                          | Trzyma czerwone wiadro chłopak.<br>Płynie łodzią po wodzie kilku mężczyzn. |
| ORIGINAL SENTENCE                                                                                           |                                                                            | TRANSLATION                                                                                                  |                                                                            |
| Chłopak w czapce i kamizelce trzyma czerwone wiadro, z którego ubrany w granatowe okrycie koń wyjada siano. |                                                                            | A boy in a hat and a vest is holding a red bucket, from which a horse dressed in a navy cover is eating hay. |                                                                            |
| Kilku mężczyzn płynie po wodzie łodzią, którą przewożą motocykl, rowery i wiele innych rzeczy.              |                                                                            | A few men are sailing on a boat on which they carry a motorbike, bicycles and many other things.             |                                                                            |

**Table 4.2:** Example sentences from the 6 differently-ordered Polish counterfactual treebanks (top) and the sentences they were based on (bottom).

clauses are very uncommon (Figure 4.2).

However, morphology is complex and word order is not – at least for neural models based on RNNs and/or transformers, both of which are by design predisposed to pay attention to word order. RNNs do so through the sequential nature of their processing (see Appendix C.3), while transformers have direct access to positional information in their input embeddings (see Appendix C.4). In contrast, morphs are not singled-out in the models’ inputs. In addition, in the rare cases where morphosyntactic ambiguity does emerge, word order is a likely disambiguator, especially for Polish, Russian and Finnish (see Figure 4.3). This creates a pull away from morphology and towards reliance on word order. Also note that recognising morphosyntactic ambiguity is non-trivial, as evidenced in Figure 4.1, and that the models might not recognise that the requirement to rely on word order is only conditional.

**Insufficient evaluation** Given (i) the difference in the complexity of the two signals, (ii) the disambiguating role of word order and (iii) the strong correlation between the pre/post-verbal positions and subject/objecthood in the training data, it is not clear which one of the two – morphology or word order – is a more attractive signal to the models. Further, because the skew towards SVO applies also to the development and test data, standard evaluation on the treebanks does not effectively disambiguate between the two types of generalisation. A drop in performance on the less typical orders would have a negligible influence on the overall accuracy. Indeed, this is perhaps what we observe in the morphologically blind baseline results from Section 3.3; while it is likely that the baselines rely on a range of signals in the absence of morphology, word order is a likely contributor

|          |                                                                                                     |
|----------|-----------------------------------------------------------------------------------------------------|
| Polish   | PDB (Wróblewska, 2018), LFG (Patejuk and Przepiórkowski, 2018)                                      |
| Russian  | SynTagRus (Droganova et al., 2018), GSD <sup>6</sup> , Taiga (Shavrina and Shapovalova, 2017)       |
| Finnish  | TDT (Haverinen et al., 2014), FTB <sup>7</sup> (conversion of the FinnTreeBank (Voutilainen, 2011)) |
| Estonian | EDT (Muischnek et al., 2014), EWT (Muischnek et al., 2019)                                          |

**Table 4.3:** Treebanks that form the basis of the counterfactual dataset. <sup>6</sup>[https://universaldependencies.org/treebanks/ru\\_gsd/index.html](https://universaldependencies.org/treebanks/ru_gsd/index.html), <sup>7</sup>[https://universaldependencies.org/treebanks/fi\\_ftb/index.html](https://universaldependencies.org/treebanks/fi_ftb/index.html)

|         | SVO | OVS | SOV | OSV | VSO | VOS |          | SVO | OVS | SOV | OSV | VSO | VOS |
|---------|-----|-----|-----|-----|-----|-----|----------|-----|-----|-----|-----|-----|-----|
| Polish  | 307 | 42  | 12  | 7   | 10  | 7   | Finnish  | 154 | 13  | 2   | 1   | 0   | 0   |
| Russian | 730 | 74  | 8   | 27  | 3   | 24  | Estonian | 260 | 50  | 20  | 20  | 76  | 29  |

**Table 4.4:** Counts of transitive constructions from declarative sentences in which both subject and object are nouns (development splits of the PDB, SynTagRus, TDT and EDT UD treebanks).

to their predictions. This posits a need for more targeted evaluation paradigm. I propose such paradigm in the following Section 4.2.

## 4.2 Counterfactual treebanks

To solve the evaluation issues discussed in Section 4.1.3, I propose a more thorough evaluation paradigm in which a model is evaluated on *six counterfactual versions* of a dependency treebank – each with transitive sentences transformed to exhibit a different ordering of S, v and O (see Table 4.2 for an example). This approach allows me to isolate the effects of word order on parsing performance – I ensure that the order of the verbs’ core arguments is the only difference between the six treebank versions. It also addresses frequency issues – given the limited amounts of treebanked data (see Table 4.4), the absolute number of many word orders is insufficient to evaluate on each word order independently, as they naturally occur in the UD data.

I generate the counterfactual treebanks based on 9 existing UD treebanks (see Table 4.3<sup>8</sup>) by altering the word order in transitive clauses present in those corpora, which I have adjusted to maintain acceptability after the reordering (I discuss this further in Section 4.2.2). In contrast to artificial data, such naturalistic data has a number of important benefits. First, it incorporates realistic lexical semantics – the amount of sentences describing semantically implausible scenarios is kept to minimum. Note that because the data is naturalistic this does not happen at the expense of lexical variety. Words present in the resulting treebanks belong to various lexical semantic classes, rather than a selected few. Second, word frequency distribution remains close to that of the training data (assuming the parsers are trained on the original UD treebanks), preventing such

<sup>8</sup>I omit the PUD treebanks and Finnish OOD since they do not offer the standard train/dev/test splits.

|       | Polish |     | Russian   |      |       | Finnish |      | Estonian |      |     |
|-------|--------|-----|-----------|------|-------|---------|------|----------|------|-----|
|       | PDB    | LFG | SynTagRus | GSD  | Taiga | TDT     | FTB  | EDT      | EWT  |     |
| TRAIN | 2314   | 746 |           | 4416 | 514   | 475     | 1077 | 771      | 2291 | 108 |
| DEV   | 295    | 88  |           | 632  | 96    | 21      | 127  | 90       | 258  | 22  |
| TEST  | 272    | 83  |           | 560  | 86    | 33      | 136  | 110      | 284  | 42  |
| TRAIN | 221    | 191 |           | 737  | 91    | 66      | 64   | 47       | 182  | 11  |
| DEV   | 21     | 16  |           | 83   | 21    | 3       | 6    | 5        | 23   | 1   |
| TEST  | 26     | 21  |           | 84   | 24    | 6       | 9    | 4        | 25   | 5   |

**Table 4.5:** Counterfactual treebanks’ sentence counts (per one word order). Top: unambiguous split. Bottom: ambiguous split.

lexical factors from influencing the evaluation which targets word order.

I create two versions of the counterfactual dataset – one that contains only *morphosyntactically unambiguous* transitive clauses and one that contains only *ambiguous* transitive clauses. Together, both versions allow for answering a range of questions regarding a model’s reliance on word order vs morphology (see the introduction). To create the treebanks in either version I follow a three-step process that involves: (i) treebank filtering (ii) sentence simplification and (iii) sentence reordering – all described in detail in the following subsections. The sentence simplification and the reordering steps are identical for both unambiguous and ambiguous versions. The only difference lies in the filtering step. The whole process is largely automatic and could be adjusted to different treebanks and/or languages. Such adjustment is mostly straightforward, although special care is required at the sentence simplification stage which might require crafting language specific rules to avoid breaking the grammaticality of a language.

The remainder of this section is structured as follows: first, I describe the process of creating the counterfactual dataset, devoting a subsection to each of the three treebank-construction steps. Next, I discuss possible means and metrics for evaluating on the resulting data.

### 4.2.1 Treebank filtering

Throughout this thesis I focus on declarative, active transitive constructions with noun subject and object arguments. As a first step, I filter out all sentences which do not contain such construction. Specifically, I require that each sentence contains a verb with (i) one noun dependent linked through the UD *nsubj* relation, (ii) one noun dependent linked through the *obj/iobj* relation and, to keep the constructions simple, (iii) no auxiliary dependents or clausal complements. I also exclude ditransitive verbs from the evaluation.

Further, for the unambiguous version of the dataset, to ensure that all constructions are inflectionally unambiguous – i.e., there is only one correct subject and object assignment – for each transitive verb I require that *at least one* of the following holds:

- (i) The case used to mark the subject and the case used to mark the object are not the same<sup>9</sup> and are not syncretic for at least one core argument.<sup>10</sup>
- (ii) The verb agrees with only one argument.

Analogously, for the ambiguous version of the dataset, I require that for all transitive clauses *neither* of the two above condition holds. Note that the above conditions mirror the ambiguity conditions from Figure 4.1. As in Section 4.1.1, to detect whether either is the case I make use of Wiktionary. If it is not possible to deduce whether the verb’s argument structure is ambiguous, due to Wiktionary missing the relevant lexemes, I exclude the sentence that contains it from both the ambiguous and the unambiguous treebanks.

I present the sentence counts for each treebank after the filtering – in both unambiguous and ambiguous versions – in Table 4.5. Note that, due to the infrequent instances of full morphosyntactic ambiguity in the UD treebanks (see Section 4.1.1), for some counterfactual treebanks in the ambiguous version of the dataset, the number of resulting sentences is very low. While this prevents standard evaluation on the original test/development splits of those treebanks, the data can still be leveraged in meaningful way, as I discuss in Section 4.5.

## 4.2.2 Maintaining acceptability

While all six orderings of S, V and O are grammatical for each language I experiment with, word order remains guided by pragmatic factors (Kisselev, 2019, Neeleman and Van de Koot, 2016). As a consequence, altering the word order in dependent clauses – situated within a particular context of the main clause – can result in unacceptable utterances. Acceptability can also be compromised if the reordered arguments are associated with extensive subtrees, e.g., due to many modifiers. In such case, the reordering can break the principle of the syntactically *heavier*, long constituents appearing after the *lighter*, shorter ones (Quirk et al., 1972, Siewierska, 1993). It can also give rise to long dependencies, which can substantially increase the complexity of a sentence (Futrell et al., 2015, Gibson, 1998, Gibson et al., 2019). Consider, for instance, the Polish SVO sentence in the top of Figure 4.5. Transforming it into SOV (bottom of the figure) introduces very long dependencies between the verb and both of its arguments, making it hard to process.

To avoid such acceptability issues, I simplify the sentences before the reordering by (i) setting the transitive verb as the root of the sentence and (ii) truncating its dependency subtree by only allowing the dependents linked through selected dependency types. At the truncation stage, I keep the relations which are required to maintain grammaticality

<sup>9</sup>Note that this can happen e.g., in Finnish (see example (14) in Chapter 2).

<sup>10</sup>That is, I avoid instances where both nouns have the same form in the case used to mark the subject (usually nominative) and the case used to mark the object, e.g., accusative or partitive.

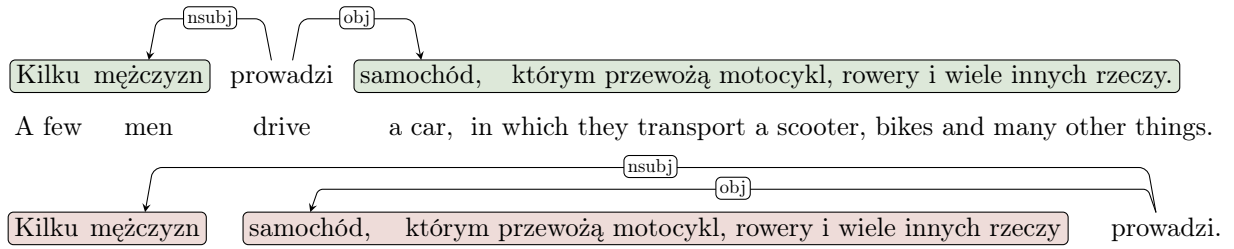


**Figure 4.5:** A clause with a *heavy* object and its example, acceptability decreasing reordering.

| Accepted relations                                                                                                                                              |                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>nsubj, obj, iobj, aux, expl:pv, obl:arg, expl:pv, nmod:gobj, nmod:gsubj, nmod:poss, nmod:arg, fixed, flat, compound, amod, nummod, case, det, cc:preconj</i> |                                                                                                                                                                                                        |
| Conditionally accepted relations                                                                                                                                |                                                                                                                                                                                                        |
| LABEL                                                                                                                                                           | CONDITION                                                                                                                                                                                              |
| <i>conj</i>                                                                                                                                                     | the head is the transitive verb's subject (this maintains subject–verb agreement)                                                                                                                      |
| <i>cc, mark</i>                                                                                                                                                 | the head is not the (new) root verb                                                                                                                                                                    |
| <i>punct</i>                                                                                                                                                    | the head is linked to its head through the <i>conj</i> relation <i>or</i> the dependent is a quotation mark                                                                                            |
| <i>nmod, obl</i>                                                                                                                                                | the size of the dependent's subtree is $\leq 3$                                                                                                                                                        |
| <i>advmod</i>                                                                                                                                                   | <i>either</i> the head is the root verb and the dependent is a Polish/Russian negation particle (nie/ nie) <i>or</i> the head is not the root verb and the size of the dependent's subtree is $\leq 3$ |
| <i>amod</i>                                                                                                                                                     | the dependent's subtree is of size 1                                                                                                                                                                   |

**Table 4.6:** Relations accepted and conditionally accepted during the sentence simplification.

and those which typically do not introduce large sub-trees. The latter is done to keep as much of the original tree as possible, in order to keep the task non-trivial. The decisions regarding which exact relations to keep at the truncation stage were made by analysing the PDB, SynTagRus, TDT and EDT treebanks, and are described in more detail in Appendix E. I present the list of all accepted and conditionally accepted relations in Table 4.6, and examples of sentences before and after the simplification in Table 4.7. Note that some relations, like *aux*, although accepted, are very unlikely to occur due to the filtering applied (see Section 4.2.1).

### 4.2.3 Reordering

I carry out the reordering as I recursively linearise the simplified dependency tree. Specifically, I obtain a linearisation of a subtree headed by a node by (i) deciding the order for that node and its dependents and (ii) combining the node's surface form and the linearised sub-trees of the dependents according to that order. For all nodes but the root verb, I follow the order exhibited in the original sentence. At the root node, I swap the positions of the verb, the subject and the object according to the word order chosen for

---

Meżczyzna jedzie motocyklem wraz z psem w okularach siedzącym w bocznym wózku. *Meżczyzna jedzie motocyklem.*

Chłopiec w kasku i białych goglach spędza czas na plastikowych sankach w trakcie zimy. *Chłopiec spędza czas na plastikowych sankach w trakcie zimy.*

Chłopak w czapce i kamizelce trzyma czerwone wiadro, z którego ubrany w granatowe okrycie koń wyjada siano. *Chłopak trzyma czerwone wiadro.*

---

Алгоритм содержит ошибки, если приводит к получению неправильных результатов либо не даёт результатов вовсе. *Алгоритм содержит ошибки.*

Однако вопрос о его смысле длительное время вызывал ожесточённые споры. *Вопрос о его смысле длительное время вызывал ожесточённые споры.*

Около 1250 года английский астроном и математик Иоанн Сакробоско написал труд по арифметике *Algorismus vulgaris*, на столетия ставший основным учебником по вычислениям в десятичной позиционной системе счисления во многих европейских универси тетах. *Около 1250 года английский астроном и математик написал труд по арифметике Algorismus vulgaris.*

---

Yleensä Peppi ja Mimmi syleilevät ihmisiä, mutta minä tarvitsin kunnon rutistuksen. *Peppi ja Mimmi syleilevät ihmisiä.*

Tosi asiassa päiväpeittonikin huutaa värikkäitä koristetyynyjä, joten ehkäpä pitää sittenkin miettiä, mikä tyynyt olisi se kaikista kivoin. *Tosi asiassa päiväpeittonikin huutaa värikkäitä koristetyynyjä.*

Tuota tekisi mieli sovittaa, sillä vartalon tarvitsen aivan tietyn mallisen mekon, jotta se näyttää hyvältä. *Vartalon tarvitsen mekon.*

---

Siis me otsustasimegi Andriesseniga, et ooperisse tuleb ainult kolm tegelast – kolm naist, sest Vermeer maalis alati naisi. *Vermeer maalis naisi.*

Ma tahaksin, et muusika teeniks minu ideid. *Muusika teeniks minu ideid.*

Ka uus ooper tuleb omamoodi multimeedia, diaproktsioonid etendavad näiteks Delfti linna põlemist. *Diaproktsioonid etendavad näiteks Delfti linna põlemist.*

---

**Table 4.7:** Example sentences before and after simplification. From top: Polish, Russian, Finnish, Estonian.

the counterfactual treebank. I maintain the original relative positions of the remaining dependents of the verbs, with the exception of those linked via *aux*, *expl* and *advmod* relations which I keep next to the verb to maintain grammaticality. Figure 4.6 presents two examples of the reordering.

After the reordering, I manually check the resulting Polish sentences to ensure their acceptability. As such, the Polish counterfactual treebanks can be viewed as *gold data*.<sup>11</sup>

#### 4.2.4 Proposed evaluation

Both the ambiguous and the unambiguous splits of the dataset hold 6 versions of each of the 9 treebanks – one version per word order. The sentences in each of these versions are counterfactual variants of some sentence *s* in the original, source treebank. As a result, there is a one-to-one correspondence between any two treebank versions, with the paired sentences differing *only* in the word order of the core constituents. This setting allows for a number of different evaluation approaches/metrics. One possible approach is to evaluate

---

<sup>11</sup>As to the remaining treebanks, small samples of perturbed sentences have been checked by native speakers of those languages to identify any issues. Any issues identified were fixed, but since the data was not thoroughly checked, these treebanks might contain some residual errors.

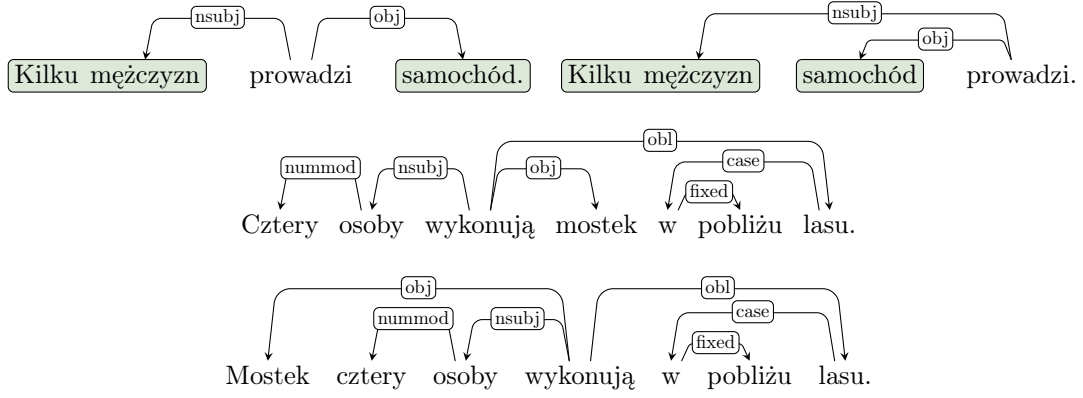


**Figure 4.6:** Two examples of sentence reordering. At the top, the simplified version of the SVO sentence from Figure 4.5 is reordered to SOV. At the bottom, a Polish SVO sentence which translates to “Four persons do a bridge [exercise] near the forest.” is reordered to OSV.

using a standard metric (e.g., LAS/UAS) on each of the treebank versions individually and report all 6 results. Alternatively (or additionally) one can employ a new metric that would quantify the difference in performance across the word order categories. In this thesis, I report LAS results for each word order individually and propose two *word order bias* metrics: **SVO attachment bias** and **SVO probability bias**, to complement the LAS results.

**SVO attachment bias** is based on a model’s *predictions* and calculates the average of the differences between the ‘privileged’ SVO UAS/LAS and UAS/LAS for other word orders. Let  $O$  be the set of all word orders, excluding SVO. The metric can be then formalised as follows:

$$\text{SVO attachment bias (UAS)} = \frac{1}{|O|} \sum_{o \in O} |\text{UAS}_{\text{svo}} - \text{UAS}_o| \quad (4.1)$$

$$\text{SVO attachment bias (LAS)} = \frac{1}{|O|} \sum_{o \in O} |\text{LAS}_{\text{svo}} - \text{LAS}_o| \quad (4.2)$$

**SVO probability bias** closely resembles the former, but it is based on the *probabilities* a model assigns to the *gold* head/label, rather than its predictions. This allows the metric to also capture more subtle differences in performance which do not lead to different predictions. The metric leverages the fact that all treebank versions consist of *matched* sentences with identical (albeit reordered) tokens and calculates the average probability difference between the equivalent tokens for two treebanks – one treebank exhibiting the SVO word order and one exhibiting any other word order. Such probability difference score is obtained for each of the non-SVO word orders – the final metric score is their average.

Let  $n$  be the token count (same in each treebank version) and  $T^x = [t_1^x, t_2^x, \dots, t_n^x]$  be a list of tokens in the treebank version  $x$ .  $T^x$  is ordered such that token  $t_k^i$  corresponds to the same dependency tree node (and consequently word form) as token  $t_k^j$  for any  $k \leq n$  and any  $i, j \in \{\text{SVO}, \text{OVS}, \text{SOV}, \text{OSV}, \text{VSO}, \text{VOS}\}$ . Sentences in each treebank version naturally follow the same order. The same order of *tokens* across treebank versions is obtained via a top-down *deterministic* dependency tree traversal.<sup>12</sup> Next, let  $f_H(t)$  be the probability assigned by a model to the *gold head* for token  $t$  and  $f_L(t)$  the probability assigned to the *gold label*. Given this setup, the SVO probability bias is defined as follows:

$$\text{SVO probability bias (head)} = \frac{1}{|O|} \frac{1}{n} \sum_{o \in O} \sum_{i=1}^n |f_H(t_i^{\text{SVO}}) - f_H(t_i^o)| \quad (4.3)$$

$$\text{SVO probability bias (label)} = \frac{1}{|O|} \frac{1}{n} \sum_{o \in O} \sum_{i=1}^n |f_L(t_i^{\text{SVO}}) - f_L(t_i^o)| \quad (4.4)$$

where  $O$  is, as before, a set of all non-SVO word orders.

Both metrics can be calculated on all relations in the dataset, or on any selected subset.

**Alternative metrics** The above two metrics were inspired by metrics used to quantify social bias in neural NLP models – and by [Dixon et al. \(2018\)](#) and [Garg et al. \(2019\)](#) in particular. Indeed, the counterfactual setting described above has many parallels with evaluation of social bias in neural NLP. Like in the treebanks proposed here, where data is split between word order categories, work on fairness in NLP often operates on data split between categories corresponding to different protected social groups ([Borkan et al., 2019](#), [Dixon et al., 2018](#), [Garg et al., 2019](#), [Gaut et al., 2020](#), [Huang et al., 2020](#)). Further, fairness in NLP work also frequently makes use of paired/matched versions of sentences which differ only in the mentioned protected social group ([Garg et al., 2019](#), [Huang et al., 2020](#), [Kiritchenko and Mohammad, 2018](#)). Because of those similarities, many metrics used to quantify differences in a model’s behavior across social groups – see [Czarnowska et al. \(2021\)](#) for an overview – could be straightforwardly adapted to evaluate on treebanks I propose here. That is to say, there are many other possible metrics that could be used on the counterfactual treebanks, beyond those proposed here, and one could look to the fairness research for inspiration.

## 4.3 Experimental details

**Models** For my experiments I select four different parsers. Two of the parsers are the same as in the experiments from Section 3.3 (base models); these are both based on

---

<sup>12</sup>To make tree traversal deterministic, I employ an algorithm that visits the children of a node in the alphabetical order of their dependency labels concatenated with the word form of the child.

the Dozat and Manning (2017)’s parser (DM), but take different inputs: (i) fine-tuned BERT and (ii) embeddings constructed with a CNN-over-characters followed by a highway network, as in Kim et al. (2016). In addition, I also include a parser with the same architecture which takes as input FASTTEXT embeddings<sup>13</sup> (Bojanowski et al., 2017b) and a parser which is purely BERT-based; i.e., the biLSTM from the Dozat and Manning’s model is dropped and it is the BERT outputs that are fed into the ReLU layers.<sup>14</sup> I experiment with the latter to isolate the effects of BERT from the effects of the LSTM.

As in Section 3.3, I use a separate, monolingual BERT for each language – Polish PolBERT (Kłeczek, 2020), Russian RuBERT (Kuratov and Arkhipov, 2019), Finnish FinBERT (Virtanen et al., 2019) and Estonian EstBERT (Tanvir et al., 2021). For each BERT I use its base, cased version. I do not use POS embeddings as inputs to the model, since in my preliminary experiments *gold* POS led to inconsistent and only minor improvements.

All three: char-CNN, FASTTEX and BERT are common inputs in neural NLP and each represents a different approach to handling morphology. They all allow the parser to access morphosyntactic subword features – at least in principle. BERT does so through operating on the level of subword units, called word-pieces (Schuster and Nakajima, 2012, Wu et al., 2016) – an automatically learned vocabulary of likely informative word segments. CNN gives the parser direct access to words’ characters and, since it is trained from scratch on the task, it can specialise in retrieving the relevant inflectional features. Lastly, since FASTTEXT embeddings are summations of the character n-gram embeddings, they are likely to encode a lot of relevant subword information; although, unlike for the other two inputs, using pre-trained FASTTEXT word embeddings does not grant the parser access to the individual n-gram embeddings, consequently preventing the model from tweaking individual subword representations to better suit the task. For more discussion on how these and other neural models allow for morphology in NLP see Appendix D.

**Data** In all experiments presented in this chapter I train the models on the biggest, *original* (i.e., unaltered) treebank available for the language and evaluate them on a *counterfactual* development split of the same treebank, unless stated otherwise.<sup>15</sup> I also experimented with a mixed-treebank condition, in which the models are trained on the concatenation of all treebanks in the language and evaluated on the concatenation of the counterfactual development splits of those treebanks; since I observed very similar performance tendencies in this condition, I omit those experiments from the discussion in

<sup>13</sup>FASTTEXT is an extension of Skip-gram (Mikolov et al., 2013a,b) (see Appendix C.1) which constructs word representations via summing the embeddings of words’ character n-grams. I use the FASTTEXT vectors trained on Wikipedia – <https://fasttext.cc/docs/en/pretrained-vectors.html>.

<sup>14</sup>See Section 3.3.1 for the Dozat and Manning (2017)’s model description.

<sup>15</sup>In this *single-treebank* condition I only experiment with the largest treebanks because other treebanks have smaller amounts of counterfactual evaluation data available (see Table 4.5) due to the filtering applied to create the counterfactual treebanks (see Section 4.2.1).

|                |     | Polish | Russian | Finnish | Estonian |
|----------------|-----|--------|---------|---------|----------|
| DM -LSTM +BERT | LAS | 90.5   | 93.4    | 92.5    | 88.8     |
|                | UAS | 95.5   | 95.4    | 94.5    | 91.9     |
| DM +BERT       | LAS | 90.5   | 92.9    | 91.9    | 86.9     |
|                | UAS | 95.5   | 95.1    | 94.1    | 90.6     |
| DM +CNN        | LAS | 81.8   | 87.6    | 77.5    | 78.8     |
|                | UAS | 89.9   | 91.2    | 83.2    | 84.5     |
| DM +FASTTEXT   | LAS | 81.8   | 87.3    | 74.0    | 74.5     |
|                | UAS | 89.0   | 90.7    | 80.8    | 81.3     |

**Table 4.8:** LAS & UAS on the development splits of the PDB, SynTagRus, TDT and EDT treebanks (standard evaluation).

this chapter and move them to Appendix G instead (Table G.3, Table G.4).

**Hyperparameters and optimization** For the details on hyperparameters and optimisation see the baseline experiments in Section 3.3. As in Section 3.3, here I also train four models for each language and type of input and report average scores, unless stated otherwise. In Table 4.8, I present the labeled and unlabeled attachment scores (LAS, UAS) on the development splits of the UD treebanks (standard evaluation) for all parsers.

**Metrics** I report results for the SVO *attachment bias* and SVO *probability bias* metrics (Section 4.2.4) calculated based on predictions for the core verbal arguments. I also report UAS/LAS results calculated individually for each of the counterfactual treebank versions.

I calculate SVO attachment bias based on the average UAS/LAS results across instances of the same model trained with a different seed (four models). For SVO probability bias, I similarly average the scores across all instances of the model. Note that, while the scores for the first metric are easily interpretable, this is not the case for the latter – it is not always clear what constitutes a good result for the probability-based metric. For this reason, to put the scores for SVO probability bias in context, I also report results for the **probability bias control** (PB. control), which provides a lower-bound. PB. control uses the same equation as the SVO probability bias (Equation (4.3)/(4.4)) but instead of being calculated for the same model(s) across differently ordered versions of the treebank, it is calculated for the SVO version of the treebank across *different instances of the same model* (i.e., trained with different seeds).

## 4.4 Experiment I: Evaluation on unambiguous sentences

The first experiment targets **CONDITION I** posed in the introduction; i.e., it tests if the models’ performance is invariant to the change of order of the core elements in *inflectionally unambiguous* clauses. It involves training four different neural parsers (see

previous section) and evaluating them on the unambiguous counterfactual treebanks, proposed in Section 4.2. I discuss this evaluation in Section 4.4.1, which is followed by an error analysis, in Section 4.4.2. In the last (Section 4.4.3), I situate these results within the evaluation on *naturally occurring* sentences, exhibiting different word orders.

### 4.4.1 Results

**Bias metrics** In Table 4.9 I present the bias metrics’ results on the development splits of the counterfactual treebanks. The metrics have been calculated based only on the core arguments of transitive verbs with explicit subject and object.<sup>16</sup> Evaluation on those relations alone provides a concise signal which indicates whether the parsers’ performance on detecting subject and object is independent of the word order exhibited in a sentence.

Results from Table 4.9 suggest that, for all four parsers, the performance is *not* independent of word order. The CNN and FASTTEXT-based models stand out as the most affected. The average LAS difference between the CNN’s performance on SVO treebank version and differently ordered versions ranges between 31.5 and 43.0 LAS (!) across the four languages. For FASTTEXT these numbers are similarly high. Compared to LAS, the UAS difference is a bit lower, but still very prominent – for CNN it does not fall below 15.5; for FASTTEXT it is never below 21.4. The large differences noted for the models’ predictions are also reflected in the scores for the probability-based metric. The prediction errors naturally translate to the differences in probabilities, suggesting that when the models’ predictions on an SVO and a non-SVO clause differ, it has high confidence in making both predictions. Further, in the light of SVO attachment bias results (UAS), the scores for the SVO probability bias for head assignment (H) suggest that there are probability differences even when the models make the same predictions for SVO and non-SVO clauses.<sup>17</sup>

Results for the transformer-based models also reveal a difference in performance between SVO and other word orders. When compared to the aforementioned LAS differences noted for CNN and FASTTEXT, scores noted for BERT models, in the 6.0–13.0 range, might look pale (figuratively and literally, given the heatmap highlighting). However, these differences are high enough to be meaningful and point to the models being erroneously influenced by word order. Note that this split of the counterfactual dataset contains simple and relatively short sentences (see Section 4.2.2) in which morphology unambiguously points out the correct interpretation. If the BERT models correctly relied on morphology, the difference in performance across word orders should be close to 0 LAS. The scores for the

<sup>16</sup>To see more fine-grained results, calculated for all tokens and for each of the core relations independently, see Appendix G, Table G.5.

<sup>17</sup>Let’s assume that the head assignment is never correct for a non-SVO sentence when it is *incorrect* for its SVO alternative. Given this assumption, if the probabilities assigned to gold tokens differed only when a model made a different prediction for an SVO and a non-SVO sentence, the maximum score on the probability-based metric would be equal to 100 (maximum probability difference per token)  $\times$  the score on the prediction-based metric.

|                | Att. bias |      | Prob. bias |      | PB. control |      |
|----------------|-----------|------|------------|------|-------------|------|
|                | UAS       | LAS  | H          | L    | H           | L    |
| DM -LSTM +BERT |           |      |            |      |             |      |
| PL             | 5.8       | 13.0 | 0.06       | 0.12 | 0.00        | 0.01 |
| RU             | 2.4       | 6.0  | 0.03       | 0.06 | 0.00        | 0.01 |
| FI             | 4.4       | 8.0  | 0.04       | 0.06 | 0.00        | 0.00 |
| ET             | 6.8       | 8.5  | 0.08       | 0.06 | 0.01        | 0.01 |
| DM +BERT       |           |      |            |      |             |      |
| PL             | 5.7       | 12.6 | 0.06       | 0.12 | 0.00        | 0.01 |
| RU             | 2.6       | 6.3  | 0.03       | 0.06 | 0.00        | 0.01 |
| FI             | 3.7       | 6.1  | 0.04       | 0.05 | 0.00        | 0.00 |
| ET             | 7.4       | 9.7  | 0.10       | 0.07 | 0.02        | 0.02 |

|         | Att. bias |      | Prob. bias |      | PB. control |      |
|---------|-----------|------|------------|------|-------------|------|
|         | UAS       | LAS  | H          | L    | H           | L    |
| DM +CNN |           |      |            |      |             |      |
| PL      | 22.1      | 43.0 | 0.25       | 0.36 | 0.03        | 0.09 |
| RU      | 15.5      | 35.0 | 0.18       | 0.30 | 0.03        | 0.11 |
| FI      | 20.8      | 40.6 | 0.26       | 0.33 | 0.05        | 0.10 |
| ET      | 17.8      | 31.5 | 0.23       | 0.27 | 0.05        | 0.12 |
| DM +FT  |           |      |            |      |             |      |
| PL      | 21.4      | 30.2 | 0.27       | 0.27 | 0.05        | 0.10 |
| RU      | 25.7      | 34.8 | 0.28       | 0.27 | 0.02        | 0.06 |
| FI      | 33.7      | 42.2 | 0.39       | 0.33 | 0.07        | 0.10 |
| ET      | 29.1      | 30.8 | 0.36       | 0.27 | 0.07        | 0.13 |

**Table 4.9:** Bias metrics’ results when evaluated *on core arguments only* (see Appendix G, Table G.5 for more results). *H* stands for head, *L* stands for label.

probability-based metric also point to differences in the models’ behaviour across the word orders – for each language they are noticeably higher than the PB. control. Considering the results of the prediction-based metric, the picture painted by the SVO probability bias scores corresponds to the situation where the probability difference is in the high 90s for each token mis-predicted for the non-SVO orders and the tokens in the SVO ordered clauses are rarely mis-predicted.

The prominent difference between the transformer and non-transformer-based models revealed here gives weight to my earlier argument regarding the importance of linguistically-targeted evaluation (see Section 3.4). While standard evaluation does record a difference between the transformer-based and CNN/FASTTEXT-based models (see Table 4.8), the gap I reveal here is much bigger and points to an important difference between the models. Although all three types of input embeddings encode subword-level knowledge, the parsers based on CNNs and FASTTEXT hardly rely on morphological signals. This is a significant result but further investigation regarding its cause lies outside of the scope of this thesis. Potentially the mechanisms through which these models process language are radically different to that of BERT, but it could also be a *pre-training artifact*. At pre-training BERT models are exposed to vast amounts of data which includes domains likely to exhibit more word order variability than the UD treebanks. Further, for each (largest) UD treebank but SynTagRus<sup>18</sup> I registered an overlap between the texts it covers and the corresponding BERT’s pre-training data. Consequently, the models might have seen the source versions of the development and test sentences with their original word order (most commonly SVO), which could have non-trivial effects on their performance in Table 4.9. While I partially address the latter concern in the next chapter, where I account for lexical semantics, I leave deeper investigation of this result for future work.

<sup>18</sup>Kuraton and Arkhipov (2019) do not list sources for their news data so I was unable to detect whether there is an overlap with SynTagRus.

|                | Polish<br>295 clauses |       |      |      | Russian<br>632 clauses |       |      |      | Finnish<br>127 clauses |       |      | Estonian<br>258 clauses |       |      |
|----------------|-----------------------|-------|------|------|------------------------|-------|------|------|------------------------|-------|------|-------------------------|-------|------|
|                | all                   | nsubj | obj  | iobj | all                    | nsubj | obj  | iobj | all                    | nsubj | obj  | all                     | nsubj | obj  |
| DM -LSTM +BERT |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO            | 94.2                  | 99.5  | 98.8 | 89.8 | 97.7                   | 98.9  | 99.2 | 96.2 | 98.0                   | 100   | 99.2 | 94.1                    | 98.0  | 98.0 |
| OVS            | 91.5                  | 91.9  | 92.9 | 79.9 | 96.3                   | 95.4  | 94.2 | 95.2 | 96.5                   | 97.5  | 97.5 | 92.9                    | 94.4  | 94.6 |
| SOV            | 91.4                  | 95.6  | 84.3 | 75.8 | 95.9                   | 96.2  | 93.8 | 87.8 | 94.8                   | 92.9  | 92.3 | 89.2                    | 89.5  | 88.8 |
| OSV            | 90.9                  | 91.1  | 90.6 | 77.0 | 95.9                   | 94.0  | 95.8 | 95.7 | 96.0                   | 97.6  | 93.1 | 88.8                    | 91.2  | 81.3 |
| VSO            | 84.5                  | 67.0  | 76.6 | 75.8 | 93.1                   | 85.4  | 86.3 | 80.8 | 94.4                   | 90.9  | 94.1 | 92.0                    | 90.3  | 94.6 |
| VOS            | 88.7                  | 76.9  | 94.7 | 84.8 | 95.9                   | 91.7  | 97.9 | 95.7 | 90.3                   | 76.8  | 83.3 | 88.5                    | 84.4  | 86.2 |
| DM +BERT       |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO            | 94.2                  | 99.8  | 98.3 | 88.5 | 97.4                   | 99.5  | 99.0 | 95.7 | 97.5                   | 99.8  | 98.8 | 93.9                    | 97.5  | 97.2 |
| OVS            | 91.5                  | 93.9  | 92.0 | 78.3 | 96.0                   | 95.1  | 93.3 | 95.2 | 96.4                   | 98.4  | 99.0 | 93.0                    | 95.0  | 94.2 |
| SOV            | 90.6                  | 96.4  | 83.9 | 71.3 | 95.9                   | 97.7  | 94.8 | 88.3 | 95.2                   | 96.7  | 92.7 | 88.2                    | 87.9  | 86.3 |
| OSV            | 91.0                  | 92.6  | 90.1 | 78.3 | 95.5                   | 94.3  | 94.9 | 93.1 | 96.2                   | 99.0  | 94.5 | 86.7                    | 90.0  | 76.0 |
| VSO            | 84.0                  | 69.6  | 75.1 | 70.1 | 92.3                   | 85.0  | 84.2 | 80.3 | 94.6                   | 90.4  | 96.7 | 91.1                    | 89.8  | 92.8 |
| VOS            | 88.6                  | 78.1  | 93.8 | 84.4 | 95.5                   | 91.1  | 98.1 | 95.2 | 90.6                   | 75.4  | 89.4 | 87.3                    | 82.3  | 82.8 |
| DM +CNN        |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO            | 86.1                  | 93.5  | 89.2 | 72.5 | 92.2                   | 91.9  | 91.7 | 67.6 | 86.2                   | 92.3  | 86.7 | 84.6                    | 87.5  | 83.6 |
| OVS            | 70.0                  | 45.7  | 41.9 | 43.9 | 80.7                   | 52.6  | 58.0 | 62.2 | 72.7                   | 60.4  | 53.1 | 77.5                    | 69.8  | 59.2 |
| SOV            | 72.4                  | 84.3  | 35.9 | 32.0 | 81.8                   | 71.5  | 58.2 | 43.6 | 63.8                   | 58.5  | 27.0 | 68.8                    | 51.4  | 47.9 |
| OSV            | 66.5                  | 43.8  | 38.2 | 38.1 | 79.5                   | 47.1  | 65.5 | 62.8 | 70.0                   | 85.2  | 25.0 | 67.8                    | 63.3  | 25.9 |
| VSO            | 66.4                  | 38.3  | 46.4 | 38.5 | 75.7                   | 35.3  | 52.2 | 46.8 | 65.9                   | 41.9  | 48.8 | 73.5                    | 52.9  | 67.7 |
| VOS            | 68.2                  | 17.4  | 78.4 | 65.6 | 81.1                   | 37.3  | 83.9 | 63.3 | 66.9                   | 37.2  | 51.0 | 71.7                    | 53.0  | 50.1 |
| DM +FASTTEXT   |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO            | 86.2                  | 88.7  | 88.4 | 77.9 | 93.1                   | 91.9  | 95.3 | 88.3 | 82.5                   | 88.8  | 84.5 | 79.4                    | 80.2  | 84.4 |
| OVS            | 71.4                  | 61.7  | 53.4 | 27.4 | 82.7                   | 66.2  | 62.6 | 60.7 | 66.8                   | 45.1  | 49.6 | 72.5                    | 69.1  | 56.6 |
| SOV            | 74.3                  | 72.0  | 62.0 | 42.6 | 83.8                   | 65.5  | 74.5 | 63.8 | 66.9                   | 49.6  | 64.4 | 68.6                    | 47.1  | 71.7 |
| OSV            | 67.5                  | 67.3  | 41.0 | 22.5 | 81.5                   | 70.4  | 61.9 | 38.8 | 66.9                   | 78.0  | 26.1 | 65.2                    | 69.3  | 33.0 |
| VSO            | 68.2                  | 56.6  | 56.1 | 47.5 | 64.2                   | 39.9  | 43.6 | 46.8 | 46.6                   | 27.8  | 44.5 | 51.2                    | 38.1  | 55.1 |
| VOS            | 73.4                  | 42.8  | 77.9 | 61.5 | 70.1                   | 43.8  | 59.1 | 60.6 | 50.2                   | 20.6  | 38.2 | 53.8                    | 36.0  | 39.4 |

**Table 4.10:** LAS scores on the development splits of the *unambiguous* counterfactual treebanks. All results are averaged across four models, trained with different random seeds.

**Full LAS breakdown** While informative, the above bias metrics’ results offer only an initial insight. What they do not uncover is whether there are any word orders/relations that are affected more than others. They also do not reveal the direction of the bias – the metrics use absolute differences and are based on the assumption that SVO yields the best results, motivated by the treebanks’ word order statistics (see Section 4.1.2). In this paragraph, I address both of these gaps by looking into the full breakdown of LAS performance, across the different word orders and relations.

Table 4.10 presents the full performance trend on the development splits of the counterfactual treebanks (see Table G.6, in Appendix G for standard deviation scores). For each language and word order, I report a separate LAS for *nsubj*, *obj* and *iobj* dependents of transitive verbs, as well as the accumulated LAS, for all relations (*all*).<sup>19</sup> I also report the number of transitive constructions that have contributed to each score. Note that, while reporting results for each [language, model, word order, relation] combination amounts to quite a lot of numbers in Table 4.10, it is not the individual numbers that matter the

<sup>19</sup>I exclude the *iobj* relation from the tables for Finnish and Estonian since it does not appear in those treebanks.

|          |   | SVO         | OVS         | SOV         | OSV         | VSO         | VOS         |
|----------|---|-------------|-------------|-------------|-------------|-------------|-------------|
| Polish   | H | 0.99 (0.02) | 0.98 (0.02) | 0.95 (0.12) | 0.91 (0.14) | 0.94 (0.12) | 0.92 (0.14) |
|          | L | 0.91 (0.14) | 0.91 (0.16) | 0.86 (0.19) | 0.86 (0.19) | 0.88 (0.17) | 0.87 (0.18) |
| Russian  | H | 0.93 (0.05) | 0.91 (0.17) | 0.93 (0.14) | 0.92 (0.14) | 0.94 (0.12) | 0.94 (0.12) |
|          | L | 0.92 (0.13) | 0.96 (0.10) | 0.93 (0.13) | 0.92 (0.15) | 0.93 (0.14) | 0.92 (0.15) |
| Finnish  | H | 0.77 (0.00) | 0.92 (0.15) | 0.92 (0.13) | 0.99 (0.02) | 0.92 (0.14) | 0.96 (0.11) |
|          | L | 0.89 (0.17) | 0.98 (0.07) | 0.92 (0.17) | 0.95 (0.13) | 0.94 (0.12) | 0.89 (0.18) |
| Estonian | H | 0.81 (0.19) | 0.88 (0.17) | 0.85 (0.19) | 0.88 (0.17) | 0.84 (0.19) | 0.85 (0.19) |
|          | L | 0.87 (0.15) | 0.89 (0.17) | 0.87 (0.18) | 0.89 (0.17) | 0.90 (0.17) | 0.90 (0.17) |

**Table 4.11:** DM +BERT; average confidence score for heads (H) on mis-attached core tokens and for labels (L) on mis-labeled core tokens. I report standard deviation in brackets.

most, but the pattern of performance differences revealed through the heatmap shading.

Across all languages and all models, I observe over-fitting to the dominant SVO word order, with SVO yielding the best performance for every considered relation. LAS degradation applies not only to the particularly rare orders, such as VOS, but also to the more frequent ones, like OVS. But while all of the non-SVO orders are affected, some appear to be more challenging than others. Frequency of the order in the treebank seems to be one contributing factor – most frequent word orders, as per Figure 4.4, tend to yield highest overall results (*all*). Notably, for some [model, language] combinations, such as Russian/Finnish DM –LSTM +BERT, the ranking of word orders according to LAS on *all* relations, is identical to the frequency-based ranking from Figure 4.4. Another noteworthy observation is that a drop in performance for one of the two core relations does not always go in hand with a performance drop for the other – see e.g., Polish VOS order results. It also appears that the relative ordering of clausal elements influences which type of relation is most affected – I will come back and add to this discussion in Section 4.4.2.

Similarly to the bias metrics, LAS breakdown from Table 4.10 uncovers a prominent difference in performance of the BERT-based and the non-BERT-based models. While the first do suffer from a notable performance decline for the particularly rare orders, this decline is nowhere as substantial as the one for the models trained with CNN and FASTTEXT inputs. Performance on the *nsubj* relations in Polish VOS clauses serves as an illustrative example of that difference. On those relations, BERT models suffer a drop of  $\sim 20$  LAS, compared to the SVO order. For FASTTEXT this drop reaches 45.9 LAS and for the CNN 75.6 LAS (!). As noted in the discussion of bias metrics’ results; this result calls for further investigation but investigating it is outside this thesis’ scope.

Finally, it is important to address the overall low results of the Polish models on the *iobj* relation. These are likely due to the PDB treebank going against the UD guideline to label exclusive objects within a clause as *obj*, regardless of the morphological case or

semantic role<sup>20</sup>, which makes the distinction between the *iobj* and *obj* labels less clear.

**Confidence scores** As a final step in this subsection, I inspect the confidence scores for the models’ *erroneous* predictions for the core dependents of transitive verbs, across the different orders. In Table 4.11, I present these results for DM +BERT model, which serves as a representative of a general trend that holds also for other models. The table reveals that word order of the core constituents has little to no effect on models’ confidence scores – regardless of the ordering, the models assign very high, 80%+ probabilities to the incorrect heads and labels. This puts results from Table 4.10 in a new light; not only are the models (in particular CNN-based and FASTTEXT-based models) often wrong when sentences do not follow SVO, but they make their erroneous predictions with high confidence.

#### 4.4.2 Error analysis

To get a better understanding of errors made by the models I inspect the confusion matrices for the *nsubj*, *obj* and *iobj* relations. In Figure 4.7 and Figure 4.8 I present those matrices for the Slavic and Finnic language pairs, respectively. I report both the *label* confusion matrix for the correctly attached tokens (i.e., correctly identified head), as well as the counts of tokens which were incorrectly attached (a column marked with *WH* for “wrong head”). In this evaluation I present the results for only one version of each model, trained with a random seed 1. Since in Section 4.4.1 I did not observe any meaningful difference between the BERT-based parser that makes use of an LSTM and one that does not, in this section I only consider the first.

The remainder of this subsection is structured as follows: first, I describe the general error trends observed for the subject and object relations, across all three models. Then, I move to discuss the differences between the models.

**Subject relations** For Slavic language pair, performance on subject relations is best on subject-initial orders (SVO, SOV) and worst on verb-initial orders, for all three models. Verb-initial orders are also the most challenging for the Finnic languages, especially for the FASTTEXT-based model. For Finnish, models are much more likely to correctly identify the subject if it directly precedes the verb – SVO, OSV. Those two word orders also stand out as the best performing in Estonian, but so does the OVS. For both language families, the most common types of errors include (i) mislabeling the subject as the object and (ii) attaching the subject to an incorrect head. The first is most likely to happen when the subject directly follows the verb, especially for Slavic languages. Note that mislabeling subject as object in OVS and VSO is one of the most common errors made by Slavic BERT-based models. This error is also relatively common for VOS orders where the subject is the last

---

<sup>20</sup><https://universaldependencies.org/u/dep/iobj.html>

|       |      |      |     |      |       |    |      |     |      |       |    |      |     |      |       |     |      |     |      |       |     |      |     |      |       |     |      |     |      |       |     |
|-------|------|------|-----|------|-------|----|------|-----|------|-------|----|------|-----|------|-------|-----|------|-----|------|-------|-----|------|-----|------|-------|-----|------|-----|------|-------|-----|
|       |      | SVO  |     |      |       |    | OVS  |     |      |       |    | SOV  |     |      |       |     | OSV  |     |      |       |     | VSO  |     |      |       |     | VOS  |     |      |       |     |
|       |      | nsbj | obj | iobj | other | WH | nsbj | obj | iobj | other | WH | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  |
| BERT  | nsbj | 295  | 0   | 0    | 0     | 0  | 278  | 13  | 3    | 0     | 1  | 287  | 3   | 1    | 2     | 2   | 272  | 2   | 0    | 0     | 21  | 230  | 44  | 15   | 2     | 4   | 238  | 12  | 2    | 2     | 41  |
|       | obj  | 1    | 231 | 2    | 0     | 1  | 17   | 211 | 4    | 2     | 1  | 5    | 192 | 7    | 1     | 30  | 15   | 211 | 2    | 2     | 5   | 24   | 167 | 8    | 2     | 34  | 7    | 221 | 2    | 2     | 3   |
|       | iobj | 0    | 6   | 53   | 2     | 0  | 3    | 4   | 47   | 6     | 1  | 1    | 1   | 42   | 1     | 16  | 2    | 3   | 44   | 8     | 4   | 1    | 1   | 42   | 3     | 14  | 1    | 4   | 52   | 4     | 0   |
| CNN   | nsbj | 275  | 10  | 1    | 0     | 9  | 106  | 158 | 14   | 3     | 14 | 244  | 16  | 5    | 2     | 28  | 128  | 25  | 3    | 3     | 136 | 91   | 155 | 16   | 12    | 21  | 42   | 57  | 11   | 2     | 183 |
|       | obj  | 4    | 213 | 3    | 3     | 12 | 121  | 97  | 3    | 1     | 13 | 28   | 71  | 4    | 6     | 126 | 115  | 89  | 3    | 1     | 27  | 3    | 93  | 1    | 10    | 128 | 12   | 188 | 5    | 6     | 24  |
|       | iobj | 2    | 9   | 47   | 2     | 1  | 23   | 4   | 29   | 3     | 2  | 4    | 1   | 22   | 5     | 29  | 21   | 3   | 28   | 3     | 6   | 1    | 2   | 26   | 3     | 29  | 4    | 6   | 43   | 3     | 5   |
| FASTT | nsbj | 253  | 21  | 3    | 0     | 18 | 167  | 88  | 13   | 6     | 21 | 199  | 22  | 2    | 3     | 69  | 165  | 32  | 4    | 1     | 93  | 128  | 76  | 10   | 35    | 46  | 103  | 40  | 4    | 7     | 141 |
|       | obj  | 10   | 208 | 3    | 3     | 11 | 71   | 131 | 2    | 1     | 30 | 27   | 127 | 1    | 2     | 78  | 75   | 84  | 1    | 3     | 72  | 6    | 100 | 0    | 4     | 125 | 9    | 187 | 2    | 23    | 14  |
|       | iobj | 4    | 7   | 46   | 4     | 0  | 19   | 9   | 18   | 4     | 11 | 6    | 0   | 23   | 5     | 27  | 21   | 5   | 13   | 2     | 20  | 1    | 4   | 24   | 2     | 30  | 3    | 6   | 39   | 6     | 7   |

|       |      |      |     |      |       |    |      |     |      |       |    |      |     |      |       |     |      |     |      |       |     |      |     |      |       |     |      |     |      |       |     |
|-------|------|------|-----|------|-------|----|------|-----|------|-------|----|------|-----|------|-------|-----|------|-----|------|-------|-----|------|-----|------|-------|-----|------|-----|------|-------|-----|
|       |      | SVO  |     |      |       |    | OVS  |     |      |       |    | SOV  |     |      |       |     | OSV  |     |      |       |     | VSO  |     |      |       |     | VOS  |     |      |       |     |
|       |      | nsbj | obj | iobj | other | WH | nsbj | obj | iobj | other | WH | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  | nsbj | obj | iobj | other | WH  |
| BERT  | nsbj | 630  | 0   | 0    | 1     | 1  | 603  | 23  | 0    | 3     | 3  | 615  | 9   | 0    | 2     | 6   | 591  | 10  | 0    | 3     | 28  | 553  | 59  | 8    | 10    | 2   | 579  | 3   | 0    | 5     | 45  |
|       | obj  | 4    | 579 | 0    | 3     | 0  | 41   | 540 | 0    | 2     | 3  | 6    | 555 | 0    | 7     | 18  | 17   | 549 | 2    | 12    | 6   | 53   | 465 | 4    | 8     | 56  | 8    | 571 | 3    | 2     | 2   |
|       | iobj | 0    | 0   | 45   | 2     | 0  | 0    | 0   | 45   | 1     | 1  | 0    | 0   | 39   | 2     | 6   | 0    | 0   | 43   | 3     | 1   | 0    | 0   | 30   | 1     | 16  | 0    | 0   | 45   | 1     | 1   |
| CNN   | nsbj | 543  | 28  | 5    | 22    | 34 | 362  | 204 | 6    | 42    | 18 | 436  | 85  | 21   | 35    | 55  | 292  | 75  | 1    | 32    | 232 | 236  | 274 | 20   | 51    | 51  | 244  | 58  | 4    | 29    | 297 |
|       | obj  | 30   | 515 | 8    | 20    | 13 | 187  | 335 | 11   | 22    | 31 | 78   | 321 | 6    | 41    | 140 | 157  | 349 | 19   | 33    | 28  | 50   | 285 | 14   | 34    | 203 | 61   | 470 | 9    | 28    | 18  |
|       | iobj | 0    | 3   | 33   | 9     | 2  | 5    | 2   | 28   | 10    | 2  | 1    | 2   | 16   | 4     | 24  | 6    | 1   | 30   | 8     | 2   | 1    | 1   | 13   | 4     | 28  | 8    | 1   | 31   | 5     | 2   |
| FASTT | nsbj | 598  | 6   | 2    | 11    | 15 | 416  | 184 | 1    | 17    | 14 | 464  | 42  | 4    | 24    | 98  | 436  | 90  | 1    | 6     | 99  | 213  | 54  | 1    | 37    | 327 | 227  | 14  | 1    | 13    | 377 |
|       | obj  | 11   | 559 | 2    | 3     | 11 | 225  | 329 | 3    | 7     | 22 | 75   | 430 | 2    | 5     | 74  | 133  | 326 | 3    | 14    | 110 | 44   | 162 | 0    | 12    | 368 | 59   | 284 | 1    | 30    | 212 |
|       | iobj | 3    | 0   | 41   | 2     | 1  | 8    | 3   | 26   | 5     | 5  | 2    | 0   | 26   | 4     | 15  | 9    | 1   | 19   | 6     | 12  | 0    | 0   | 8    | 1     | 38  | 2    | 0   | 28   | 8     | 9   |

**Figure 4.7:** Polish (top) and Russian (bottom) confusion matrices for the development splits of counterfactual treebanks. WH stands for wrong head.

core element of the clause – this is especially so for the CNN-based models, across all languages. The second type of error – misattachment – is most frequent for orders in which the subject directly precedes or follows the object. After closer inspection of those instances I notice that for the majority of errors, the subject has been incorrectly identified as the object’s modifier. Another error that stands out for FASTTEXT is misrecognising the subject as the root – this happens with high frequency for the verb-initial orders (VSO, VOS) for all languages but Polish.

**Object relations** Similarly to the subject relations, the most common error types for objects include mislabeling them as subjects and attaching them to wrong heads. In Slavic languages, the first occurs mostly when the object is the first core element of the clause – OVS and OSV. Note that in the light of the strong results of Slavic models on subject assignment in the *subject-initial* clauses (previous paragraph) this suggests that the initial position in a clause is an important cue to subjecthood for these models. For Finnic language pair, the models seem to rely more on the relative position of the noun with respect to the verb – the mislabeling of objects as subjects is most common for OVS

|        |       | SVO   |     |       |    | OVS   |     |       |    | SOV   |     |       |     | OSV   |     |       |     | VSO   |     |       |     | VOS   |     |       |     |
|--------|-------|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|
|        |       | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH  | nsubj | obj | other | WH  | nsubj | obj | other | WH  | nsubj | obj | other | WH  |
| BERT   | nsubj | 251   | 0   | 2     | 5  | 242   | 6   | 2     | 8  | 225   | 1   | 1     | 31  | 223   | 0   | 2     | 33  | 227   | 2   | 3     | 26  | 209   | 6   | 4     | 39  |
|        | obj   | 1     | 253 | 3     | 1  | 5     | 246 | 3     | 4  | 8     | 224 | 2     | 24  | 8     | 199 | 2     | 49  | 0     | 239 | 9     | 10  | 6     | 215 | 3     | 34  |
| CNN    | nsubj | 224   | 6   | 12    | 16 | 166   | 54  | 10    | 28 | 125   | 6   | 5     | 122 | 153   | 14  | 12    | 79  | 140   | 30  | 27    | 61  | 144   | 29  | 14    | 71  |
|        | obj   | 9     | 218 | 18    | 13 | 74    | 150 | 21    | 13 | 80    | 119 | 23    | 36  | 58    | 69  | 20    | 111 | 16    | 173 | 35    | 34  | 30    | 134 | 17    | 77  |
| FASTT. | nsubj | 196   | 13  | 22    | 27 | 180   | 37  | 11    | 30 | 120   | 11  | 14    | 113 | 180   | 17  | 19    | 42  | 81    | 5   | 7     | 165 | 92    | 10  | 13    | 143 |
|        | obj   | 16    | 207 | 19    | 16 | 46    | 142 | 53    | 17 | 37    | 170 | 18    | 33  | 35    | 70  | 30    | 123 | 6     | 81  | 28    | 143 | 11    | 63  | 14    | 170 |

|        |       | SVO   |     |       |    | OVS   |     |       |    | SOV   |     |       |    | OSV   |     |       |    | VSO   |     |       |    | VOS   |     |       |    |
|--------|-------|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|
|        |       | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH |
| BERT   | nsubj | 126   | 0   | 0     | 1  | 125   | 0   | 0     | 2  | 120   | 1   | 0     | 6  | 125   | 0   | 2     | 0  | 115   | 3   | 1     | 8  | 96    | 6   | 13    | 12 |
|        | obj   | 0     | 126 | 1     | 0  | 0     | 126 | 1     | 0  | 0     | 116 | 3     | 8  | 0     | 121 | 1     | 5  | 0     | 121 | 5     | 1  | 0     | 110 | 7     | 10 |
| CNN    | nsubj | 118   | 2   | 0     | 7  | 87    | 28  | 3     | 9  | 73    | 1   | 8     | 45 | 111   | 1   | 6     | 9  | 58    | 34  | 5     | 30 | 52    | 37  | 10    | 28 |
|        | obj   | 1     | 106 | 12    | 8  | 42    | 65  | 13    | 7  | 52    | 23  | 17    | 35 | 7     | 27  | 16    | 77 | 10    | 56  | 21    | 40 | 3     | 64  | 17    | 43 |
| FASTT. | nsubj | 110   | 2   | 4     | 11 | 54    | 41  | 18    | 14 | 63    | 4   | 5     | 55 | 96    | 8   | 5     | 18 | 14    | 10  | 12    | 91 | 15    | 21  | 16    | 75 |
|        | obj   | 0     | 102 | 10    | 15 | 37    | 75  | 6     | 9  | 15    | 74  | 5     | 33 | 6     | 36  | 8     | 77 | 0     | 36  | 12    | 79 | 0     | 37  | 5     | 85 |

**Figure 4.8:** Estonian (top) and Finnish (bottom) confusion matrices for the development splits of counterfactual treebanks. WH stands for wrong head.

and SOV – orders in which the object directly precedes the verb. Again, this aligns with the insights from the previous paragraph.

When it comes to the misattachments, such errors are most common when the object directly precedes/follows the subject. For Slavic languages they appear mostly in the latter case (SOV and VSO orders), although for FASTTEXT-based models they are also common for OSV (Polish and Russian) and VOS (Russian). For Finnic languages, the reverse is true; the objects are most likely to be misattached when they *precede* the subject. After closer inspection, I notice that for most of such errors the object is incorrectly interpreted as the subject’s modifier. I also find that for the UD corpora the models were trained on, in Slavic languages nominal modifiers typically follow the modified noun, while for Finnic languages they typically precede it; the models seem to have learned to rely on those tendencies as signals to dependencies. Other re-occurring misattachment errors involve misattaching the object as an object of the subject/other token or, for Finnic languages, as a part of a compound noun or a multi-word-expression. Finally, FASTTEXT often mislabels the object as the root in the VOS and VOS ordered clauses.<sup>21</sup>

**Model differences** For all three models I observe similar error patterns; they differ mainly with respect to the frequency with which they make different kinds of errors. BERT-based models seem more prone to make misattachment errors than mislabeling errors, especially when parsing Finnic languages. Note, for example, that the Finnish

<sup>21</sup>CNN also mislabels objects and subjects as a root at times, but much less frequently than FASTTEXT.

BERT-based model never mislabels the object as the subject. Many of those misattachment errors seem to stem from misinterpreting one of the core arguments as a modifier of the other. Given that in all four languages noun modifier relations can be communicated through the same inflectional markings as the core relations<sup>22</sup>, these errors can stem from the models mis-interpreting morphological signals, rather than ignoring them. On the other hand, confusing the subject and the object – a common error for the Slavic BERT models, especially for the VSO order – points to the models ignoring the morphological signal when making their predictions.

The CNN-based models stand out in making the most mislabeling errors – consider for example OVS and VSO Polish word orders, where the model makes more mislabeling errors than correct predictions. This result is somewhat surprising, given that the CNN-based model is, at least in principle, better suited to make use of morphological signals than FASTTEXT – in contrast to FASTTEXT, it has direct access to all the characters in a word and the processing of those characters is directly trained for the parsing task. One possible explanation is that FASTTEXT-based model performs better because it can tap into the lexical information encoded in the pre-trained embeddings, which can serve as an additional signal to dependencies. In the next chapter, I directly test this hypothesis, by accounting for lexical semantics in my experiments. Also note that while they are particularly prone to mislabelling, CNN models also make a substantial amount of misattachment errors.

While the CNN is the worst when it comes to mislabeling, FASTTEXT is the worst when it comes to misattachment errors. At times, the model makes a substantial amount of these errors in cases where the other two models barely make any – see e.g., Russian object relations for VOS word order. In some of those instances the model is making the same types of errors as the other models, just to a larger extent. In other cases, like the aforementioned Russian VOS case, FASTTEXT stands out in making a lot of mislabeling-as-root errors. This is a surprising phenomenon, especially given the scarcity of such errors for the other two models.

### 4.4.3 Naturally occurring sentences

For completeness, in Table 4.12, I also present LAS results for each word order, on the differently ordered transitive constructions from the *original*, unfiltered and unaltered treebanks.<sup>23</sup> Since the presence/absence of the LSTM did not have much influence on the BERT-based model in previous evaluation – the models perform almost identically – I

---

<sup>22</sup>E.g., in Polish and Russian noun modifiers are communicated through the genitive case that is at times syncretic with the core accusative case.

<sup>23</sup>Note that if a sentence has both a transitive and an intransitive verb with a *nsubj* dependent, the intransitive verb's *nsubj* dependent is not included in the calculation of the *nsubj* LAS reported in the table. Further, if a sentence has 2 transitive verbs, one exhibiting e.g., SVO and the other VOS word order, the total attachment score of that sentence is counted towards the *all* LAS for both SVO and VOS orders.

|              | Polish |          |      |      |     | Russian |          |      |      |     | Finnish |          |      |     | Estonian |          |      |     |
|--------------|--------|----------|------|------|-----|---------|----------|------|------|-----|---------|----------|------|-----|----------|----------|------|-----|
|              | all    | nsubjobj | iobj | #    |     | all     | nsubjobj | iobj | #    |     | all     | nsubjobj | #    |     | all      | nsubjobj | #    |     |
| DM +BERT     |        |          |      |      |     |         |          |      |      |     |         |          |      |     |          |          |      |     |
| svo          | 90.7   | 99.4     | 98.3 | 87.7 | 307 | 93.9    | 99.1     | 99.0 | 95.9 | 730 | 92.1    | 99.6     | 98.7 | 154 | 87.9     | 96.9     | 94.5 | 270 |
| ovs          | 88.0   | 97.6     | 87.8 | 83.3 | 42  | 94.1    | 95.5     | 96.8 | 100  | 74  | 94.8    | 100      | 100  | 13  | 86.2     | 91.3     | 97.3 | 50  |
| sov          | 90.9   | 100      | 86.7 | 100  | 12  | 91.5    | 91.7     | 95.2 | 100  | 8   | 89.7    | 100      | 100  | 2   | 86.4     | 93.3     | 96.7 | 20  |
| osv          | 87.6   | 100      | 83.3 | 0.0  | 7   | 94.4    | 86.4     | 91.7 | 100  | 27  | 94.9    | 100      | 100  | 1   | 84.3     | 100      | 100  | 2   |
| vso          | 83.7   | 90.0     | 55.6 | 40.0 | 10  | 84.5    | 55.6     | 66.7 | 0.0  | 3   | -       | -        | -    | 0   | 87.1     | 97.8     | 96.5 | 76  |
| vos          | 86.2   | 85.7     | 83.3 | 83.3 | 7   | 93.5    | 95.8     | 100  | 100  | 24  | -       | -        | -    | 0   | 89.9     | 97.7     | 89.7 | 29  |
| DM +CNNs     |        |          |      |      |     |         |          |      |      |     |         |          |      |     |          |          |      |     |
| svo          | 81.5   | 92.5     | 87.8 | 73.3 | 307 | 88.8    | 92.4     | 91.3 | 74.0 | 730 | 77.4    | 92.6     | 80.3 | 154 | 79.7     | 88.9     | 80.9 | 270 |
| ovs          | 71.0   | 49.2     | 48.9 | 45.2 | 42  | 88.5    | 72.5     | 79.0 | 66.6 | 74  | 74.1    | 71.8     | 84.6 | 13  | 76.4     | 76.7     | 65.3 | 50  |
| sov          | 82.7   | 94.5     | 63.3 | 77.8 | 12  | 81.9    | 83.3     | 66.6 | 100  | 8   | 88.5    | 100      | 50.0 | 2   | 83.5     | 71.7     | 83.3 | 20  |
| osv          | 79.4   | 80.9     | 33.3 | 33.3 | 7   | 87.6    | 61.8     | 73.6 | 60.0 | 27  | 71.8    | 100      | 100  | 1   | 72.6     | 100      | 50.0 | 2   |
| vso          | 76.3   | 76.7     | 61.1 | 40.0 | 10  | 75.2    | 55.6     | 0.0  | 0.0  | 3   | -       | -        | -    | 0   | 79.3     | 84.2     | 77.2 | 76  |
| vos          | 77.1   | 57.1     | 75.0 | 66.7 | 7   | 87.7    | 65.3     | 73.3 | 83.3 | 24  | -       | -        | -    | 0   | 81.6     | 71.3     | 63.2 | 29  |
| DM +FASTTEXT |        |          |      |      |     |         |          |      |      |     |         |          |      |     |          |          |      |     |
| svo          | 82.0   | 89.3     | 91.6 | 81.0 | 307 | 88.5    | 93.5     | 94.8 | 89.0 | 730 | 73.9    | 87.0     | 81.4 | 154 | 75.8     | 80.3     | 80.1 | 270 |
| ovs          | 72.5   | 72.2     | 52.2 | 31.0 | 42  | 85.8    | 72.1     | 74.2 | 76.9 | 74  | 66.9    | 58.9     | 76.9 | 13  | 71.2     | 71.3     | 76.0 | 50  |
| sov          | 77.0   | 88.9     | 73.3 | 55.6 | 12  | 78.7    | 83.3     | 76.2 | 0.0  | 8   | 77.0    | 83.3     | 66.7 | 2   | 75.6     | 70.0     | 85.0 | 20  |
| osv          | 83.2   | 90.5     | 61.1 | 33.3 | 7   | 86.0    | 76.6     | 70.8 | 60.0 | 27  | 79.5    | 100      | 100  | 1   | 70.6     | 100      | 100  | 2   |
| vso          | 71.5   | 70.0     | 44.4 | 33.3 | 10  | 72.1    | 33.3     | 100  | 0.0  | 3   | -       | -        | -    | 0   | 73.5     | 79.8     | 71.9 | 76  |
| vos          | 71.6   | 33.4     | 66.7 | 58.3 | 7   | 86.8    | 81.9     | 85.0 | 83.3 | 24  | -       | -        | -    | 0   | 77.8     | 64.4     | 85.1 | 29  |

**Table 4.12:** LAS on subsets of the *original treebanks*, each with differently ordered transitive constructions. This is evaluation on naturally occurring, unaltered sentences. Column # reports the construction counts for each subset.

omit the DM -LSTM +BERT from Table 4.12. and, likewise, will not consider it in the subsequent Section 4.5.

Note that given the insufficient coverage of many word orders (see construction counts reported in Table 4.12) it is impossible to draw reliable conclusions from these results alone. In addition, other factors beyond the word order, such as sentence complexity, may have also contributed to the performance differences – unlike in my dataset where all such factors are kept constant across different word order splits. Nevertheless, in Table 4.12 I also observe a variability in performance across word orders, which resembles that observed in Table 4.10. This provides additional reassurance for the validity of my approach.

## 4.5 Experiment II: Evaluation on ambiguous sentences

The second experiment targets **CONDITION II** from the introduction; i.e., it tests if the models’ performance is affected by the change of order of the core elements of a clause in *inflectionally ambiguous* sentences. In contrast to the evaluation in Section 4.4, where good performance equates to *no* difference across the word orders, here performance gaps between the SVO word order and other word orders are expected. Also note that, unlike

in the previous section, for the following evaluation there is no such thing as a good solution. It is not clear to what extent a model should rely on word order in the presence of ambiguity in a situation where the lexicosemantic signal is pulling its predictions in the opposite direction.<sup>24</sup>

Nevertheless, evaluating on the ambiguous data allows us to gain further insight into the inner workings of a model if the results are compared with those for the unambiguous sentences (Section 4.4). What we should expect from a well-performing model is performance no-better than in the unambiguous setting. Further, visibly worse performance on non-SVO clauses would be a good sign, since it would suggest that a model can identify instances of morphosyntactic ambiguity and, importantly, that ambiguity affects its parsing strategy. Lastly, comparing the results on ambiguous SVO clauses to those for unambiguous SVO clauses allows us to get insight into whether the *alignment of signals* has an affect on parsing performance.

With the above expectations in mind, I evaluate three of the parsers from Section 4.3 – those employing the full DM architecture – on the *ambiguous* counterfactual treebanks and present the results in the following Section 4.5.1.

### 4.5.1 Results

Because of the very low percentage of the ambiguous transitive clauses in the original treebanks (see Section 4.1.1), the ambiguous counterfactual treebanks contain many fewer sentences than the unambiguous alternatives. This translates to very few sentences in the development and test splits for many of these treebanks (see Table 4.5). For this reason, to avoid evaluating on too little data, I evaluate the models on the concatenation of the development and test splits, for each language.<sup>25</sup> While for Polish, Russian and Estonian the quantity of the resulting data (for the largest treebanks) is sufficient to draw initial conclusions from the results, this is not the case for Finnish. I do include Finnish results in all the subsequent tables, but I raise a note of caution that these results are very preliminary. Also note that similarly preliminary are the results for the Polish and Russian *iobj* relations, since the ambiguous sentences include only a few indirect object instances.

In Table 4.13 I present the full LAS breakdown across the different relations and word orders for the development+test split concatenations of the ambiguous counterfactual treebanks (see Table G.7, in Appendix G for standard deviation scores). What is most important here, is how this table compares to Table 4.10, from the previous section. For

---

<sup>24</sup>Indeed, I am not aware of any human studies that analyse the relative reliance on the two disambiguating strategies for case-marking languages. In the future, it would be interesting to evaluate human performance on the ambiguous split of the counterfactual dataset.

<sup>25</sup>To check that combining the development and test data does not influence the results in unpredictable ways, I also reproduce my results from Section 4.4 while evaluating on the dev+test splits. While I do not report these results here, the observed trends and absolute values closely resembled those from Section 4.4.

|              | Polish<br>47 clauses |       |      |      | Russian<br>168 clauses |       |      |      | Finnish<br>15 clauses |       |      | Estonian<br>48 clauses |       |      |
|--------------|----------------------|-------|------|------|------------------------|-------|------|------|-----------------------|-------|------|------------------------|-------|------|
|              | all                  | nsubj | obj  | iobj | all                    | nsubj | obj  | iobj | all                   | nsubj | obj  | all                    | nsubj | obj  |
| DM +BERT     |                      |       |      |      |                        |       |      |      |                       |       |      |                        |       |      |
| svo          | 91.5                 | 95.2  | 97.0 | 80.0 | 94.4                   | 97.8  | 97.9 | 100  | 84.1                  | 93.3  | 98.3 | 82.4                   | 87.5  | 95.8 |
| ovs          | 86.3                 | 77.7  | 57.1 | 60.0 | 88.9                   | 60.4  | 52.1 | 100  | 77.9                  | 53.3  | 51.6 | 76.8                   | 74.5  | 69.8 |
| sov          | 87.0                 | 91.5  | 73.2 | 50.0 | 90.7                   | 88.5  | 74.8 | 75.0 | 78.9                  | 73.3  | 51.7 | 74.8                   | 59.9  | 68.2 |
| osv          | 86.6                 | 81.9  | 74.4 | 40.0 | 90.0                   | 78.9  | 69.8 | 91.7 | 79.9                  | 85.0  | 75.0 | 73.9                   | 72.4  | 50.0 |
| vso          | 85.6                 | 51.0  | 75.0 | 45.0 | 90.6                   | 72.6  | 67.6 | 75.0 | 81.0                  | 61.7  | 90.0 | 79.1                   | 69.8  | 90.7 |
| vos          | 87.5                 | 58.5  | 92.9 | 65.0 | 91.5                   | 77.5  | 85.6 | 100  | 80.8                  | 40.0  | 93.3 | 75.6                   | 53.1  | 77.1 |
| DM +CNN      |                      |       |      |      |                        |       |      |      |                       |       |      |                        |       |      |
| svo          | 82.6                 | 91.0  | 94.0 | 50.0 | 89.3                   | 89.6  | 92.5 | 83.3 | 70.5                  | 78.3  | 83.3 | 71.8                   | 74.5  | 76.6 |
| ovs          | 72.4                 | 36.2  | 20.2 | 35.0 | 80.4                   | 29.5  | 28.8 | 100  | 61.3                  | 13.3  | 16.7 | 64.6                   | 38.0  | 33.4 |
| sov          | 72.7                 | 81.4  | 28.6 | 35.0 | 81.5                   | 65.5  | 42.8 | 58.4 | 61.2                  | 28.3  | 25.0 | 63.0                   | 31.8  | 30.7 |
| osv          | 70.6                 | 29.2  | 20.2 | 20.0 | 80.5                   | 37.5  | 40.5 | 58.4 | 62.4                  | 41.6  | 36.7 | 62.0                   | 45.3  | 13.0 |
| vso          | 74.2                 | 33.0  | 50.0 | 25.0 | 82.4                   | 34.1  | 42.5 | 50.0 | 63.7                  | 20.0  | 51.6 | 67.9                   | 36.0  | 70.3 |
| vos          | 75.8                 | 16.5  | 81.6 | 40.0 | 83.4                   | 25.9  | 77.6 | 50.0 | 65.3                  | 6.7   | 63.4 | 64.7                   | 27.1  | 45.8 |
| DM +FASTTEXT |                      |       |      |      |                        |       |      |      |                       |       |      |                        |       |      |
| svo          | 80.5                 | 79.3  | 88.1 | 50.0 | 86.0                   | 84.1  | 89.0 | 66.7 | 65.5                  | 70.0  | 71.7 | 65.4                   | 63.0  | 67.7 |
| ovs          | 71.8                 | 39.9  | 36.9 | 45.0 | 78.3                   | 30.1  | 35.0 | 33.3 | 57.3                  | 8.3   | 23.3 | 59.6                   | 33.3  | 25.5 |
| sov          | 72.2                 | 68.1  | 56.0 | 40.0 | 79.4                   | 57.9  | 49.1 | 66.7 | 59.5                  | 43.3  | 30.0 | 58.2                   | 21.9  | 41.1 |
| osv          | 68.7                 | 38.3  | 38.1 | 20.0 | 78.5                   | 41.2  | 42.1 | 25.0 | 59.1                  | 46.7  | 21.6 | 57.5                   | 34.4  | 26.1 |
| vso          | 72.2                 | 42.0  | 59.5 | 25.0 | 76.7                   | 28.1  | 44.7 | 25.0 | 57.9                  | 18.3  | 41.7 | 59.0                   | 27.1  | 49.0 |
| vos          | 73.1                 | 22.9  | 69.7 | 50.0 | 78.8                   | 32.3  | 64.1 | 66.7 | 59.4                  | 10.0  | 48.4 | 57.7                   | 17.7  | 37.5 |

**Table 4.13:** LAS scores on the dev+test splits of the *ambiguous* counterfactual treebanks.

|                                              |     | Polish |       |      |       | Russian |       |      |       | Finnish |       |       | Estonian |       |       |
|----------------------------------------------|-----|--------|-------|------|-------|---------|-------|------|-------|---------|-------|-------|----------|-------|-------|
|                                              |     | all    | nsubj | obj  | iobj  | all     | nsubj | obj  | iobj  | all     | nsubj | obj   | all      | nsubj | obj   |
| svo attachment bias metric (LAS)             |     |        |       |      |       |         |       |      |       |         |       |       |          |       |       |
| BERT                                         | UA. | 5.1    | 13.6  | 11.3 | 12.1  | 2.3     | 6.9   | 5.9  | 5.3   | 2.9     | 7.8   | 4.4   | 4.6      | 8.6   | 10.8  |
|                                              | A.  | 4.9    | 23.0  | 22.5 | 28.0  | 4.0     | 22.1  | 27.9 | 11.7  | 4.4     | 30.6  | 26.0  | 6.4      | 21.6  | 24.7  |
| CNN                                          | UA. | 17.4   | 47.6  | 41.0 | 28.9  | 12.4    | 43.2  | 28.1 | 11.8  | 18.4    | 35.7  | 45.7  | 12.7     | 29.4  | 33.5  |
|                                              | A.  | 9.5    | 51.7  | 53.9 | 19.0  | 7.7     | 51.1  | 46.1 | 26.7  | 7.7     | 56.3  | 44.7  | 7.4      | 38.8  | 37.9  |
| FT.                                          | UA. | 15.2   | 28.6  | 30.3 | 37.6  | 16.6    | 34.7  | 35.0 | 34.1  | 23.1    | 44.6  | 39.9  | 17.2     | 28.3  | 33.3  |
|                                              | A.  | 8.9    | 37.0  | 36.1 | 14.0  | 7.7     | 46.2  | 42.0 | 23.4  | 6.9     | 44.6  | 38.7  | 7.0      | 36.2  | 31.9  |
| svo LAS difference (unambiguous - ambiguous) |     |        |       |      |       |         |       |      |       |         |       |       |          |       |       |
| BERT                                         |     | ↓2.7   | ↓4.6  | ↓1.3 | ↓8.5  | ↓3.0    | ↓1.8  | ↓1.1 | ↑4.3  | ↓13.4   | ↓6.5  | ↓0.5  | ↓11.5    | ↓10   | ↓1.4  |
| CNN                                          |     | ↓3.5   | ↓2.5  | ↑4.9 | ↓22.5 | ↓2.9    | ↓2.3  | ↑0.9 | ↑15.8 | ↓15.8   | ↓14   | ↓3.3  | ↓12.8    | ↓13   | ↓7.1  |
| FT.                                          |     | ↓5.7   | ↓9.4  | ↓0.3 | ↓27.9 | ↓7.1    | ↓7.8  | ↓6.4 | ↓21.6 | ↓17     | ↓18.8 | ↓12.8 | ↓14      | ↓17.2 | ↓16.7 |

**Table 4.14:** Top: svo attachment bias metric (LAS) for the unambiguous counterfactual treebanks (UA.) (see Section 4.4) and the ambiguous counterfactual treebanks (A.) (this section). Bottom: The difference between the svo LAS on the *unambiguous* counterfactual treebanks and the svo LAS on *ambiguous* counterfactual treebanks.

convenience of spotting the important trends, in Table 4.14 I summarise the key differences between the two sets of results. Note that because the ambiguous and the unambiguous versions of the counterfactual dataset differ in more factors than just morphosyntactic ambiguity, the differences between the individual LAS values could be attributed to more than just the presence/absence of ambiguity. However, comparing the *patterns of performance* emerging from the two types of evaluation can lead to meaningful insights.

**Differences across word orders** One of the key insights emerging from this evaluation is that the models rely on word order substantially more when parsing ambiguous clauses. The differences between performance on the SVO and the other word orders is much larger in the ambiguous setting – as captured in the SVO attachment bias scores, presented in Table 4.14 (top). When directly comparing Table 4.10 to Table 4.13, I notice that the prominent performance drops uncovered in the unambiguous evaluation remain and are accentuated in Table 4.13 – consider, for example, the Polish BERT-based model’s performance on the VSO and VOS clauses or the performance drop of the Estonian models for the OSV object. But there are also some new patterns, where a performance drop can be seen for word orders that were little affected before. One such pattern is the LAS deterioration on the OVS. In unambiguous evaluation OVS was often the second best performing order, after SVO. In Table 4.13, it is often one of the most challenging orders.

All of the above constitutes a very interesting result. The fact that the models perform worse in the ambiguous setting is a good sign, in the context of this thesis, as it signals that they are (at least to some extent) sensitive to ambiguity and that they learn to lean on word order more when it is detected. The observed differences in performance also point to the fact that the presence of unambiguous morphology does pull the models away from relying on word order. Consider, for example, the aforementioned LAS drop on OVS word order noted in the ambiguous setting. The fact that such a big drop does not occur for this word order (which closely resembles SVO, given the verb in the middle) in the unambiguous setting means that in many cases the models *can* and often do use morphology as a signal to meaning. And they do so despite never being directly pointed towards morphology, neither during the pre-training (for BERT and FASTTEXT) nor during training on the parsing task. Also notice how this result suggests the models largely rely on morphology, and not just lexical semantics, when parsing the OVS unambiguous clauses. If the reverse was true, the drop noted for OVS in the ambiguous evaluation would not be as prominent. Although, it is important to point out that lexical semantics must still have a strong effect; especially for the BERT-based models, since their performance on the non-SVO *ambiguous* clauses is larger than 70 LAS for many relations.

The above paragraph painted a rather optimistic picture. To balance it out, I find it important to note that if the models generalised in a truly human-like way, we would observe the performance differences across word orders *only* in the ambiguous condition. And while the worse performance in the ambiguous setting suggests the models pick up traces of the correct generalisation, including the conditional reliance on word order, the poor performance in the unambiguous setting means they are not quite there yet. However, given that a lot of important trends are at least partially captured, especially for BERT, perhaps the models only have to be nudged in the direction of the correct generalisation to improve their competence on the task. This is a direction I explore in Chapter 6.

**Differences in SVO performance** Experiments from Section 4.4 revealed that the models perform best on clauses which adhere to the dominant SVO order. I now ask whether the *cooperation* of the word order and morphological signals has an effect on how well the models perform on such clauses. To see whether this is the case I compare the results on the unambiguous SVO clauses from Table 4.10 to that on the ambiguous SVO clauses from Table 4.13. I report the LAS differences between the two in Table 4.14 (bottom).

Table 4.14 clearly demonstrates that the ambiguity has a negative effect on the models’ performance, even on the SVO clauses. For the vast majority of relations, I register a notable drop of LAS in the ambiguous clauses; for some relations as high 10+ LAS points. The only instances where the results go up is the Polish and Russian object relations for the CNN model, and the Russian indirect object relations for the CNN and BERT (note that the latter result on *iobj* is inconclusive, given the low frequency of those relations in the ambiguous data). This result points to the models benefiting from signal cooperation – the performance is best when the cues align, signaling the same interpretation.

This result is preliminary, given (i) that the LAS results from the two tables are not directly comparable<sup>26</sup> and (ii) the limited size of the ambiguous treebanks. However, it may point to the models learning to rely primarily on cue *gestalts*, which bares resemblance to the comprehension strategies developed by children at the earliest stages of language acquisition (Chan et al., 2009, Ibbotson and Tomasello, 2009, Krajewski and Lieven, 2014, MacWhinney and Bates, 1989) – a tendency discussed in Section 2.5.2.

## 4.6 Conclusion

In Section 4.1.3 I hypothesized that despite the strong signal coming from morphology, the easy access to word order features and word order’s correlation with grammatical function can push the models to rely on word order over morphology. My experiments from Section 4.4 suggest that, indeed, this often seems to be the case. Evaluation on the unambiguous counterfactual treebanks revealed that for every language, all four parsers over-rely on word order as a cue to core relations, over-fitting to the most frequent SVO. These results are the more meaningful, given that, as proposed in Section 4.2, my counterfactual alteration did not alter lexical semantics. This means that in the data word order is in direct competition with not just one, but two collaborating signals – morphology and lexical semantics. Despite this, the models still over-rely on word order. For CNN-based and FASTTEXT-based models this tendency is very strong – for many non-SVO word orders the performance on the core relations falls below 50 LAS, meaning

---

<sup>26</sup>As noted before, the absolute LAS values are not comparable across the ambiguous and the unambiguous settings because the versions of the dataset differ in more than just morphosyntactic ambiguity.

that in the majority of test instances word order signals overpower morphology + lexical semantics. This is in contrast to BERT-based models, the performance of which never falls below 67 LAS, for any of the core relations.

On the other hand, in my experiments on ambiguous clauses (Section 4.5), I find traces of correct linguistic generalisation, with the models conditionally increasing their reliance on word order, in the presence of morphosyntactic ambiguity. Together with the unambiguous results, this paints a nuanced picture of the models' generalisations. BERT-based models, in particular, seem to 'recognise' the role of morphology as an important signal to meaning. However, this tendency is inconsistent, with the word order signals regularly overpowering the correct generalisation.

Results from this chapter have important theoretical and practical implications. Despite some correct tendencies, none of the models develops the linguistic generalisation which is a pre-requisite to learning the grammar of a language. This has direct consequences for the robustness of these models; while morphology remains consistent across different genres and domains in all four languages, word order distribution vary (Dyakonova, 2009, Makarchuk, 2019). Consequently, relying on the latter over the first can lead to severe performance degradation in data that does not follow the word order distribution of the UD treebanks.

On a final note, while all of the above insights are important, the main contribution of this chapter was the evaluation methodology, rather than providing an exhaustive account of models' generalisations. Consequently, results from this chapter are just a foundation for further investigations into neural parsers morphological competence. One question that remains to be answered is whether it is morphology that the models rely on when they make correct predictions on the non-SVO word orders or whether it is lexical semantics. This is something I address in my next chapter, where I deepen my investigation of models' reliance on morphology, by accounting for lexicosemantic effects in my counterfactual evaluation.

---

## ACCOUNTING FOR LEXICAL SEMANTICS

---

In this chapter I build on the word order experiments from Chapter 4 by accounting for lexical semantics. Similarly to the previous chapter, my goal here is to test whether the models rely on lexicosemantic cues in situations where they should rely on morphology instead. Specifically, I ask the following question: do models’ generalisation strategies resemble those of native speakers, who rank morphology higher than lexical semantics as a signal to core sentence meaning. I attempt to answer this question through a series of experiments in which I manipulate the *lexemes* in UD sentences, to study the lexicosemantic influence on models’ predictions. Importantly, I only do so at evaluation, to test the generalisations developed by models trained on natural, unaltered data. Intuitively, if a model correctly relies primarily on morphology, alterations of lexemes within the test data should have little effect on its performance. As in Chapter 4, I experiment on two Slavic and two Finnic languages: Polish, Russian, Finnish and Estonian, and target (primarily) neural models from Section 4.3, which employ the full architecture of Dozat and Manning (2017) (DM).

The chapter begins with Section 5.1, where I propose and advocate for a new type of data perturbations – **lexeme-based perturbations** (see Figure 5.1), discussing how it differs from alternative lexically-oriented perturbation approaches proposed in earlier work. In the same section, I also introduce a new framework – LEXMIX – which provides the core support for the proposed lexeme-based alterations, facilitating all my experiments in this chapter.

The discussions of the new perturbation style and the framework are followed by three experimental sections (5.2, 5.3, 5.4) – each concerning a different experiment. In the first two experiments, I keep the word order unaltered and experiment with two different lexeme-based perturbations. The first perturbation (Section 5.2) involves permuting the lexemes of nouns in the UD corpora. Note that changing exclusively lexemes of *nouns* breaks important lexical relationships between content words in a language – including

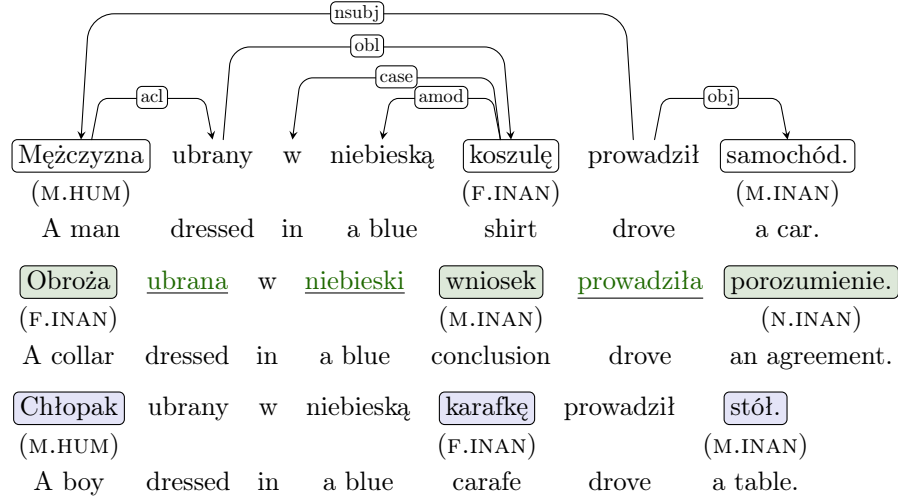


**Figure 5.1:** Lexeme-based perturbations of nouns (green, middle) vs form-based noun perturbations (blue, bottom); a Polish example. The first allow for the original nouns to be replaced by nouns of other gender and animacy.

those discussed in Section 2.4.5.2 – while maintaining grammaticality of the sentences (more on that in Section 5.1.3). This is in contrast to perturbing verbs which breaks argument structures, yielding ungrammatical sentences. The second perturbation (Section 5.3) is more targeted and involves swapping the lexemes of the subject and the object in each transitive clause with two core noun arguments. Note that, when the core relations are considered, this results in more challenging test cases, since it ensures that the lexical cues to subject/objecthood are false friends, pulling in the opposite direction to the correct assignment. In both of these experiments, I expect a morphologically competent model, as defined in Section 2.7, to be little affected by the lexical changes.

Finally, in Section 5.4, I bring together both the word order and the lexical alterations. I do so by applying the argument swapping perturbation, introduced in Section 5.3, to all differently ordered treebanks in the counterfactual dataset from Chapter 4. Combining the two perturbations puts morphology in direct opposition to both word order and lexicosemantic cues, posing it as the only linguistic signal that points out the correct subject and object. As a consequence, this makes the subject/object prediction task close to unsolvable for models which make little to no use of morphology, as I demonstrate in Section 5.4.1 by evaluating the morphologically-blind baselines introduced in Section 3.3 on the resulting data. Finally, in the last Section 5.5, in addition to evaluating the models from Section 4.3 (based on BERT, CNN and FASTTEXT), I also train and evaluate a range of alternative transformer-based models, ultimately demonstrating that neural models of all shapes and sizes fall short of true morphological competence. All models I consider under-rely on inflectional morphology as a signal to core relations.

## 5.1 LexMix: A framework to alter lexemes in UD treebanks for highly inflected languages

In this section I propose a novel way of perturbing language data to study the extent of a neural model’s reliance on lexicosemantic features (Section 5.1.1). I also propose a new framework that enables carrying out such perturbations within the UD treebanks for highly inflected languages (Section 5.1.2).

### 5.1.1 Lexeme-based vs form-based perturbations

The **lexeme-based perturbations** I propose in this chapter differ in an important way from what I refer to as **form-based lexical perturbations** employed in relevant related work (Gulordava et al., 2018, Hall Maudslay and Cotterell, 2021, Kasai and Frank, 2019, Lasri et al., 2022). The latter involve substituting inflected forms occurring in the corpus with forms belonging to a different lexeme but with *the exact same morphosyntactic description* as the originals. This means that they disallow replacing, for example, feminine noun forms with forms of masculine nouns (see Figure 5.1). For weakly inflected languages, like English, this approach has no clear drawbacks, but it is ill suited for experimentation on more inflected languages (in particular those with grammatical gender and/or animacy). This is because it skews the perturbation towards assignments in which:

- (i) adjectives used primarily for description of feminine/masculine/neuter nouns keep modifying such nouns (i.e., there is no gender variation),
- (ii) verb lexemes’ tendencies of taking subjects of particular morphosyntactic description (e.g., of specific genders/animacy) is maintained, and
- (iii) if alternative forms are drawn from the original corpus (as in e.g., Gulordava et al. (2018) and Kasai and Frank (2019)), noun lexemes rarely occupy atypical roles – lexemes which in the corpus take predominantly accusative case, will unlikely fill subject roles, which require nominative case.

In other words, form-based perturbations only break *some* of the lexical relationships, while maintaining other correlations.

In contrast, the lexeme-based mechanism I propose is free from such limitations and has an additional advantage of being more flexible, allowing for a wider variety of perturbations. It decouples the lexeme and lexeme-specific features (e.g., gender and animacy for Polish/Russian nouns) from the syntactic slot filled by the original form, manifested in the agreement features. For each slot, a new lexeme is selected based exclusively on the coarse-grained POS of the original form. Next, an appropriate form of

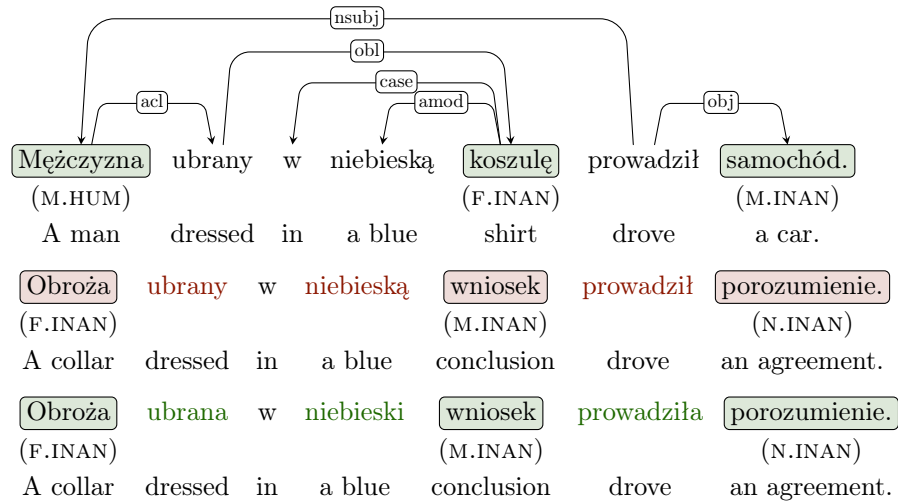


**Figure 5.2:** An example of lexeme changes (row 2) which trigger the reinflection (row 3).

the new lexeme is chosen from its complete paradigm to fit the morphosyntactic context of the original token. Note that, unlike the form-based alternative, this type of perturbation does not preserve the values of lexeme-specific features, which are often involved in agreement (like gender and animacy for Polish/Russian nouns). This necessitates a follow-up step of *agreement maintenance*, where word forms that agreed with the original token are *reinflected* to maintain grammaticality (see Figure 5.2 for an example).

Because of the reinflection step, the lexeme-based perturbations involve global, sentence level, rather than local operations (which stands in contrast to the form-based approaches). This, admittedly, makes them harder to implement. To facilitate all my experiments from this chapter, I propose a new framework, LEXMIX, which provides (i) the core support for lexeme changes within UD treebanks and (ii) easy integration of new methods of altering the data. I discuss it in more detail in the next subsection.

### 5.1.2 The LexMix framework

This subsection provides an overview of a new framework for carrying out lexeme-based perturbations in UD treebanks for highly inflected languages. For now, it only supports changing lexemes of content POS – i.e., nouns, verbs, adjectives and adverbs – within the four languages considered in this thesis, but it should be possible to straightforwardly extend it to cover more languages. Within the framework one can mix lexemes already present in a treebank, both across all sentences, as well as within a single sentence. It also provides a functionality to mix in novel lexemes, which do not appear in the treebank. This allows for easy integration of various experimental settings, including but not limited to those covered in this chapter.

**Paradigm record** Lexeme-based perturbations require access to the full paradigms of all lexemes with the POS involved in the alteration – i.e., the POS that is being mixed, as well as all POS that might be involved in agreement with that POS. For Polish this includes full paradigms for nouns, adjectives, participles, verbs, determiners and numerals. All this information is kept within the LEXMIX in a **paradigm record** – a record of paradigms for all relevant POS in the treebank for a given language, with *all* slots filled with forms. This record is constructed before any alterations take place and it takes shape of a nested dictionary, which maps lemma+POS instances to a dictionary mapping morphosyntactic tags to forms. The construction of this record is described in the following paragraph.

My goal was to create data which is as clean as possible, with minimal amount of reinflection errors. To achieve this, I construct a paradigm record for each language in a four step process, which prioritises reliable sources with *gold* inflected forms. The first step involves identifying all relevant lexemes present in a considered treebank (all splits) and creating a partially-filled record with all forms that occur in that corpus. Next, I use UniMorph<sup>1</sup> (Kirov et al., 2016, McCarthy et al., 2020) – a resource comprised of gold inflectional word paradigms for 107 languages – to fill as many gaps in the record’s paradigms as possible. Third, to fill the remaining gaps, I scrape the paradigms directly from Wiktionary.<sup>2</sup> Finally, I fill the slots which are neither in the corpus, nor retrievable from UniMorph or Wiktionary with probabilistic inflection models trained on UniMorph for a considered language (a separate model for each POS). For determiners and inflected numerals, I skip the automatic inflection stage, since Wiktionary is fairly complete when it comes to closed class words.

At the last, reinflection step I make use of the probabilistic inflector of Wu and Cotterell (2019), which obtained single-model state of the art in the SIGMORPHON 2019 Shared Task (McCarthy et al., 2019b). The model has a latent character-level monotonic alignment between the source and the target inflected forms which is jointly learned with a character-level transducer. I train it to output an inflected form given a lemma and a morphosyntactic tag.

Note that while it is possible to get complete paradigms based on the reinflection alone (i.e., omitting steps two and three), they would likely be of a lower quality, since the reinflection model is bound to make some errors. By retrieving as many forms as possible from hand-crafted resources I keep the error count to minimum. Even further, I additionally build in a functionality for checking/fixing the forms obtained from the reinflection model at the final paradigm filling stage, as well as a functionality to provide missing forms directly from a human annotator, instead of using a reinflection model. I used a combination of those methods for filling in the Polish language paradigm record. As

---

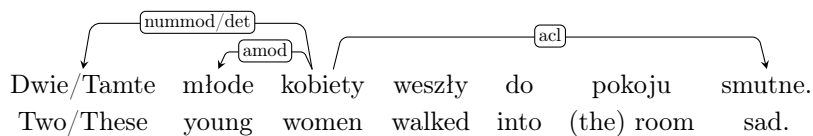
<sup>1</sup><https://unimorph.github.io/>

<sup>2</sup>Although UniMorph is based on Wiktionary it only contains a subset of lexemes present in Wiktionary.

a result, by construction, Polish lexeme-perturbed data used for experiments in Sections 5.2–5.4 contains few to none reinflection errors (which was also confirmed by manual review of the data).

**Lexeme change** Let  $t$  be the altered UD token,  $l_o$ ,  $l_n$  be the original pre-change lemma and the new lemma, and  $m_o$  the pre-change morphosyntactic tag. Both  $l_o$  and  $m_o$  are provided for each token in the UD morphosyntactic annotation (see Section 2.1.1). The new morphosyntactic tag  $m_n$  is determined by setting the values of all lexeme-specific features based on  $l_n$  and the values of all other inflectional features based on  $m_o$ . While the latter can be easily retrieved from UD annotation, the values of lexeme specific features for  $l_n$  are retrieved pre-alteration and recorded in the paradigm-record. Which features belong to which category depends on the considered language and POS. For example, for Polish and Russian nouns, lexeme-specific features include gender and animacy and other inflectional features include case and number. Once  $m_n$  is obtained, the appropriate form for  $l_n$  is retrieved from a pre-constructed **paradigm record** (see previous paragraph).

**Head and modifier reinflection** After the lexeme is changed and a form is selected to fit the original token’s position, the sentence is fixed to maintain grammaticality. This involves maintaining agreement in lexeme-specific features which might have changed during the substitution. For the Slavic language pair, the only POS for which lexeme changes can break agreement is a noun, while in the two Finnic languages no agreement involves lexeme-specific features. The modifier candidates for reinflection are determined by considering a language-specific set of relations involved in agreement with the changed POS. For Polish and Russian nouns these include: numeric modifiers (*nummod*), determiners (*det*), adjective modifiers (*amod*) and clausal modifiers of a noun (*acl*).<sup>3</sup> Below I demonstrate example uses of those relations:



Once the candidates for reinflection are identified, an additional check takes place, where they are tested for agreement with the pre-change word form. Those which agree with the form in the lexeme-specific features are selected for reinflection, which involves changing the relevant features to match the new lexeme and selecting an appropriate form from the pre-constructed paradigm record. A similar procedure is conducted to reinflect (i) the head verb, if linked through the *nsubj* relation, (ii) other verbs linked to that head

<sup>3</sup>For Polish I also consider auxiliaries (*aux*) to account for misannotations, where a copula verb ‘być’ is misattached to the subject as an auxiliary, instead of being attached to the nonverbal predicate as a copula.

verb with a *conj* or *aux* relation which agree with the pre-change noun and (iii) copula verbs, which are linked as modifiers of the nonverbal predicate heading the changed token.

If any of the above changes are unsuccessful (due to missing forms) all changes, including the initial lexeme change are reversed. While this is very rare due to the comprehensive filling of the paradigm records, it can happen when key features which determine the form of agreeing tokens are missing from the token/lexeme annotation – e.g., the gender is unknown in Russian/Polish.

### 5.1.3 Comparison with other work

Most existing work which studies the lexicosemantic effects on neural models’ performance is exclusive to English. For agreement prediction task (see Section 3.4.2), Wei et al. (2021) build an evaluation dataset based on 56 sentential templates, with missing subject and verb slots, which they fill with pre-selected noun and verb forms. Lasri et al. (2022) also focus on agreement prediction task and similarly create their test examples based on templates (they use templates from Marvin and Linzen (2018)), by filling them with pre-selected forms. Hall Maudslay and Cotterell (2021) experiment with *jabberwocky* sentences in dependency-parsing-based probing, replacing original content words in a treebank with pseudo-words that match the morphosyntactic description of the original form. Another work which considers lexical mixing for dependency parsing is that of Kasai and Frank (2019). The authors consider an alteration in which they replace each word in a UD treebank with a randomly chosen word with the same POS and morphosyntactic description. One of the few non-English centric works is that of Gulordava et al. (2018) on agreement prediction task, which considered four languages: English, Italian, Hebrew and Russian. To create their mixed, nonce sentences, they followed a procedure equivalent to that of Kasai and Frank (2019), substituting each content word in a UD treebank with a randomly chosen content word from the same treebank that matched the original’s POS and morphosyntactic description.

Note that all of the above are form-based – mixing in word form from a pre-determined list. When this list is obtained based on the original treebank, as in Gulordava et al. (2018) and Kasai and Frank (2019), it is unlikely to cover full paradigms, which in turn is associated with a number of issues (see Section 5.1.1). The framework I proposed in this section is free from such limitations. Further, mixing *all* content words, including verbs, as is commonly done (Gulordava et al., 2018, Hall Maudslay and Cotterell, 2021, Kasai and Frank, 2019) breaks the grammaticality of resulting sentences, since some verbs are unavoidably placed in ill-suited contexts which do not match their valency. In the experiments I conduct in this chapter, I avoid mixing verbs to keep the sentences grammatical.

|                                                                                                                                                                                                |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Polish                                                                                                                                                                                         |
| Biała rybka stoi w tragedii z rozpostartymi pasażerami. → A white little fish is standing in a tragedy with spread passengers.                                                                 |
| Nos w krótkich spodenkach i piosenkarzu zjeżdża na zastępcach po wodzie. → A nose in shorts and in a singer is sliding down on the replacements across the water.                              |
| Strój w przedsiębiorstwie stoi obok Eugeniusza innych ciągów wchodzących na różowe jedzenie. → An outfit in an enterprise stands next to Eugene of other strings which enter on the pink food. |
| Russian                                                                                                                                                                                        |
| Ранее часто писали “фантастика”, сейчас такое результат используется редко, но, тем не менее, имеет РСПП (например, Нормальный штаб Основы).                                                   |
| Очень распространенным было лицо о греческой стоимости проявления.                                                                                                                             |
| В этом же альпинисте покорен Луксор немецкой командой Кóшкойprzest. кóшкой Состоянием (Hermann Buhl).                                                                                          |
| Finnish                                                                                                                                                                                        |
| Arviointi, Vuosi ja Toyota juoruilivat keskellä kauppaä täysin häpeilemättä.                                                                                                                   |
| Nyt lähden kuvailemaan Enemmistön turnaukset ja siivoamaan kappaleen jälleen isään.                                                                                                            |
| Paistetu osoite passaa hyvin vaikkapa kauhan, paitojen tai vaikka kunnian satuometsäksi.                                                                                                       |
| Estonian                                                                                                                                                                                       |
| Teiseks avaldub Aru autorites dementsuse DeCSSi ja Karil üleominekupajasjärku filosoofiline Bogani.                                                                                            |
| Need 18 lehmät on saadetud Pillekriin Lülisambal, kutsudes Tähte tagasi Fellinil hemorraagiat kandma oma töö, produkt ja Bronpino eest.                                                        |
| Kui sul on juba immigrant ja loodejõud monstrumil, hakkab meditsiini kirjandus arvule.                                                                                                         |

**Table 5.1:** Example sentences resulting from a noun lexeme permutation. English translations are provided for Polish examples (grey).

## 5.2 Experiment I: Noun lexeme permutation

My first experiment involves evaluating neural parsers from Section 4.3 (those which employ the full DM architecture) on UD data with *permuted* noun lexemes (see Table 5.1 for an example). Note that this alteration maintains the original frequencies of the lexemes – the only changed variable is their placement. This allows me to decouple the effects of words’ lexical content from lexeme frequency effects, standing in contrast to the alternative, random sampling of each new lexeme with replacement. I mix exclusively nouns because this (i) is sufficient to break important lexical relationships between content words in a language (adjective-noun, verb-noun) and (ii) yields grammatical sentences. For the permutation, I consider all common nouns and proper nouns in a given treebank split, except those linked via the *fixed* relation (mixing nouns linked via *fixed* relations would break fixed expressions) and plural-only nouns (e.g., Polish *okulary*), which cannot fill the slots originally occupied by singular nouns.<sup>4</sup>

<sup>4</sup>At the moment, the framework only reinflects based on the *lexeme-specific* features, although it could be easily extended to reinflect based on the changed number if a new lexeme is plural-only.

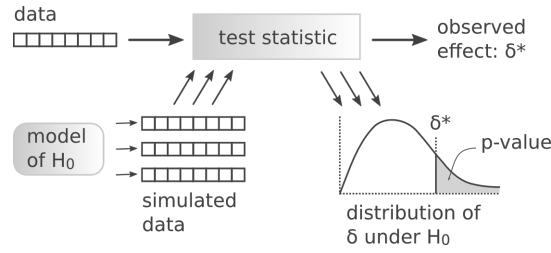


**Figure 5.3:** Simulation method framework for hypothesis testing. Figure from Downey (2016).

### 5.2.1 Simulation-based experimental design

Note that unlike for the word order alterations (Chapter 4), where there were only 6 possible versions for each sentence, the space of all possible noun lexeme permutations is too large to consider each possible one. One solution is to consider only selected few, as was done, for example, in Gulordava et al. (2018), and report the average performance. However, this presents only a limited picture of lexicosemantic effects and is associated with risks of evaluating on particularly challenging/easy lexical assignments. Here, I propose a novel type of evaluation, free from such shortcomings.

Inspired by *simulation methods* in significance testing, I design my experiments to test a null hypothesis which states that the placement of noun lexemes within a treebank has *no effect* on a neural model’s performance. To test this hypothesis I follow the steps akin to those used in simulation methods in significance testing (pictured in Figure 5.3). These steps include:

- (i) measuring the labeled attachment score (LAS), which I select as a test statistic, for the original dataset
- (ii) evaluating the model on  $n$  simulated datasets, each with differently permuted noun lexemes, to obtain a distribution of LAS under the null hypothesis
- (iv) calculating the probability of obtaining a score as good or better than the LAS recorded for the original treebank based on the said distribution

The probability from the last step corresponds to the  $p$ -value in statistical hypothesis testing and quantifies the likelihood of a result as extreme as the one originally observed given that there is no link between the placement of noun lexemes and the model’s performance.

In all my experiments, I consider  $n = 500$ ; i.e., I evaluate the models on 500 different versions of the original treebank with atypical lexical semantics.<sup>5</sup> When reporting the results I always present distributions for LAS scores obtained for all these datasets, in addition to making note of  $p$ -values obtained for each experimental setting.

<sup>5</sup>I do not include the original treebank in the 500 versions and always permute all qualifying noun lexemes.

### 5.2.1.1 Not-so syntactic relations

As in Chapter 4, my focus in this chapter is on core relations. For this reason, for my primary results I consider LAS calculated exclusively on *nsubj*, *obj* and *iobj* in transitive clauses. But since the permutation perturbation also concerns other relations, to complement the core results and put them in perspective, I also report LAS distributions for other relations.

While considering the non-core UD relations I address the fact that some of them are not purely syntactic (see Section 2.1.1). At least not in the sense of Chomsky, who claimed that “grammaticalness cannot be identified with meaningfulness” (Chomsky, 1957, p. 106). Some of those relations, such as *flat*, *fixed* and *compound* are designed specifically for *lexical phenomena*. Others, although not exclusively lexical, can be difficult or impossible to distinguish from other relations without access to lexical content. Consider for example *appos* and *nmod* – both used for nominal dependents of another noun. The first corresponds to an attribute of the modified noun or its genitive complement, while the second “serves to define, modify, name, or describe that noun”<sup>6</sup>. These two are impossible to distinguish given randomised lexical semantics.

Given this lexical dependence, it is unreasonable to expect a model will perform equally well on such ‘lexically influenced’ relations after a lexeme perturbation. Consequently, evaluating on *all* relations is ill-suited for the lexically perturbed data, since it obscures the true picture of a model’s performance. To avoid such unfair evaluation and gain better insights, I separate the set of UD relations into two categories – (i) one that covers relations that are likely to be affected, like *flat* or *nmod* (i.e., this is the expected behaviour even for a well performing model) and (ii) one covering relations that should not be affected – and evaluate on each category separately. I present the categorisation in Table 5.2; for more information about the rationale behind individual, per-relation decisions see Appendix F.

To my knowledge, I am the first to consider lexical effects in UD annotation while evaluating lexicosemantic effects in UD parsing. Related work typically considers *all* relations, including the more lexical ones (Hall Maudslay and Cotterell, 2021, Kasai and Frank, 2019) which introduces noise and distorts the results.

### 5.2.1.2 Data and models

I experiment on three models from the previous chapter which employ the full DM architecture of (see Section 4.3) and evaluate on the largest available UD treebank for each language – Polish PDB (Wróblewska, 2018), Russian SynTagRus (Droganova et al., 2018), Finnish TDT (Haverinen et al., 2014) and Estonian EDT (Muischnek et al., 2014). To reduce computational costs, for each [language, input] combination I consider only one version of a model, trained with the random seed set to 1.

---

<sup>6</sup><https://universaldependencies.org/u/dep/appos.html>

| Should not be affected (SA)                                                                                                                                                                                                                                                                                                                                                                                                                                             | Likely to be affected (LA)                                                                                                                                  |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| punct, root, aux, aux:clitic, aux:cnd, aux:pass, aux:imp, advmod, advmod:neg, advmod:arg, compound:prt, list, csubj, ccomp, xcomp, expl:pv, nmod:pred, case, det, det:poss, det:numgov, det:nummod, cop, mark, advmod:emph, amod, discourse, discourse:emo, discourse:intj, conj, cc, cc:preconj, parataxis, parataxis:obj, parataxis:insert, obl:arg, fixed, nummod, nsubj:cop, xcomp:ds, advcl, acl, acl:relcl, csubj:cop, xcomp:ds, xcomp:pred, nsubj:pass, obl:cmpr | orphan, amod:flat, flat, flat:name, flat:foreign, compound, compound:nn, obl, obl:agent, nmod, nmod:arg, nmod:flat, nmod:poss, nmod:gobj, nmod:gsubj, appos |

**Table 5.2:** Lexical categorisation of UD dependency relations.

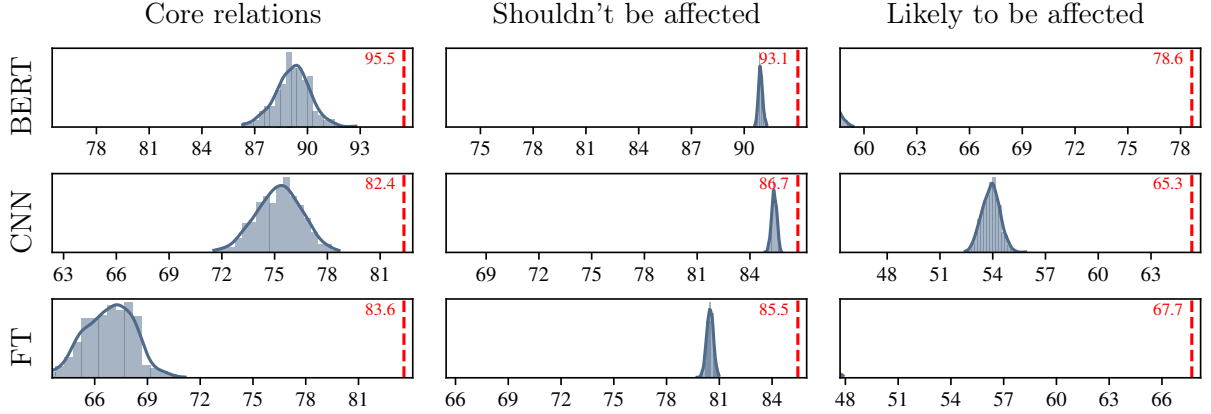


**Figure 5.4:** Polish: LAS distribution on data with permuted noun lexemes (500 runs). Note that the FASTTEXT results for the LA relations fall out of the  $[LAS_o - 20, LAS_o + 0.5]$  range (see text).

## 5.2.2 Results

### 5.2.2.1 LAS distributions

I present the results for Polish, Russian, Finnish and Estonian in Figure 5.4, Figure 5.5, Figure 5.6 and Figure 5.7, respectively. Each figure presents the distribution of LAS on the perturbed versions of the development data for: (i) core relations in transitive clauses, (ii) relations that should not be affected (SA; see Table 5.2) and (iii) relations that are likely to be affected (LA). I mark the LAS scores measured on unaltered corpora ( $LAS_o$ ) with red, dashed lines and, for each plot, I set the range of the x axis to  $[LAS_o - 20, LAS_o + 0.5]$ . The latter is done to ease the comparison of the observed decrease in LAS across the plots. Note that it is this drop – from LAS on the original data to LAS on the perturbed data – that is of primary interest here, rather than the absolute LAS performance. The better model is the one that exhibits a smaller performance drop on the core relations and SA relations in the face of a lexeme permutation.

**First insights** Across all experimental conditions (language x input) I observe a notable decrease in performance when comparing LAS for core relations on perturbed vs non-

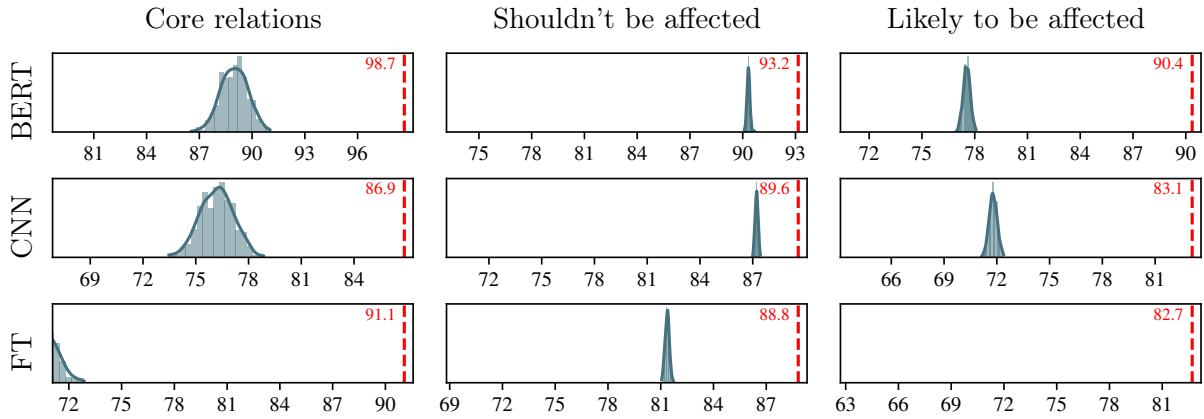


**Figure 5.5:** Russian: LAS distribution on data with permuted noun lexemes (500 runs).

perturbed data. The prominence of this drop depends on the language and the type of input – the average drop (across the 500 runs) ranges from 6.3 LAS for Polish BERT-based model to 21.1 for Russian FASTTEXT-based parser. None of the perturbed datasets yields performance equal to or larger than their respective  $LAS_o$ , which means that  $p = 0$  for each condition. While the roundness of this result is largely due to the limited number of runs (500), given the large shift of the distributions with respect to  $LAS_o$ ,  $p$  would likely be close to 0 even with increased number of runs.

As expected, LA relations are particularly challenging. In all conditions, the average LAS drop is  $> 10$  and in some cases it is so large that the results do not fit on a plot (see e.g., Finnish BERT or FASTTEXT). SA relations are affected to a much lesser extent – for many conditions the average performance drop is  $< 3$  LAS points. LAS distributions for SA are also notably more narrow than those for core relations and LA, suggesting that the drop on SA relations can be largely attributed to the *removal of the original noun*; i.e., the lexical properties of the original noun that the models pick up on fit a very narrow class of nouns, so it is unlikely that a noun with similar properties will fall into that slot. The difference between trends observed here for SA relations and LA relations highlights the importance of separating the two while examining lexicosemantic effects in UD.<sup>7</sup>

**Core vs other relations** Across all experimental conditions, the performance decrease on core relations sits between that for SA relations and that for LA relations. LAS distributions for core relations are also always the widest. On one hand, the fact that the drops observed for core relations are always smaller than those for LA relations is a good sign – since LA encapsulates relations which often require the models to rely on lexical signals, performance on this category can be thought of as a baseline in this

<sup>7</sup>Notably, the performance deteriorations observed here for SA relations are much lower than those reported in related work for English (Hall Maudslay and Cotterell, 2021, Kasai and Frank, 2019). While this could be due to language differences, separation of LA and SA relations could be another contributing factor.

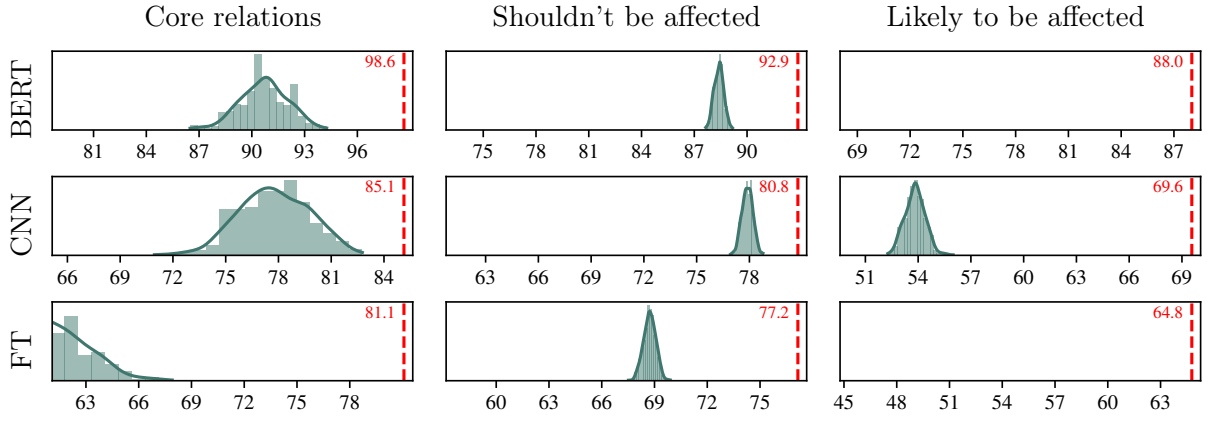


**Figure 5.6:** Finnish: LAS distribution on data with permuted noun lexemes (500 runs).

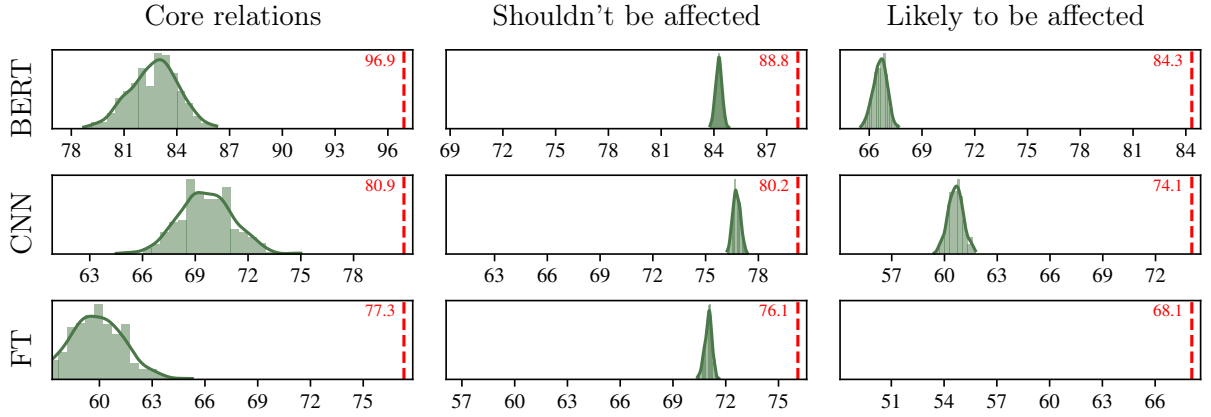


**Figure 5.7:** Estonian: LAS distribution on data with permuted noun lexemes (500 runs).

context. On the other hand, the difference between the core results and SA results signals that core relations are an area of particular difficulty – the models are affected by lexical reassignments to a greater extent while parsing core relations, compared to parsing other relations that should not be affected by lexical changes. The wider distributions point to the fact that performance largely depends on what kind of noun replaced the original one. This suggests that, for core relations, the models rely on more general lexical properties, which fit a broad set of nouns. I hypothesise that one of those properties is animacy. While testing this hypothesis is beyond the scope of this thesis, it could be tested in future work by directly controlling for animacy of the mixed-in nouns; note that such alterations are naturally supported by the LEXMIX framework.

**Comparison across the models** Both BERT-based and CNN-based models seem to rely on lexicosemantic signals to a very similar extent. This is demonstrated in the LAS drop on both the core relations and the SA relations.<sup>8</sup> For Polish and Russian, the trends for both models are nearly identical (on core relations the average LAS decrease is 6.3 and 7.1 for Polish BERT and CNN, respectively; for Russian I note 9.7 for BERT and

<sup>8</sup>The patterns for LA relations cannot lead to meaningful conclusions here.



with the pre-log values noted on a reference axis (top of the figure), for ease of reading. In the plots, each relation is colored based on its category from Table 5.2 – relations categorised as SA are cool-colored (hues of blue), those categorised as LA are warm-colored (hues of red). Within each category, the darker the color the higher the drop *on the Polish treebank*. Following the same coloring across all plots facilitates cross-language comparisons. The core relations are colored grey.

A quick glance at Figure 5.8 allows to notice the separation between the SA and the LA relations. The cut between the two is the cleanest for Finnish, with only one SA relation – *nummod* – passing the boundary of  $\sim 8$  LAS decrease.<sup>9</sup> I notice that there are many similarities between the languages when it comes to the LAS drops noted for individual relations. Some relations are universally challenging. These include the *flat*, *appos*, *comopund(:nn)*, *nmod*, categorised as LA. Out of SA relations the hardest ones, with average LAS drop  $> 5$ , include: *nummod(:gov)*, *acl(:relcl)*, *conj*, *parataxis(:insert)*, and for Finnish and Estonian also *cop*, *nsubj:cop* and *case*. While examining the causes of the larger drop on those SA relations is outside the scope of this thesis, for *acl* the drop could be caused by its use with special subset of nouns like *fact* or *report* (see Appendix F). Other relations are universally very little affected, such as *advmod*, *aux*, *xcomp*, *mark* and *advcl*. There are also some language peculiarities, such as the prominent drop on Russian *nsubj:pass* or Polish *xcomp:pred*.

For each language, the core relations sit roughly in the middle of all relations ranked according to the average LAS difference. In all cases, the performance drop on *nsubj* is lesser than for *obj*. This gap between subjects and objects is most prominent for Russian.

## 5.3 Experiment II: Core argument rotation

In the previous section I experimented with random lexeme reassignments via a perturbation that affected the majority of nouns in a corpus. Here, I experiment with a more targeted perturbation, which swaps the lexemes of the subject and object in transitive sentences. This perturbation, which I term **core argument rotation** is, at least in theory, more challenging than random lexeme replacements because it ensures that the lexicosemantic cues pull in the opposite direction from the correct interpretation signalled by morphology.

### 5.3.1 Experimental details

**Models and data** As in Section 5.2, I experiment on the models from the previous chapter (see Section 4.3) trained on the largest UD treebank available for the language, but

<sup>9</sup>One possible reason for this drop on *nummod* is that in Finnish the lexeme can influence the number of a whole phrase – “Finnish numerals occur with singular nouns [...] However, public events such as funerals, weddings, sport events, etc. are often denoted by plural nouns.” (Karttunen, 2006).

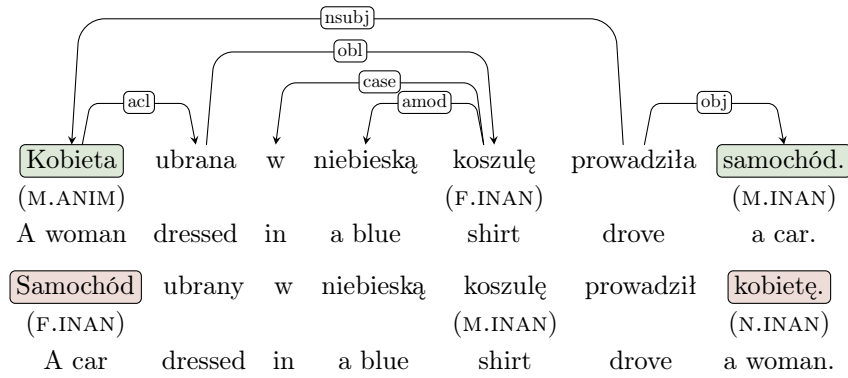


**Figure 5.9:** An example of core argument rotation perturbation.

this time I report the results averaged across 4 versions of each model, trained with different random seeds. I also evaluate on the largest UD treebanks but, since the perturbation concerns only transitive clauses, I only consider sentences which include at least one transitive clause. To filter the data, I follow the same steps as those followed while creating counterfactual treebanks from Chapter 4 (see Section 4.2.1 for details). Importantly, I only repurpose the *filtering* step and do not simplify or reorder the sentences.

I also follow my experiments from Chapter 4 in controlling for morphosyntactic ambiguity in evaluation data. I create two versions of the filtered data: one that contains only morphosyntactically unambiguous transitive clauses and one that contains only ambiguous clauses. Note that the lack/presence of ambiguity is naturally preserved even after the rotation of argument lexemes. This is unlike the permutation perturbation, where each permutation results in a different set of ambiguous sentences (which is why I did not control for ambiguity in those experiments).<sup>10</sup> Mirroring Chapter 4, the expectation here is that if a model’s generalisations resemble those of native speakers, who rank morphology higher than lexical semantics, the parsers *should not* rely on lexical semantics when parsing unambiguous clauses (equivalent of **CONDITION I** in Chapter 4), but they *could* rely on lexical cues when parsing ambiguous clauses (equivalent of **CONDITION II** in Chapter 4); see 4.5 for a more detailed discussion.

**The perturbation** Figure 5.9 provides an example of the perturbation. Note that (i) like in the case of lexeme permutations (Section 5.2), swapping the lexemes of core arguments often requires inflection and (ii) unlike the lexeme permutations, the rotation perturbation alters *only* the lexemes of the core noun arguments. Also note that I do not rotate the whole NPs, but only the lexemes of nouns which head the NPs. This arguably results in an easier evaluation condition than if the whole NPs were rotated. This is because it is not only the nouns, but also their modifiers that can be correlated with

<sup>10</sup>While changing the set of evaluation sentences for each permuted dataset would be possible, the results across runs would not be directly comparable.

subject/objecthood. Consider, for example, the modifier ‘dressed in a blue shirt’ from Figure 5.9. This modifier alone suggests that the modified noun is animate and neural models could rely on this cue even when the subject noun itself is not (typically) animate (e.g., ‘a car’). As a result, the experimentation in this section provides a lower-bound on the model’s tendencies to rely on lexicosemantic cues when identifying core arguments.

### 5.3.2 Results

In Table 5.3 and Table 5.4 I present the core relation LAS results for the unambiguous clauses and the ambiguous clauses, respectively (see Table G.8 and Table G.9, in Appendix G for standard deviation scores). I report the performance on unaltered sentences (O), sentences with rotated core arguments (R) and the difference between the two (O - R). Note that although I evaluate on the same subsets of sentences as in Section 4.4 (transitive unambiguous) and in Section 4.5 (transitive ambiguous), the transitive clause counts differ because here I do not simplify the sentences.

#### 5.3.2.1 Unambiguous sentences

Although the rotation perturbation should be, in principle, more challenging for the models than random lexeme reassignments, this appears to be the case only for the Slavic language pair. For both Polish and Russian I note a larger decline in Table 5.3 than the average decline observed in Figure 5.4 and Figure 5.5 in all conditions, except the Russian CNN. In contrast, for Finnish and Estonian, rotation of the core lexemes causes a *smaller* LAS drop than noun lexeme permutations in all conditions, except the Finnish CNN. In general, all Finnic models suffer smaller drops than their Slavic equivalents. This stands in contrast to what could be observed for core relations in Section 5.2, where Estonian models were among those most affected. One potential reason for these differences could be the absence of ditransitive sentences in the filtered data – it could be that for the Finnic language pair, lexeme alterations are particularly challenging in ditransitive clauses – or the occasional presence of ambiguity in the permuted data from Section 5.2. Alternatively, the unaltered modifying clauses of the rotated nouns could be providing some signal (see the discussion of the perturbation in Section 5.3.1).

As in the permutation experiments from Section 5.2, there is little difference between the performance decrease observed for BERT-based and the CNN-based parsers. Similarly, FASTTEXT remains the worst performing model. Interestingly, the CNN suffers from a lesser drop than BERT for the Slavic languages and a larger drop for the Finnic languages – this is a reverse of what I observed for permutation experiments.

|      |     | Polish<br>302 clauses |       |       |       | Russian<br>654 clauses |       |       |       | Finnish<br>129 clauses |       |       | Estonian<br>268 clauses |       |       |
|------|-----|-----------------------|-------|-------|-------|------------------------|-------|-------|-------|------------------------|-------|-------|-------------------------|-------|-------|
|      |     | core                  | nsubj | obj   | iobj  | core                   | nsubj | obj   | iobj  | core                   | nsubj | obj   | core                    | nsubj | obj   |
| BERT | O   | 96.0                  | 98.9  | 96.1  | 81.5  | 99.2                   | 99.3  | 99.2  | 96.3  | 99.2                   | 99.8  | 98.6  | 96.6                    | 98.0  | 95.1  |
|      | R   | 86.3                  | 88.8  | 84.7  | 80.6  | 87.0                   | 89.5  | 84.0  | 93.2  | 94.5                   | 95.3  | 93.6  | 89.8                    | 93.2  | 86.2  |
|      | O-R | ↓9.7                  | ↓10.2 | ↓11.4 | ↓0.8  | ↓12.2                  | ↓9.9  | ↓15.3 | ↓3.1  | ↓4.7                   | ↓4.5  | ↓5.0  | ↓6.8                    | ↓4.8  | ↓8.9  |
| CNN  | O   | 82.8                  | 87.3  | 82.0  | 64.1  | 88.9                   | 89.9  | 89.3  | 71.3  | 86.3                   | 91.8  | 80.8  | 83.7                    | 87.6  | 79.8  |
|      | R   | 75.7                  | 82.3  | 71.9  | 58.1  | 78.4                   | 81.6  | 76.1  | 64.6  | 77.8                   | 84.1  | 71.5  | 73.3                    | 81.3  | 65.3  |
|      | O-R | ↓7.2                  | ↓5.0  | ↓10.1 | ↓6.0  | ↓10.5                  | ↓8.3  | ↓13.2 | ↓6.8  | ↓8.5                   | ↓7.8  | ↓9.3  | ↓10.3                   | ↓6.2  | ↓14.5 |
| FT.  | O   | 84.6                  | 87.3  | 84.2  | 72.6  | 91.5                   | 91.2  | 92.5  | 84.3  | 83.2                   | 85.7  | 80.8  | 81.1                    | 81.5  | 80.5  |
|      | R   | 63.4                  | 66.3  | 64.3  | 46.0  | 68.5                   | 70.9  | 66.8  | 56.2  | 66.7                   | 78.1  | 55.2  | 65.8                    | 74.2  | 57.3  |
|      | O-R | ↓21.1                 | ↓21.0 | ↓19.9 | ↓26.6 | ↓23.0                  | ↓20.2 | ↓25.6 | ↓28.1 | ↓16.6                  | ↓7.6  | ↓25.5 | ↓15.3                   | ↓7.3  | ↓23.3 |

**Table 5.3:** Unambiguous data: LAS scores on unaltered transitive sentences (O) and their equivalents with rotated core arguments (R). Column *core* reports LAS for all core relations.

|      |     | Polish<br>49 clauses |       |       |       | Russian<br>181 clauses |       |       |       | Finnish<br>15 clauses |       |       | Estonian<br>49 clauses |       |       |
|------|-----|----------------------|-------|-------|-------|------------------------|-------|-------|-------|-----------------------|-------|-------|------------------------|-------|-------|
|      |     | core                 | nsubj | obj   | iobj  | core                   | nsubj | obj   | iobj  | core                  | nsubj | obj   | core                   | nsubj | obj   |
| BERT | O   | 95.7                 | 97.0  | 96.0  | 80.0  | 97.4                   | 97.7  | 97.1  | 100   | 95.0                  | 91.7  | 98.3  | 91.9                   | 87.8  | 95.9  |
|      | R   | 55.4                 | 60.2  | 53.4  | 25.0  | 70.3                   | 73.1  | 67.4  | 75.0  | 83.3                  | 80.0  | 86.7  | 67.6                   | 64.3  | 70.9  |
|      | O-R | ↓40.3                | ↓36.8 | ↓42.6 | ↓55.0 | ↓27.0                  | ↓24.6 | ↓29.7 | ↓25.0 | ↓11.7                 | ↓11.7 | ↓11.7 | ↓24.3                  | ↓23.5 | ↓25.0 |
| CNN  | O   | 88.8                 | 89.8  | 92.0  | 50.0  | 90.9                   | 89.2  | 92.4  | 100   | 81.7                  | 80.0  | 83.3  | 68.9                   | 66.8  | 70.9  |
|      | R   | 67.1                 | 63.8  | 77.3  | 10.0  | 79.4                   | 79.6  | 79.5  | 66.7  | 68.3                  | 66.7  | 70.0  | 56.6                   | 61.2  | 52.0  |
|      | O-R | ↓21.7                | ↓26.0 | ↓14.8 | ↓40.0 | ↓11.5                  | ↓9.6  | ↓12.9 | ↓33.3 | ↓13.3                 | ↓13.4 | ↓13.3 | ↓12.2                  | ↓5.6  | ↓18.9 |
| FT.  | O   | 78.6                 | 77.0  | 84.1  | 45.0  | 86.1                   | 84.2  | 88.9  | 33.3  | 68.3                  | 66.7  | 70.0  | 58.9                   | 56.6  | 61.2  |
|      | R   | 56.1                 | 53.6  | 64.2  | 10.0  | 68.5                   | 66.5  | 70.3  | 75.0  | 65.0                  | 68.3  | 61.7  | 43.9                   | 51.5  | 36.2  |
|      | O-R | ↓22.4                | ↓23.4 | ↓19.9 | ↓35.0 | ↓17.6                  | ↓17.7 | ↓18.6 | ↑41.7 | ↓3.3                  | ↑1.7  | ↓8.3  | ↓15.0                  | ↓5.1  | ↓25.0 |

**Table 5.4:** Ambiguous data: LAS scores on unaltered transitive sentences (O) and their equivalents with rotated core arguments (R). Column *core* reports LAS for all core relations.

### 5.3.2.2 Ambiguous sentences

As in Section 4.5 (discussing word order experiments on ambiguous clauses), because of the very low percentage of ambiguous transitive clauses in the original treebanks, for each language, I evaluate the models on the concatenation of the development and test splits of the filtered treebanks. As before, even after merging the two splits, the quantity of the resulting data for Finnish is insufficient to draw any reliable conclusions. I do include those results in the tables, but caution that these results are very preliminary. This also applies to the Slavic *iobj* relations, which are very sparse.

Just like for the word order experiments, for evaluation on ambiguous clauses here, there is no such thing as a good solution. It is not clear to what extent a model should rely on lexical semantics in the presence of ambiguity, in a situation where word order cues might be pulling its predictions in the opposite direction. However, relevant insights can be drawn by comparing the results on ambiguous clauses (Table 5.4) to those for unambiguous clauses (Table 5.3).

**Comparing LAS decrease (O - R)** This comparison reveals that CNN and BERT rely on lexical cues to a far greater extent in ambiguous clauses. This applies particularly

to the BERT-based models – the difference between the drop on unambiguous sentences and the drop on ambiguous sentences can be as large as 30 LAS (noted for the Polish core relations). Indeed, in evaluation on ambiguous sentences, BERT stands out as the most affected model. This is a very interesting finding, especially in the light of results on unambiguous data, where BERT was affected by the rotations to a similar extent as the lexically-subpar, non-pre-trained CNN. This adds to the findings from Section 4.5, signaling that the BERT-based parsers are, at least to some extent, sensitive to morphosyntactic ambiguity and that it affects their parsing strategies to a larger extent than the alternative models.

For FASTTEXT, on the other hand, in most conditions I note a *smaller* performance drop for ambiguous sentences and in some cases even a performance increase (although this increase is for evaluation conditions with only a few data instances – Finnish and Russian *iobj*). This suggests that morphosyntactic ambiguity does not have an effect on the extent to which these models rely on lexical semantics, which stands in contrast to the other two models.

These findings, resemble those for word order in Section 4.5, with two exceptions: (i) in Section 4.5 all models were affected, including FASTTEXT and (ii) the effect observed here for CNN and BERT is more prominent, which might mean that the disambiguation strategies of those models are more heavily based on lexical semantics than word order. Similarly, my conclusions here resemble those from Section 4.5; it is a good sign that the performance of CNN and BERT-based parsers suffers more from argument lexeme rotation in the absence of morphosyntactic signals. This suggests that they can recognise the instances of morphosyntactic ambiguity and that the ambiguity affects their parsing strategy. It also points to the fact that the presence of morphology pulls the models away from relying on lexical semantics.

## 5.4 Experiment III: Morphology vs word order + lexical semantics

In this section I bring together the word order alterations from Chapter 4 and the core argument lexeme rotations from Section 5.3. I do this by introducing another version of the counterfactual dataset from Section 4.2 – one in which all transitive clauses in differently ordered treebanks have their arguments rotated (see Table 5.5 for an example). In this experiment, morphology is placed in direct opposition to both word order and lexicosemantic cues, consequently becoming the only linguistic signal pointing to the correct interpretation. More precisely, this applies to the non-SVO orders and to the OVS order in particular, since it breaks all of the word order tendencies – i.e., (i) the subject does not precede the verb, (ii) the object does not follow the verb, (iii) the subject is not

|         | COUNTERFACTUAL<br>TREEBANKS (Chapter 4)                                                                     | ROTATED COUNTERFACTUAL<br>TREEBANKS (Section 5.4)                                                            |
|---------|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Meaning | A boy is holding a red bucket.                                                                              | A bucket is holding a red boy.                                                                               |
| svo     | Chłopak trzyma czerwone wiadro.                                                                             | Wiadro trzyma czerwonego chłopaka.                                                                           |
| ovs     | Czerwone wiadro trzyma chłopak.                                                                             | Czerwonego chłopaka trzyma wiadro.                                                                           |
| sov     | Chłopak czerwone wiadro trzyma.                                                                             | Wiadro czerwonego chłopaka trzyma.                                                                           |
| osv     | Czerwone wiadro chłopak trzyma.                                                                             | Czerwonego chłopaka wiadro trzyma.                                                                           |
| vso     | Trzyma chłopak czerwone wiadro.                                                                             | Trzyma wiadro czerwonego chłopaka.                                                                           |
| vos     | Trzyma czerwone wiadro chłopak.                                                                             | Trzyma czerwonego chłopaka wiadro.                                                                           |
|         | ORIGINAL SENTENCE                                                                                           | TRANSLATION                                                                                                  |
|         | Chłopak w czapce i kamizelce trzyma czerwone wiadro, z którego ubrany w granatowe okrycie koń wyjada siano. | A boy in a hat and a vest is holding a red bucket, from which a horse dressed in a navy cover is eating hay. |

**Table 5.5:** Difference between the counterfactual treebanks from Chapter 4 and the rotated counterfactual treebanks from Section 5.4, demonstrated using a Polish sentence as an example.

the first core element in a clause, (iv) the object is not the last core element in a clause.

The result is a challenging evaluation setting in which only the models capable of (i) retrieving the relevant subword information *and* (ii) recognising it as the primary meaning indicator can excel. It is also a more pure test of the models morphological competence than those explored in Section 4.4, Section 4.5, Section 5.2 and Section 5.3, in which morphological signal was always accompanied by lexical or word order cues, strengthening the overall signal.

This section is split into three subsections. In the first (Section 5.4.1), I empirically demonstrate that a model without access to morphology cannot solve the task proposed here, by presenting and discussing LAS results on the **rotated counterfactual treebanks** for the two lemmatising, morphologically blind baselines from Section 3.3. These baseline results provide context for experimentation on the 12 parsers from Section 4.3 (based on DM), which I discuss in the following Section 5.4.2. Finally in Section 5.4.3, I discuss results from Section 5.4.2 in the light of my earlier experiments which involved only one type of perturbation (word order or lexical); this discussion provides insights into the relative ranking of signals within the individual models.

### 5.4.1 Lemmatising baselines

In the top section of Table 5.6 I present the results on the rotated version of the *unambiguous* split<sup>11</sup> of the counterfactual treebanks for the baselines processing lemmatised language. In addition to reporting the results for each differently ordered, rotated treebank (SVO, OVS,

<sup>11</sup>While I do not include those results here, I have also run experiments on the *ambiguous* version of the rotated counterfactual treebanks. As expected, given the lack of morphological information, the presence/absence of morphosyntactic ambiguity had no regular effect on the baselines performance – I observed similar LAS performance for the ambiguous treebanks to that for the unambiguous treebanks.

etc.) I also report results for their unaltered equivalent, which holds the same *simplified* sentences as each of the reordered versions but with lexical semantics and word order unaltered. These results are presented in the row labelled *Sim.*

While in Table 5.6 I report results for all core relations in each reordered treebank, it is important to highlight that in this experiment the following evaluation conditions are of particular importance:

- for *nsubj*: OVS, VSO, VOS
- for *(i)obj*: OVS, SOV, OSV

These are the conditions in which both the absolute position and the position relative to the verb stand in opposition to the statistical tendencies (see Section 4.1.2). To differentiate those conditions from others in the discussion, I refer to them as *relevant* or *challenging*.

**The challenging conditions** As expected, both BERT-based and CNN-based baselines perform *very* poorly on all of the conditions listed above. For *nsubj* the performance is very close to 0 LAS (!) for some word orders – these include Finnish OVS and VOS (1.4 LAS and 0.4 LAS for BERT, 0.6 and 0.4 for CNN), and Polish VOS (4.8 LAS for BERT, 1.9 for CNN). Performance on the object relations is marginally better throughout, but remains low – I record the highest result of 30.5 LAS for Estonian SOV. As predicted, OVS proves to be the most challenging ordering for the baselines, with very low performance for all core relations.

These results validate the experimental approach I propose here. The difference between the *Sim.* results and the results on the challenging evaluation conditions is huge, revealing the baselines’ inability to rely on morphology, despite achieving good performance on standard evaluation. In the few cases that the models get right, they are likely relying on spurious, non-linguistic signals. Another possibilities are that (a) the models rely on lexical signals from the arguments’ modifiers and that (b) in those cases both candidate lexemes are plausible subjects/objects.

**Other conditions** For conditions in which some ordering signal is preserved (i.e., those not listed above), performance is generally better, and in some instances notably better. For subjects, those instances include performance on the Slavic subject-initial word orders (SVO and SOV), on the Finnish word orders in which subject directly precedes the verb (SVO, OSV) and Estonian SVO, SOV and OSV. For objects, in all three: Polish, Russian and Finnish, both baselines pick up on the relative position of the object with respect to the verb (best performance on SVO and VOS). For Estonian, it is harder to spot a pattern – the best performance is noted for SVO, followed by (the more challenging) SOV. Note that those patterns largely mirror those observed in Section 4.4.2 for morphologically-aware

| Polish<br>clause count: 295          |      |       |      | Russian<br>clause count: 632 |      |       |      | Finnish<br>clause count: 127 |      |       | Estonian<br>clause count: 258 |      |       |      |
|--------------------------------------|------|-------|------|------------------------------|------|-------|------|------------------------------|------|-------|-------------------------------|------|-------|------|
|                                      | all  | nsubj | obj  | iobj                         | all  | nsubj | obj  | iobj                         | all  | nsubj | obj                           | all  | nsubj | obj  |
| MORPHOLOGICALLY BLIND BASELINES      |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| DM +BERT                             |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| Sim.                                 | 89.6 | 89.4  | 91.0 | 70.9                         | 95.0 | 94.2  | 92.3 | 83.0                         | 91.7 | 86.7  | 91.5                          | 82.2 | 74.0  | 75.7 |
| svo                                  | 69.8 | 41.1  | 41.8 | 39.8                         | 70.5 | 35.8  | 25.4 | 42.6                         | 61.3 | 26.4  | 35.0                          | 62.9 | 30.4  | 36.4 |
| ovs                                  | 58.6 | 5.1   | 5.8  | 4.9                          | 63.2 | 7.2   | 4.8  | 10.1                         | 50.8 | 1.4   | 3.3                           | 55.6 | 8.0   | 11.3 |
| sov                                  | 58.5 | 27.7  | 10.0 | 5.7                          | 63.0 | 20.8  | 5.5  | 23.9                         | 52.2 | 5.1   | 14.6                          | 56.2 | 17.6  | 25.2 |
| osv                                  | 56.9 | 7.5   | 21.6 | 11.1                         | 62.1 | 10.8  | 12.0 | 15.4                         | 52.0 | 14.0  | 14.0                          | 52.1 | 11.6  | 12.6 |
| vso                                  | 56.6 | 5.6   | 8.5  | 16.0                         | 59.2 | 5.5   | 3.5  | 5.3                          | 51.8 | 2.8   | 17.9                          | 56.7 | 8.2   | 22.4 |
| vos                                  | 63.8 | 4.8   | 56.4 | 52.9                         | 63.8 | 14.0  | 26.8 | 45.8                         | 52.7 | 0.4   | 21.1                          | 55.4 | 4.3   | 22.3 |
| DM +CNN                              |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| Sim.                                 | 81.4 | 79.5  | 78.4 | 55.4                         | 86.1 | 80.6  | 76.4 | 46.3                         | 75.1 | 72.5  | 62.4                          | 71.8 | 56.2  | 61.5 |
| svo                                  | 74.3 | 66.3  | 67.5 | 47.1                         | 80.6 | 70.6  | 67.8 | 44.1                         | 64.6 | 48.0  | 52.4                          | 63.7 | 42.8  | 46.1 |
| ovs                                  | 55.9 | 7.5   | 9.6  | 5.7                          | 60.6 | 4.6   | 4.6  | 7.5                          | 50.0 | 0.6   | 12.8                          | 50.2 | 3.6   | 9.5  |
| sov                                  | 61.0 | 59.9  | 11.1 | 7.8                          | 66.5 | 52.0  | 10.2 | 11.2                         | 51.9 | 14.6  | 26.6                          | 55.5 | 26.6  | 30.5 |
| osv                                  | 54.4 | 9.6   | 18.1 | 5.3                          | 61.0 | 13.3  | 12.3 | 4.2                          | 50.0 | 28.9  | 7.1                           | 49.1 | 18.6  | 9.3  |
| vso                                  | 55.1 | 12.4  | 14.7 | 4.9                          | 59.7 | 7.6   | 5.6  | 7.5                          | 49.6 | 2.2   | 17.9                          | 53.6 | 6.1   | 25.9 |
| vos                                  | 63.0 | 1.9   | 74.8 | 45.1                         | 68.4 | 3.8   | 68.7 | 31.4                         | 52.9 | 0.4   | 35.8                          | 52.3 | 2.1   | 27.1 |
| MODELS PROCESSING UNALTERED LANGUAGE |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| DM +BERT                             |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| Sim.                                 | 93.8 | 98.8  | 97.8 | 86.1                         | 97.5 | 99.2  | 98.6 | 93.1                         | 97.7 | 99.8  | 98.8                          | 94.7 | 98.3  | 97.0 |
| svo                                  | 87.6 | 89.7  | 87.4 | 83.1                         | 90.7 | 87.3  | 82.1 | 94.3                         | 92.3 | 94.1  | 95.1                          | 89.3 | 93.7  | 89.9 |
| ovs                                  | 82.0 | 68.4  | 70.4 | 72.1                         | 85.8 | 73.2  | 65.6 | 87.2                         | 88.8 | 84.2  | 83.7                          | 87.6 | 87.3  | 81.1 |
| sov                                  | 75.3 | 71.8  | 38.0 | 41.5                         | 78.8 | 67.3  | 43.1 | 80.2                         | 83.0 | 78.9  | 66.1                          | 81.8 | 87.1  | 65.9 |
| osv                                  | 80.9 | 70.0  | 70.4 | 66.7                         | 83.7 | 70.0  | 66.1 | 82.3                         | 81.8 | 86.2  | 56.3                          | 78.1 | 86.0  | 50.8 |
| vso                                  | 67.8 | 35.4  | 23.8 | 47.0                         | 73.0 | 42.5  | 32.3 | 78.0                         | 83.8 | 68.2  | 78.8                          | 84.5 | 83.7  | 76.5 |
| vos                                  | 78.7 | 47.2  | 85.0 | 78.1                         | 82.2 | 56.2  | 72.5 | 90.1                         | 75.2 | 41.5  | 58.1                          | 78.7 | 71.9  | 59.9 |
| DM +CNN                              |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| Sim.                                 | 83.8 | 86.0  | 81.7 | 66.4                         | 91.2 | 87.5  | 88.6 | 65.4                         | 85.6 | 91.2  | 85.2                          | 83.9 | 84.2  | 78.2 |
| svo                                  | 81.8 | 86.4  | 81.9 | 59.0                         | 86.5 | 79.9  | 77.4 | 65.5                         | 80.5 | 86.6  | 78.1                          | 79.8 | 82.0  | 68.3 |
| ovs                                  | 64.9 | 29.2  | 30.0 | 28.3                         | 73.3 | 34.1  | 37.6 | 55.9                         | 69.7 | 55.7  | 46.3                          | 72.0 | 58.0  | 42.8 |
| sov                                  | 68.3 | 80.1  | 25.1 | 19.3                         | 75.0 | 63.6  | 32.7 | 46.3                         | 60.5 | 60.4  | 21.2                          | 65.8 | 58.0  | 35.9 |
| osv                                  | 61.5 | 32.6  | 30.6 | 23.4                         | 73.4 | 36.0  | 48.2 | 55.8                         | 65.2 | 79.3  | 17.9                          | 62.4 | 62.0  | 15.1 |
| vso                                  | 61.0 | 27.2  | 30.5 | 27.9                         | 68.2 | 28.7  | 22.2 | 43.6                         | 61.2 | 42.3  | 35.8                          | 68.8 | 48.7  | 49.0 |
| vos                                  | 65.1 | 11.2  | 70.5 | 51.2                         | 74.6 | 26.9  | 66.6 | 60.6                         | 62.5 | 34.2  | 38.0                          | 65.7 | 48.6  | 32.5 |
| DM +FASTTEXT                         |      |       |      |                              |      |       |      |                              |      |       |                               |      |       |      |
| Sim.                                 | 85.3 | 85.7  | 85.0 | 71.7                         | 91.8 | 89.6  | 91.9 | 79.3                         | 81.6 | 85.7  | 82.5                          | 77.1 | 77.7  | 78.9 |
| svo                                  | 72.0 | 66.1  | 73.6 | 49.6                         | 80.3 | 73.6  | 69.5 | 62.4                         | 73.3 | 80.3  | 64.6                          | 74.2 | 77.8  | 65.9 |
| ovs                                  | 58.0 | 30.6  | 32.5 | 16.8                         | 66.7 | 35.0  | 27.9 | 28.4                         | 58.0 | 29.1  | 29.3                          | 64.7 | 53.2  | 35.2 |
| sov                                  | 60.4 | 55.7  | 41.3 | 18.0                         | 69.5 | 54.0  | 37.8 | 41.1                         | 59.2 | 44.7  | 31.7                          | 64.0 | 52.7  | 49.4 |
| osv                                  | 54.7 | 37.8  | 25.9 | 13.5                         | 65.5 | 40.9  | 32.1 | 16.3                         | 59.0 | 65.5  | 14.8                          | 58.3 | 60.0  | 17.4 |
| vso                                  | 56.1 | 32.7  | 33.8 | 20.5                         | 57.7 | 26.5  | 18.1 | 36.2                         | 40.4 | 19.5  | 21.2                          | 44.5 | 30.3  | 32.0 |
| vos                                  | 59.4 | 21.6  | 56.7 | 33.6                         | 61.6 | 23.6  | 38.0 | 41.1                         | 42.9 | 14.3  | 20.3                          | 45.0 | 24.4  | 20.2 |

**Table 5.6:** LAS scores on the development splits of the *unambiguous* counterfactual treebanks with core arguments *rotated* (see the right-hand side of Table 5.5). *Sim.* stands for the simplified but otherwise unaltered version of the sentences.

models. This points to the fact that the morphologically aware models rely on the same word ordering cues as the lemmatising models. Further, such cues can outweigh the lexicosemantic alternatives – if that were not the case the results would remain low across all word orders. Notably, CNN-based baseline seems to tap into word order to a greater extent, compared to its BERT-based counterpart. This is manifested in its (i) much better

overall performance on the SVO treebanks, (ii) better performance on *nsubj* for treebanks in which the relative position of the subject matches SVO (namely: SOV and OSV) and (iii) better performance on *obj* for treebanks in which the relative position of the object matches SVO (VSO, VOS). The latter two tendencies are particularly prominent for the Slavic languages – compare, for example, 27.7 LAS on *nsubj* for SOV reached by BERT and the corresponding 59.9 LAS achieved by the CNN. This difference between BERT and CNN inputs likely arises due to the richer lexical knowledge of the first, which causes it to tap into lexical semantics to a greater extent.

## 5.4.2 Morphologically aware parsers

### 5.4.2.1 LAS results

Following the results for the lemmatising baselines, I now discuss the results for the models processing unaltered language (i.e., with full access to morphology). These results are presented under the baseline results in Table 5.6 (see Table G.10, in Appendix G for standard deviation scores). I focus my discussion on the most relevant conditions, in which only morphology points to the correct interpretation (see previous section).

Across all relations, both CNN and BERT-based morphologically-aware parsers perform better than their baseline counterparts. This alone is not a particularly high achievement, since the latter have *no access* to morphology. For the CNN-based parser the improvement over the baseline is in many instances quite modest – e.g., +14.8 LAS and +21.1 for Polish and Russian *nsubj* VSO. Also notice how for Finnish *obj* relations in SOV clauses the CNN lemmatising baseline performs *better* than the model processing inflected forms. Providing BERT with morphological access yields greater improvements, all of which can be attributed to the model relying on case and agreement markings. However, in multiple conditions of interest the performance remains < 50 LAS.

The BERT-based parser stands out as the overall best performing model. The CNN-based parser takes the second place in the performance ranking, with the difference between FASTTEXT and CNN most prominent on *nsubj* relations. BERT also stands out as having a relatively good performance on OVS, which also happens to be the order for which I note the largest improvement for BERT over its lemmatising counterpart. This is the more relevant, given that OVS was the most challenging word order for the baseline. Perhaps the improvement over the baseline is highest for OVS because the separation of the two arguments nudges BERT towards paying attention to morphology, but it could be also attributed to a lower likelihood of misattachment-as-nominal-modifier errors for OVS, compared to the other non-SVO orders.

|       |       | SVO   |     |      |       |    | OVS   |     |      |       |    | SOV   |     |      |       |     | OSV   |     |      |       |     | VSO   |     |      |       |     | VOS   |     |      |       |     |
|-------|-------|-------|-----|------|-------|----|-------|-----|------|-------|----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|
|       |       | nsubj | obj | iobj | other | WH | nsubj | obj | iobj | other | WH | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  |
| BERT  | nsubj | 271   | 21  | 2    | 0     | 2  | 209   | 73  | 12   | 1     | 1  | 238   | 44  | 6    | 0     | 8   | 230   | 31  | 5    | 1     | 29  | 124   | 148 | 19   | 1     | 4   | 147   | 63  | 28   | 5     | 53  |
|       | obj   | 18    | 208 | 7    | 0     | 3  | 67    | 161 | 3    | 3     | 2  | 26    | 67  | 2    | 0     | 141 | 53    | 160 | 5    | 2     | 16  | 21    | 45  | 14   | 5     | 151 | 16    | 196 | 13   | 2     | 9   |
|       | iobj  | 2     | 3   | 50   | 5     | 1  | 9     | 4   | 43   | 4     | 1  | 3     | 1   | 23   | 3     | 31  | 14    | 4   | 39   | 2     | 2   | 0     | 2   | 27   | 3     | 29  | 3     | 4   | 49   | 2     | 3   |
| CNN   | nsubj | 261   | 21  | 5    | 1     | 8  | 59    | 198 | 21   | 6     | 12 | 239   | 27  | 5    | 2     | 23  | 85    | 34  | 5    | 4     | 168 | 73    | 172 | 16   | 13    | 22  | 29    | 73  | 7    | 8     | 179 |
|       | obj   | 7     | 207 | 3    | 3     | 16 | 140   | 77  | 6    | 1     | 12 | 36    | 50  | 5    | 1     | 144 | 123   | 86  | 1    | 1     | 25  | 18    | 56  | 3    | 5     | 154 | 27    | 175 | 5    | 9     | 20  |
|       | iobj  | 5     | 15  | 35   | 2     | 4  | 37    | 1   | 17   | 0     | 6  | 10    | 3   | 14   | 0     | 34  | 32    | 2   | 14   | 2     | 11  | 2     | 3   | 16   | 2     | 38  | 12    | 8   | 31   | 2     | 8   |
| FASTT | nsubj | 180   | 48  | 5    | 2     | 61 | 77    | 150 | 14   | 24    | 31 | 170   | 43  | 3    | 0     | 80  | 79    | 68  | 9    | 2     | 138 | 64    | 147 | 8    | 44    | 33  | 47    | 65  | 0    | 1     | 183 |
|       | obj   | 9     | 176 | 5    | 26    | 20 | 97    | 66  | 3    | 0     | 70 | 19    | 61  | 0    | 2     | 154 | 77    | 43  | 2    | 5     | 109 | 6     | 44  | 0    | 4     | 182 | 12    | 142 | 3    | 58    | 21  |
|       | iobj  | 10    | 8   | 33   | 6     | 4  | 23    | 7   | 11   | 5     | 15 | 6     | 1   | 12   | 1     | 41  | 27    | 4   | 10   | 3     | 17  | 0     | 2   | 10   | 1     | 48  | 5     | 14  | 25   | 12    | 5   |

|       |       | SVO   |     |      |       |     | OVS   |     |      |       |     | SOV   |     |      |       |     | OSV   |     |      |       |     | VSO   |     |      |       |     | VOS   |     |      |       |     |
|-------|-------|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|-------|-----|------|-------|-----|
|       |       | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  | nsubj | obj | iobj | other | WH  |
| BERT  | nsubj | 563   | 48  | 1    | 16    | 5   | 494   | 101 | 4    | 30    | 4   | 416   | 161 | 2    | 40    | 14  | 468   | 63  | 0    | 31    | 71  | 312   | 243 | 14   | 48    | 16  | 383   | 64  | 2    | 37    | 147 |
|       | obj   | 81    | 459 | 5    | 30    | 12  | 187   | 359 | 2    | 29    | 10  | 83    | 247 | 0    | 25    | 232 | 119   | 369 | 5    | 63    | 31  | 115   | 150 | 7    | 22    | 293 | 75    | 398 | 19   | 51    | 44  |
|       | iobj  | 3     | 0   | 42   | 1     | 1   | 2     | 3   | 39   | 1     | 2   | 3     | 1   | 35   | 1     | 7   | 4     | 2   | 37   | 3     | 1   | 2     | 0   | 27   | 0     | 18  | 1     | 0   | 42   | 1     | 3   |
| CNN   | nsubj | 470   | 89  | 5    | 41    | 28  | 268   | 266 | 5    | 62    | 32  | 402   | 136 | 5    | 52    | 38  | 234   | 141 | 3    | 46    | 209 | 196   | 335 | 4    | 63    | 35  | 189   | 107 | 6    | 39    | 292 |
|       | obj   | 74    | 418 | 23   | 42    | 30  | 249   | 230 | 21   | 35    | 52  | 73    | 177 | 14   | 28    | 295 | 193   | 247 | 31   | 44    | 72  | 61    | 126 | 18   | 14    | 368 | 100   | 370 | 15   | 39    | 63  |
|       | iobj  | 3     | 2   | 34   | 6     | 2   | 6     | 3   | 30   | 5     | 3   | 1     | 2   | 16   | 5     | 23  | 6     | 2   | 30   | 4     | 5   | 2     | 1   | 15   | 3     | 26  | 7     | 2   | 32   | 3     | 3   |
| FASTT | nsubj | 481   | 34  | 0    | 28    | 90  | 244   | 263 | 2    | 41    | 83  | 374   | 96  | 2    | 25    | 136 | 262   | 163 | 3    | 35    | 170 | 164   | 80  | 1    | 96    | 292 | 152   | 47  | 0    | 48    | 386 |
|       | obj   | 48    | 389 | 3    | 36    | 111 | 262   | 146 | 2    | 27    | 150 | 66    | 218 | 0    | 19    | 284 | 130   | 173 | 1    | 27    | 256 | 38    | 58  | 0    | 24    | 467 | 37    | 170 | 1    | 136   | 243 |
|       | iobj  | 5     | 1   | 27   | 8     | 6   | 15    | 5   | 11   | 8     | 8   | 3     | 0   | 14   | 4     | 26  | 18    | 1   | 7    | 4     | 17  | 1     | 0   | 7    | 2     | 37  | 4     | 0   | 11   | 12    | 20  |

**Figure 5.10:** Polish (top) and Russian (bottom) confusion matrices for the development splits of the *unambiguous* rotated counterfactual treebanks. WH stands for wrong head.

#### 5.4.2.2 Error analysis

To get a deeper understanding of the models’ errors, as in Section 4.4.2, I inspect the confusion matrices for the *nsubj*, *obj* and *iobj* relations. In Figure 5.10 and Figure 5.11 I present those matrices for the Slavic and Finnic language pairs, respectively. As before, I report both the *label* confusion matrix for the correctly attached tokens and the counts of tokens which were incorrectly attached (a column marked with *WH* for “wrong head”). I present the results for only one version of each model, trained with a random seed 1.

In Figures 5.10 and 5.11 I observe the same general error tendencies as for the word order experiments in Section 4.4.2. Here, I only briefly recap those tendencies and focus on the new types of errors that were caused by introducing the rotation of argument lexemes.

**Maintained tendencies** As before, the most common types of errors for both subjects and objects include (i) mislabeling as the other core argument and (ii) attaching to a wrong head. The misattachments are most common for word orders in which object and subject are not separated by the verb. For the Slavic languages it is mostly the later argument being attached to the preceding one (e.g., subjects incorrectly attached to objects in VOS

|       |       | SVO   |     |       |    | OVS   |     |       |    | SOV   |     |       |    | OSV   |     |       |     | VSO   |     |       |     | VOS   |     |       |     |
|-------|-------|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|-------|-----|
|       |       | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH  | nsubj | obj | other | WH  | nsubj | obj | other | WH  |
| BERT  | nsubj | 240   | 9   | 3     | 6  | 223   | 22  | 7     | 6  | 223   | 7   | 2     | 26 | 217   | 6   | 2     | 33  | 213   | 19  | 4     | 22  | 180   | 26  | 16    | 36  |
|       | obj   | 7     | 231 | 4     | 16 | 27    | 215 | 4     | 12 | 17    | 167 | 8     | 66 | 23    | 126 | 0     | 109 | 12    | 192 | 11    | 43  | 23    | 157 | 4     | 74  |
| CNN   | nsubj | 214   | 15  | 8     | 21 | 134   | 78  | 17    | 29 | 152   | 9   | 13    | 84 | 157   | 19  | 12    | 70  | 120   | 47  | 34    | 57  | 124   | 56  | 17    | 61  |
|       | obj   | 21    | 180 | 29    | 28 | 99    | 107 | 27    | 25 | 68    | 87  | 27    | 76 | 62    | 39  | 14    | 143 | 24    | 127 | 37    | 70  | 46    | 86  | 24    | 102 |
| FASTT | nsubj | 193   | 16  | 36    | 13 | 153   | 60  | 21    | 24 | 144   | 10  | 31    | 73 | 178   | 31  | 22    | 27  | 69    | 9   | 11    | 169 | 60    | 18  | 33    | 147 |
|       | obj   | 42    | 163 | 31    | 22 | 85    | 85  | 66    | 22 | 68    | 118 | 30    | 42 | 63    | 41  | 37    | 117 | 9     | 35  | 27    | 187 | 31    | 28  | 24    | 175 |

|       |       | SVO   |     |       |    | OVS   |     |       |    | SOV   |     |       |    | OSV   |     |       |    | VSO   |     |       |    | VOS   |     |       |    |
|-------|-------|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|-------|-----|-------|----|
|       |       | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH | nsubj | obj | other | WH |
| BERT  | nsubj | 118   | 0   | 2     | 7  | 110   | 8   | 5     | 4  | 101   | 8   | 4     | 14 | 111   | 8   | 2     | 6  | 97    | 13  | 1     | 16 | 61    | 34  | 19    | 13 |
|       | obj   | 2     | 118 | 5     | 2  | 14    | 108 | 2     | 3  | 12    | 80  | 1     | 34 | 1     | 71  | 3     | 52 | 1     | 91  | 12    | 23 | 4     | 71  | 7     | 45 |
| CNN   | nsubj | 104   | 4   | 6     | 13 | 79    | 26  | 12    | 10 | 75    | 3   | 9     | 40 | 98    | 2   | 10    | 17 | 47    | 47  | 10    | 23 | 46    | 46  | 18    | 17 |
|       | obj   | 4     | 101 | 10    | 12 | 53    | 58  | 10    | 6  | 49    | 19  | 7     | 52 | 13    | 18  | 7     | 89 | 12    | 39  | 16    | 60 | 9     | 48  | 9     | 61 |
| FASTT | nsubj | 95    | 13  | 2     | 17 | 24    | 72  | 15    | 16 | 52    | 17  | 5     | 53 | 73    | 17  | 7     | 30 | 10    | 23  | 31    | 63 | 7     | 24  | 24    | 72 |
|       | obj   | 0     | 89  | 16    | 22 | 49    | 53  | 8     | 17 | 30    | 46  | 4     | 47 | 14    | 28  | 6     | 79 | 1     | 20  | 17    | 89 | 1     | 28  | 24    | 74 |

**Figure 5.11:** Estonian (top) and Finnish (bottom) confusion matrices for the development splits of the *unambiguous* rotated counterfactual treebanks. WH stands for wrong head.

and OSV). The mislabeling errors seem to be largely caused by relying on the following signals:

- for subjects: (a) occupying the initial position and (b) directly preceding the verb
- for objects: (a) occupying the last position and (b) directly following the verb

All four signals appear to have an effect on each of the 12 models. However, for most Slavic models signal (a) for subjects and (b) for objects seem to be the most important (with the most errors following those patterns). Finnic models, on the other hand, seem to be most sensitive to signal (b) for subjects and signal (a) for objects.

**Errors introduced by rotation** I now highlight error types which were enhanced by introducing lexeme rotation to already reordered treebanks, by considering both the absolute and the relative differences between the models’ error counts in Figure 5.10 and Figure 5.11 and those from Section 4.4.2.

I notice that, across the models, SVO accrued the most errors in *relative* terms, while the other word orders accumulated most errors in *absolute* terms. For the latter, introducing variation in lexical semantics appears to enhance already existing error patterns. For example, I note a substantial amount of new misattachment errors for objects. For the Slavic languages these are most prominent for SOV and VSO word orders. For the Finnic languages, I note the largest increase in OSV, followed by VOS, SOV and VSO. All these

misattachment errors are the cause of the most substantial performance drops observed on object relations for BERT models in Table 5.6.

BERT-based parsers stand out as the most affected – with the largest absolute and relative increase in errors. Many of those are new types of errors for BERT; i.e., not evidenced in experiments on original counterfactual treebanks. These are dominated by mislabelings as the other core argument. The increase in mislabeling errors is most prominent for subjects in the VSO order in Slavic languages (104 new mislabelings of subject as object for the Polish BERT and 184 new mislabelings for the Russian BERT) and for subjects in VOS order in Finnish and Estonian (28 and 20 new mislabelings).

While there are some similarities between the error types that increase the most for BERT and for the other models, there is also large variety. For example, the Slavic FASTTEXT models accrued notable amount of misattachment errors for subject relations, while for the Slavic CNN models the misattachments for subject relations decreased, across most word orders. This aligns with the increased tendency to make misattachment errors observed for FASTTEXT in Section 4.4.2.

### 5.4.3 Discussion in the light of earlier results

By comparing the experimental results from this section to those from Section 4.4 (word order variations alone) and Section 5.3 (argument lexeme variations alone) it is possible to gain insight into which of the three – morphology, word order, lexical semantics – is *on average* the most important signal to core relations for a neural model. Note how each of the three experiments placed one signal in opposition to the other two. In Section 4.4 word order was placed in opposition to morphology + lexical semantics. The prominence of the word order pull was quantified through the decrease in LAS on subject/object relations in non-SVO clauses. In Section 5.3 lexical semantics was placed in opposition to morphology + word order<sup>12</sup>. There, the prominence of the lexicosemantic pull was quantified via the decrease in LAS on the rotated version of the data. Finally, in this section morphology is placed in opposition to word order + lexical semantics and the prominence of morphological pull is quantified by LAS performance on the *relevant conditions*; i.e., those where the word ordering directly opposes the statistical tendencies.

What follows is that if LAS for the relevant conditions in Table 5.6 is higher than the performance decrease observed for that model in Table 4.10 (word order results; unambiguous counterfactual treebanks) and Table 5.3 (rotation results; unambiguous filtered treebanks), one can infer that morphology is the most important signal. Conversely if, e.g., the differences between the SVO and the non-SVO orders in Table 4.10 are larger than the drop observed for the rotated data in Section 5.3 and the LAS recorded for the

---

<sup>12</sup>Since most of the UD clauses adhere to the dominant SVO ordering (see Section 4.1.2).

relevant conditions in Table 5.6, it is the word order that emerges as the most important linguistic signal. Such analysis reveals that:

- For BERT-based parsers, morphological signals are more important than word order and lexicosemantic signals. While the models do experience performance drops on the counterfactual treebanks and on the rotated treebanks, these drops (which can be attributed to relying on word order and lexical semantics, respectively) remain smaller than the performance on the rotated counterfactual treebanks (which can be attributed to relying on morphology), even for the most challenging conditions. However, even though the models rely on morphology to a larger extent than the alternatives (when all are considered in isolation), their dependence on morphology is far from near-categorical.
- For CNN and FASTTEXT-based parsers, word order is likely the most important cue. The drops observed across the non-SVO orders largely outweigh the performance on the rotated counterfactual treebanks from this section and the performance drops on the rotated treebanks in Section 5.3.

## 5.5 Other transformer models

In this section I demonstrate that the tendencies I observe in my evaluation from Section 5.4 are not specific to the models from Section 4.3, but rather exemplify a tendency to under-rely on morphology exhibited in a wider range of state-of-the-art neural models. I do so by fine-tuning a selection of alternative transformer-based models (on unaltered data) and evaluating them on the rotated counterfactual treebanks. To fine-tune the models on the dependency parsing task, I employ the bi-affine transformations of DM, but do not introduce a biLSTM; i.e., the fine-tuning setup here is equivalent to the DM -LSTM +BERT from Section 4.4.

I select a number of language-specific and multilingual models available in the Hugging Face (Wolf et al., 2019) record of pre-trained models<sup>13</sup>:

1. **RoBERTa** (Liu et al., 2019b), which uses the same architecture as BERT, but is pre-trained within a different scheme that makes use of different hyperparameters, altered objective and BPE tokenisation (Gage, 1994, Sennrich et al., 2016), instead of word-pieces (Johnson et al., 2017, Wu et al., 2016). It is also trained on different corpora. I evaluate the following RoBERTa instances:

**Polish** “sdadas/polish-roberta-base-v2”<sup>14</sup>

---

<sup>13</sup><https://huggingface.co/models>

<sup>14</sup><https://huggingface.co/sdadas/polish-roberta-base-v2>

|                                                | Polish<br>clause count: 295 |       |      |      | Russian<br>clause count: 632 |       |      |      | Finnish<br>clause count: 127 |       |      | Estonian<br>clause count: 258 |       |      |
|------------------------------------------------|-----------------------------|-------|------|------|------------------------------|-------|------|------|------------------------------|-------|------|-------------------------------|-------|------|
|                                                | all                         | nsubj | obj  | iobj | all                          | nsubj | obj  | iobj | all                          | nsubj | obj  | all                           | nsubj | obj  |
| DM +BERT (4 models), copied from Table 5.6     |                             |       |      |      |                              |       |      |      |                              |       |      |                               |       |      |
| svo                                            | 87.6                        | 89.7  | 87.4 | 83.1 | 90.7                         | 87.3  | 82.1 | 94.3 | 92.3                         | 94.1  | 95.1 | 89.3                          | 93.7  | 89.9 |
| ovs                                            | 82.0                        | 68.4  | 70.4 | 72.1 | 85.8                         | 73.2  | 65.6 | 87.2 | 88.8                         | 84.2  | 83.7 | 87.6                          | 87.3  | 81.1 |
| sov                                            | 75.3                        | 71.8  | 38.0 | 41.5 | 78.8                         | 67.3  | 43.1 | 80.2 | 83.0                         | 78.9  | 66.1 | 81.8                          | 87.1  | 65.9 |
| osv                                            | 80.9                        | 70.0  | 70.4 | 66.7 | 83.7                         | 70.0  | 66.1 | 82.3 | 81.8                         | 86.2  | 56.3 | 78.1                          | 86.0  | 50.8 |
| vso                                            | 67.8                        | 35.4  | 23.8 | 47.0 | 73.0                         | 42.5  | 32.3 | 78.0 | 83.8                         | 68.2  | 78.8 | 84.5                          | 83.7  | 76.5 |
| vos                                            | 78.7                        | 47.2  | 85.0 | 78.1 | 82.2                         | 56.2  | 72.5 | 90.1 | 75.2                         | 41.5  | 58.1 | 78.7                          | 71.9  | 59.9 |
| RoBERTa (1 model)                              |                             |       |      |      |                              |       |      |      |                              |       |      |                               |       |      |
| svo                                            | 89.4                        | 89.8  | 88.9 | 88.5 | 87.1                         | 81.8  | 78.5 | 72.3 | 90.9                         | 98.4  | 92.9 | 92.2                          | 95.7  | 92.6 |
| ovs                                            | 84.5                        | 74.6  | 79.6 | 72.1 | 71.9                         | 35.4  | 29.2 | 59.6 | 82.3                         | 81.9  | 73.2 | 88.6                          | 85.7  | 79.1 |
| sov                                            | 79.3                        | 69.2  | 54.5 | 67.2 | 75.3                         | 74.1  | 23.0 | 40.4 | 80.5                         | 91.3  | 54.3 | 86.4                          | 91.1  | 71.7 |
| osv                                            | 81.6                        | 69.5  | 71.5 | 72.1 | 72.5                         | 46.2  | 35.2 | 48.9 | 77.5                         | 94.5  | 40.9 | 80.6                          | 81.4  | 53.1 |
| vso                                            | 73.5                        | 47.1  | 41.3 | 59.0 | 70.0                         | 43.4  | 18.3 | 36.2 | 77.8                         | 60.6  | 68.5 | 87.7                          | 86.4  | 78.7 |
| vos                                            | 78.5                        | 46.1  | 80.0 | 78.7 | 74.8                         | 32.1  | 60.4 | 66.0 | 69.6                         | 41.7  | 48.0 | 81.9                          | 72.9  | 63.2 |
| GPT2 (1 model)                                 |                             |       |      |      |                              |       |      |      |                              |       |      |                               |       |      |
| svo                                            | 82.3                        | 84.4  | 84.7 | 77.0 | 90.5                         | 83.9  | 89.8 | 91.5 | 87.5                         | 95.3  | 82.7 | —                             | —     | —    |
| ovs                                            | 73.2                        | 54.9  | 60.4 | 47.5 | 81.8                         | 60.8  | 65.4 | 66.0 | 81.2                         | 74.8  | 69.3 | —                             | —     | —    |
| sov                                            | 69.6                        | 81.0  | 25.1 | 27.9 | 74.7                         | 64.2  | 28.5 | 59.6 | 76.4                         | 83.5  | 46.5 | —                             | —     | —    |
| osv                                            | 67.4                        | 45.8  | 53.2 | 39.3 | 76.1                         | 47.8  | 54.8 | 61.7 | 75.1                         | 91.3  | 35.4 | —                             | —     | —    |
| vso                                            | 63.4                        | 33.2  | 28.1 | 41.0 | 71.9                         | 42.2  | 29.9 | 66.0 | 66.0                         | 37.8  | 36.2 | —                             | —     | —    |
| vos                                            | 70.4                        | 29.2  | 84.7 | 68.9 | 80.4                         | 45.3  | 80.9 | 89.4 | 68.2                         | 40.9  | 48.0 | —                             | —     | —    |
| MBERT (bert-base-multilingual-cased) (1 model) |                             |       |      |      |                              |       |      |      |                              |       |      |                               |       |      |
| svo                                            | 87.5                        | 91.5  | 90.6 | 83.6 | 90.6                         | 87.2  | 86.0 | 87.2 | —                            | —     | —    | 86.8                          | 91.1  | 82.6 |
| ovs                                            | 80.4                        | 70.5  | 65.1 | 67.2 | 86.2                         | 74.4  | 71.2 | 74.5 | —                            | —     | —    | 83.9                          | 79.8  | 71.7 |
| sov                                            | 75.9                        | 75.6  | 42.1 | 45.9 | 81.6                         | 77.1  | 51.5 | 72.3 | —                            | —     | —    | 77.7                          | 77.1  | 59.3 |
| osv                                            | 77.9                        | 68.8  | 63.8 | 65.6 | 85.2                         | 73.7  | 74.1 | 72.3 | —                            | —     | —    | 72.9                          | 75.2  | 40.3 |
| vso                                            | 68.3                        | 35.3  | 35.7 | 47.5 | 74.8                         | 47.0  | 35.8 | 68.1 | —                            | —     | —    | 78.9                          | 76.4  | 62.0 |
| vos                                            | 75.2                        | 43.4  | 74.9 | 78.7 | 83.9                         | 61.9  | 79.0 | 83.0 | —                            | —     | —    | 77.6                          | 76.4  | 50.8 |

**Table 5.7:** Alternative transformer models: LAS scores on the development splits of the *unambiguous* counterfactual treebanks with core arguments *rotated*.

**Russian** “sberbank-ai/ruRoberta-large”<sup>15</sup>

**Finnish** “Finnish-NLP/roberta-large-finnish-v2”<sup>16</sup>

**Estonian** “EMBEDDIA/est-roberta”<sup>17</sup>

2. **GPT-2** (Radford et al., 2019); unlike BERT, which uses the encoder block from the original transformer model of Vaswani et al. (2017) (see Appendix C.4), GPT-2 is an autoregressive model which uses the original transformer’s decoder block (i.e., the model is unidirectional). It is trained with a language modelling objective and uses BPE tokenisation. I experiment on the following GPT-2 instances:

**Polish** “flax-community/papuGaPT2”<sup>18</sup>

**Russian** “sberbank-ai/rugpt3large\_based\_on\_gpt2”<sup>19</sup>

<sup>15</sup><https://huggingface.co/sberbank-ai/ruRoberta-large>

<sup>16</sup><https://huggingface.co/Finnish-NLP/roberta-large-finnish-v2>

<sup>17</sup><https://huggingface.co/EMBEDDIA/est-roberta>

<sup>18</sup><https://huggingface.co/flax-community/papuGaPT2>

<sup>19</sup>[https://huggingface.co/sberbank-ai/rugpt3large\\_based\\_on\\_gpt2](https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2)

3. **mBERT** (Devlin et al., 2019a) – a multilingual version of BERT trained on 104 languages with largest Wikipedias, including Polish, Russian, Finnish and Estonian.<sup>21</sup>

**Results** I present the LAS results for the new models on the development splits of rotated counterfactual treebanks in Table 5.7. What emerges from these results is that *all* the selected models are affected by the perturbations, regardless of the particulars of the architecture, hyper-parameters, pre-training objective/data or whether they are mono or multilingual. This signals that issues uncovered in Chapter 4 and in the previous sections from this chapter are not specific to the models evaluated in these experiments, but point to a more general trend that applies to many neural models forming the basis of modern state-of-the-art NLP. The results also reveal that other transformer-based models perform on-par with BERT, notably outperforming the alternative CNN and FASTTEXT inputs.

## 5.6 Conclusion

In this chapter I proposed two different methods of targeted evaluation of lexicosemantic effects in neural dependency parsing. The first method (Section 5.2) involved a simulation-based experimental design to study noun lexeme permutation effects on parsing performance. The second (Section 5.3), directly targeted core relation parsing and involved rotation of core argument lexemes. Both experiments involved *lexeme-based* perturbations – a novel way of perturbing data, which I proposed and advocated for in Section 5.1.

Together with the evaluation design proposed in Chapter 4, these methods form a comprehensive evaluation framework which allows us to estimate a model’s reliance on different linguistic signals to core grammatical relations, providing insights into its generalisation strategies. Within this framework I have shown that:

1. Many of the modern neural networks are subject to over-reliance on both word order and lexicosemantic cues to core relations when trained on UD dependency parsing in case-marking languages. This result emerged from experiments on 72 different models for 4 different languages<sup>22</sup> – *all* of which were substantially affected by changes in the core word order and lexical content of the verb’s arguments.
2. BERT-based (and other transformer-based models) rely on morphological signals to a far larger extent than CNN and FASTTEXT-based LSTM parsers, but they still

<sup>20</sup><https://huggingface.co/Finnish-NLP/gpt2-medium-finnish>

<sup>21</sup>See <https://github.com/google-research/bert/blob/master/multilingual.md> for more details on how the multilingual BERT was trained.

<sup>22</sup>4 x fine-tuned BERT (4 seeds), 4 x DM +BERT (4 seeds), 4 x DM +CNN, 4 x DM +FASTTEXT (4 seeds), 4 x DM +CNN (4 seeds), 4 x fine-tuned RoBERTa, 3 x fine-tuned GPT-2, 1 x mBERT

remain prone to errors; i.e., they are far from recognising morphology as the only systematic linguistic generalisation.

3. Neural models are inherently biased towards particular types of signals to core sentence meaning, with different biases exhibited across the models. For BERT-based parsers of Polish, Russian, Finnish and Estonian morphology appears to be the most important linguistic signal to core sentence meaning. This is in contrast to CNN and FASTTEXT-based models which rely on word order more than morphology.
4. The presence/absence of morphosyntactic ambiguity has an effect on the extent to which the models rely on both word order and lexical semantics. When morphology is ambiguous: (i) all models tend to rely on word order more than when it is unambiguous and (ii) BERT-based and CNN-based parsers also tend to rely more on lexical semantics. These tendencies bear a weak resemblance to the parsing strategies of native speakers (see Section 2.5).

All of the above are important insights, which highlight shortcomings and differences among the models, which do not emerge from standard, non-targeted evaluation (see Chapter 3). To the best of my knowledge, I am the first to bring these issues to light and to highlight important areas in which the models should be improved, setting a direction for building better models, with linguistic signal sensitivities suited to the language they are developed for. In my next chapter (Chapter 6) I take the first step in that direction by exploring two parallel approaches to improving the model’s morphological competence: (i) augmentation of the training data and (ii) multi-task training.

## Part II

### Particulars of models' shortcomings



---

## INCREASING MORPHOLOGICAL AWARENESS OF DEPENDENCY PARSERS

---

In previous chapters (Chapter 4 and Chapter 5) I revealed that modern neural models (both transformer and LSTM-based) tend to over-fit to word order and lexicosemantic cues to core relations while parsing case-marking languages. This consequently means that the models under-rely on morphology – the only reliable indicator of core sentence meaning (see Section 2.4).

While none of the models I experimented with in the earlier chapters passed my proposed tests for morphological competence (i.e., reached close to 100% reliance on morphology as a signal to meaning), some models exhibited promising tendencies. These include: (i) relying on morphology as the most important signal on average (BERT-based parsers only) and (ii) increasing reliance on word order (all parsers) and lexical signals (BERT and CNN-based parsers) when morphology is ambiguous. When viewed in the context of related research (see Section 3.4), these tendencies suggest that the models (in particular those based on BERT) extract a lot of relevant morphosyntactic information, but fail to *use* it as a signal to core dependencies, due to the combined effect of (i) the strong word order and lexical correlations with core sentence meaning exhibited within a language and (ii) the models’ inadequate relative input signal sensitivities (RISS, defined in Section 3.1). The above hypothesis bears close resemblance to the **missed connection hypothesis** of Min et al. (2020) and it predicts that the models can be successfully nudged towards the correct linguistic generalisation by altering their training data/objective and without any change to their architectures/inputs.

In this chapter, I directly test the above by exploring two straightforward ways of increasing neural models’ morphological competence: (i) training the models on data in which the word order and lexical correlations with subject/objecthood have been tempered through perturbations (Section 6.1), and (ii) introducing another, morphologically-oriented

training objective – morphosyntactic tagging (Section 6.2). Through experimentation on two languages – Polish and Finnish – on three different types of models I demonstrate that the first, albeit very simple, is often sufficient to push the models towards the correct approach, yielding better, more morphologically-oriented parsers. I also show that the second technique – introducing an additional objective – is insufficient to sway the models towards relying on morphology.

## 6.1 Breaking word order and lexical correlations

In this section I test whether reducing the word order and lexical signals in the training data can push the models towards relying on morphology to a greater extent, without affecting the overall performance of the models. I conduct a series of experiments in which I train the Polish and Finnish parsers from Section 4.3 on data with perturbed clauses. I consider three different perturbations. Two of them concern exclusively *transitive* clauses with both arguments explicit – one perturbation involves reordering of the core clausal elements, the other involves word reordering *and* argument lexeme rotation. The third perturbation concerns *all* clauses. It involves a reordering of S, V and O, as well as a random permutation of nouns linked via core relations and relations that should not be affected by lexeme changes (I introduced those relations in Section 5.2.1.1; see Table 5.2). I discuss all three perturbations in more detail in the later Section 6.1.1.1.

For each considered [language, input, training perturbation] combination I train 4 models – each trained on data with a different proportion of clauses perturbed (15%, 45%, 75%, 100%). This is done to gain insight into how much data is required to sway the models towards relying on morphology and yields 72 models in total (2 languages  $\times$  3 different inputs  $\times$  3 perturbations  $\times$  4 proportions).

Note that because I break the word order and/or lexical correlations with subject/objecthood the models will almost certainly reduce their reliance on those signals when predicting core relations. However, it is not clear whether such interventions will increase the models’ reliance on morphology and whether they will do so without hurting the overall performance.

**Related work** Training data augmentation/perturbation is a widely explored method of pushing neural models towards the correct, linguistic generalisations. Much research from the track of linguistic analysis of neural models (see Section 3.4) proved this technique to be very effective in swaying the models towards relying on the relevant linguistic signals (Davis and van Schijndel, 2021, McCoy et al., 2019, Min et al., 2020, Pham et al., 2021, Şahin and Steedman, 2018, Wu et al., 2022). McCoy et al. (2020b) also highlights such method as a way of decreasing “the probability of there being local minima that ignore

certain phenomena” in their recent exploration of generalisation of English neural models.

### 6.1.1 Experimental details

I experiment on Polish PDB (Wróblewska, 2018) and Finnish TDT (Haverinen et al., 2014) treebanks. For each, I create 12 different versions of the train split which differ in (i) the perturbation applied and (ii) how many of the sentences are perturbed. For the latter, I consider 4 different proportions: 15%, 45%, 75% and 100%. For each proportion, I randomly draw sentences to be perturbed from the sentences in the original train split. For perturbations focusing on transitive clauses, I only consider sentences which have a transitive verb with two noun arguments (i.e., 15% means that 15% of sentences with *transitive* clauses were perturbed). I only consider one randomly drawn set for each threshold to save computational resources (training only one model for each proportion already amounts to 72 models).

#### 6.1.1.1 Training-data perturbations

**PERTURBATION I. Core element reordering (transitive clauses)** This perturbation is very similar to the one used to create the counterfactual treebanks from Chapter 4. If applied to all clauses, it effectively breaks the correlation between word order and core grammatical function. For each transitive sentence selected for perturbation I create its new version by randomly sampling a new ordering of S, V and O for each verb with at least one argument and restructure the sentence accordingly (see Section 4.2.3 for more details). Importantly, to expose the models to different word orders in a broader range of sentences and not just simple constructions, I *do not* simplify the sentences to ensure their acceptability, as I did in Section 4.2. While this can yield some challenging sentences with longer dependencies, it maintains grammaticality and is unlikely to have a detrimental effect on the models’ performance on well-formed evaluation sentences.

**PERTURBATION II. Core element reordering + lexeme rotations (transitive clauses)** This perturbation is designed to break both the word order and the lexical correlation with core grammatical function. It not only randomly reorders a transitive clause (see above) but also rotates the lexemes of the transitive verb’s core arguments with a probability of 50%. The chance element of the rotation is introduced to ensure that the correlation between the lexical properties of the nouns and their grammatical function is decreased, rather than reversed. The rotation applied here is akin to that in Section 5.3, where the core argument lexemes were swapped. The difference between the two is that now the perturbed clause can be ditransitive<sup>1</sup>, so the lexemes are shifted by 1

---

<sup>1</sup>The filtering of the evaluation data in Section 5.3 excluded ditransitive clauses.

slot to the right (rotated by 1), instead of being swapped. This means that, in a typical SVO ditransitive clause, the subject becomes the direct object, the direct object becomes the indirect object and the indirect object becomes the subject.

**PERTURBATION III. Core element reordering (all clauses)** This perturbation differs from the other two in that it applies to *any clause*, including intransitive. It randomly reorders a clause (see PERTURBATION I) and permutes lexemes of nouns linked via core relations and relations that should not be affected by lexeme changes (see Table 5.2) in selected sentences. Here, I apply lexeme permutations instead of rotations because rotations can only be applied to verbs with more than one noun argument.

By including more clauses, PERTURBATION III can have a stronger effect on the model than the other two because (i) at each step it is applied to a larger portion of the training data (at 100% all sentences are perturbed) and (ii) it guides the model towards applying correct generalisation strategies while parsing a wider range of clauses, and not just those with transitive verbs.

#### 6.1.1.2 Evaluation methods

I evaluate each model within the framework proposed in the previous chapters, testing to what extent it relies on (i) word order over morphology + lexical semantics, (ii) lexical semantics over morphology + word order and (iii) morphology over word order + lexical semantics. To enable easy visual tracking of how the models' strategies change as they are trained on more perturbed data, I quantify each of the three with a single number:

I quantify the **strength of reliance on word order** using the SVO *attachment bias* (LAS) metric (see Section 4.2.4) calculated for unambiguous counterfactual treebanks (Section 4.2).

I quantify the **strength of reliance on lexicosemantic cues** using the difference between LAS calculated for the (unaltered) sentences with unambiguous transitive clauses (see Section 5.3.1) and LAS calculated for the same sentences, but with core argument lexemes rotated (this difference is equivalent to O-R from Table 5.3).

I quantify the **strength of reliance on morphology** using LAS averaged across chosen rotated counterfactual treebanks (see Section 5.4). For *nsubj* I average OVS, VSO and VOS. For *obj* I average OVS, OSV and SOV (see Section 5.4.1 for rationale behind choosing these word orders).

In addition to the above, I also evaluate the models on original, unaltered and unfiltered treebanks (i.e., standard evaluation) to track whether the perturbations applied at training hurt their overall performance in any way.

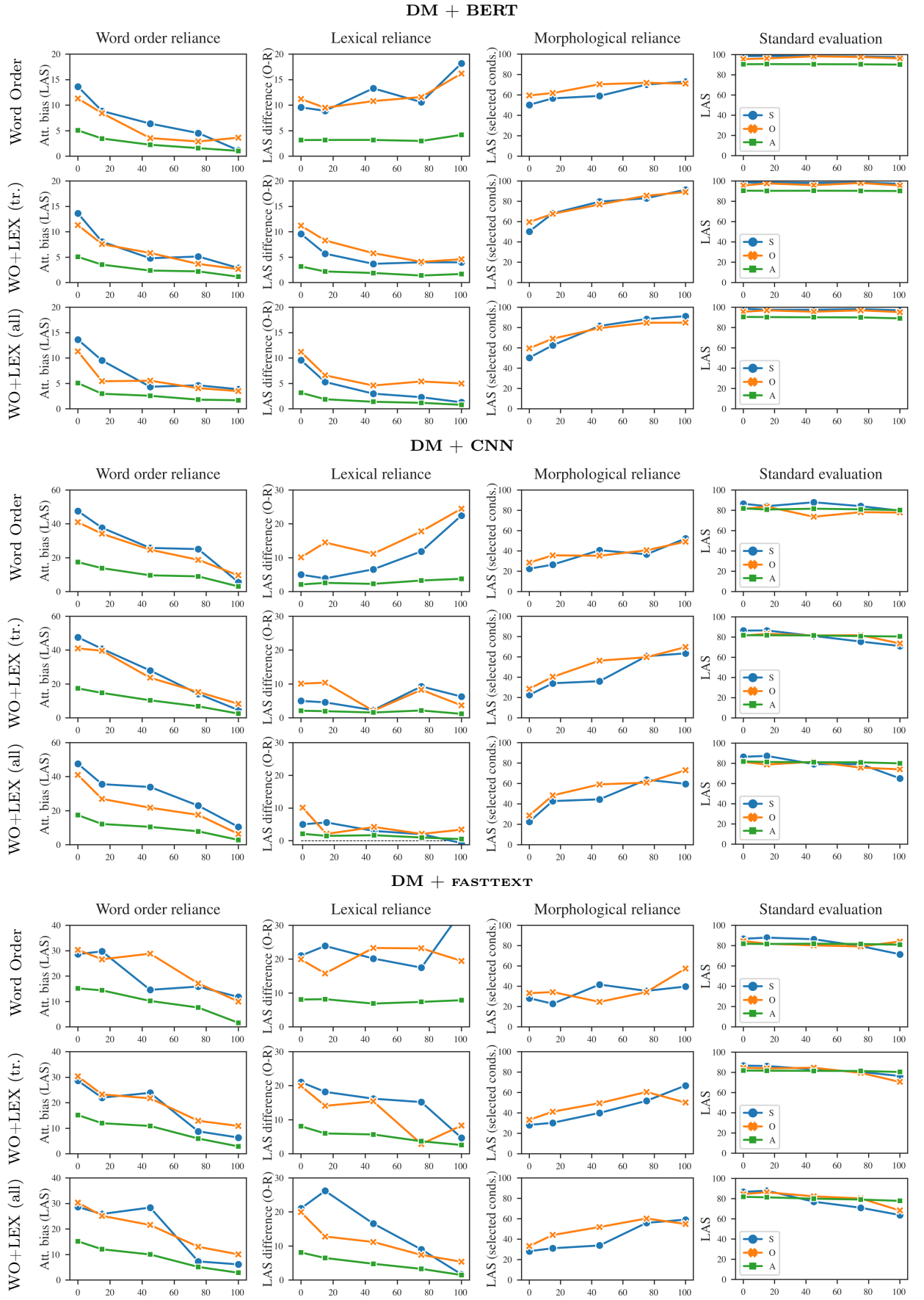


**Figure 6.1:** Changes in Polish models' generalisations as they are trained on more perturbed data. The x axes show proportion of perturbed clauses. S, O and A stand for *nsubj*, *obj* and *all*.

### 6.1.2 Results (unambiguous clauses)

In this subsection I discuss the results of evaluation on *unambiguous* clauses; see the following Section 6.1.3 for evaluation on ambiguous clauses.

In Figure 6.1 and Figure 6.2 I present the results for all Polish and Finnish models, trained on the perturbed splits of the PDB/TDT treebank and evaluated on the development splits of the *unambiguous* counterfactual treebanks (word order reliance), treebanks filtered to contain only *unambiguous* transitive sentences (lexical reliance) and rotated *unambiguous* counterfactual treebanks (morphological reliance). In most plots I present three scores: one for *nsubj* relations, one for *obj* relations and one for *all* relations.

I break down the discussion of these results as follows. First, I discuss the trends observed for the models trained on data in which only word order is altered – PERTURBATION I, rows **WORD ORDER** in the figures. Next, I discuss how introducing lexical perturbations – PERTURBATION II, rows **WO+LEX (tr.)** – changes the generalisation trends observed for the word-order-only perturbation and how perturbing all clauses – PERTURBATION III, rows **WO+LEX (all)** – differs from perturbing only transitive clauses.

#### 6.1.2.1 Word order perturbation only (PERTURBATION I)

Simple reordering of core clausal elements in transitive clauses notably decreases the models’ reliance on word order as a signal to core relations – the SVO attachment bias (LAS) drops for all considered parsers. As expected, the more clauses in the training data are perturbed the stronger this effect is. When 100% of transitive clauses are perturbed, the SVO bias is particularly low for BERT-based models (close to 0 for Finnish core relations and  $< 5$  for Polish core relations).

As the reliance on word order decreases the reliance on morphology *increases*, but so does the reliance on lexical semantics. The latter trend is particularly clear for the Polish models. Notably, this shift of generalisation strategies comes at no real performance cost for BERT-based parsers – the performance on standard evaluation remains largely constant as the proportion of perturbed training data increases. For CNN and FASTTEXT-based parsers I note a larger drop in overall performance (standard evaluation), suggesting that they cannot successfully leverage the lexical and morphological signals to the same extent as the cue gestalts, where all cues cooperate.

To provide insight into how the performance changes across the different word orders, in Table 6.1 I report the LAS results on each differently ordered counterfactual treebank for the models trained with 100% of transitive clauses perturbed. This table reveals that the word order alteration equalises the performance across the word orders, leading to substantial performance improvements on the less common orders. Importantly, this often happens at the cost of LAS performance on the SVO clauses. The drop for SVO is the

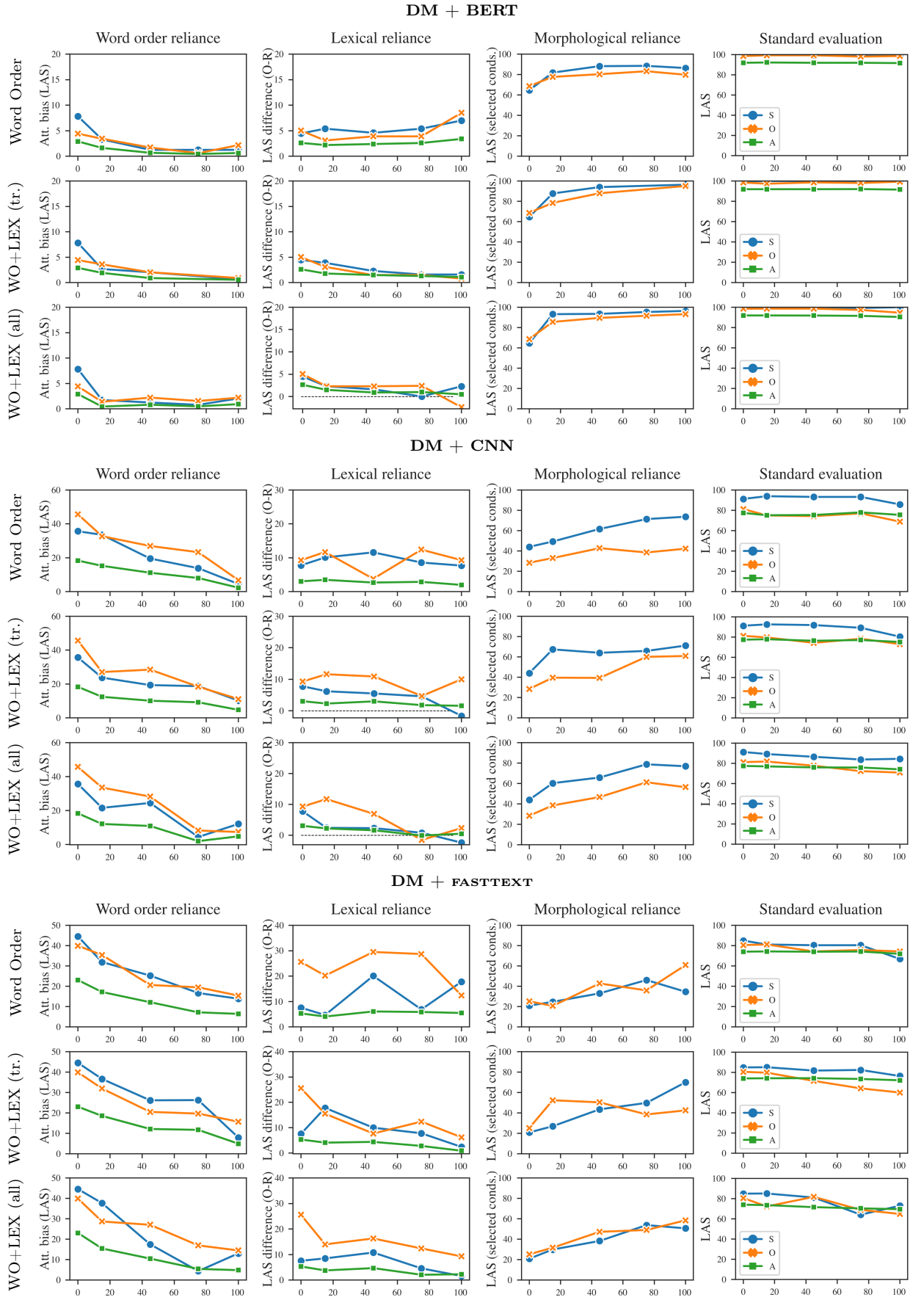


**Figure 6.2:** Changes in the Finnish models' generalisations as they are trained on more perturbed data. The x axes show proportion of perturbed clauses. S, O and A stand for *nsubj*, *obj* and *all*.

|                                                  | Polish<br>clause count: 295 |       |      |      | Finnish<br>clause count: 127 |       |      |  | Polish (gain)<br>clause count: 295 |       |       |       | Finnish (gain)<br>clause count: 127 |       |       |
|--------------------------------------------------|-----------------------------|-------|------|------|------------------------------|-------|------|--|------------------------------------|-------|-------|-------|-------------------------------------|-------|-------|
|                                                  | all                         | nsubj | obj  | iobj | all                          | nsubj | obj  |  | all                                | nsubj | obj   | iobj  | all                                 | nsubj | obj   |
| DM +BERT, trained with PERTURBATION I (100%)     |                             |       |      |      |                              |       |      |  |                                    |       |       |       |                                     |       |       |
| svo                                              | 93.5                        | 97.3  | 98.7 | 86.9 | 97.6                         | 100   | 98.4 |  | -0.7                               | -2.5  | 0.4   | -1.6  | 0.1                                 | 0.2   | -0.4  |
| ovs                                              | 92.9                        | 97.3  | 97.0 | 86.9 | 97.7                         | 99.2  | 99.2 |  | 1.5                                | 3.4   | 5.0   | 8.6   | 1.3                                 | 0.8   | 0.2   |
| sov                                              | 92.8                        | 98.0  | 96.2 | 82.0 | 97.6                         | 99.2  | 98.4 |  | 2.2                                | 1.6   | 12.3  | 10.7  | 2.4                                 | 2.5   | 5.7   |
| osv                                              | 92.7                        | 95.9  | 95.7 | 86.9 | 97.2                         | 99.2  | 96.9 |  | 1.7                                | 3.3   | 5.6   | 8.6   | 1.0                                 | 0.2   | 2.4   |
| vso                                              | 91.9                        | 96.9  | 91.9 | 77.0 | 96.9                         | 100   | 95.3 |  | 8.0                                | 27.3  | 16.8  | 6.9   | 2.3                                 | 9.6   | -1.4  |
| vos                                              | 91.9                        | 93.9  | 94.5 | 83.6 | 95.6                         | 96.1  | 92.9 |  | 3.3                                | 15.8  | 0.7   | -0.8  | 5.0                                 | 20.7  | 3.5   |
| DM +CNN, trained with PERTURBATION I (100%)      |                             |       |      |      |                              |       |      |  |                                    |       |       |       |                                     |       |       |
| svo                                              | 80.3                        | 75.3  | 79.6 | 62.3 | 78.9                         | 85.0  | 64.6 |  | -5.8                               | -18.2 | -9.6  | -10.2 | -7.3                                | -7.3  | -22.1 |
| ovs                                              | 81.1                        | 79.0  | 80.4 | 59.0 | 77.8                         | 80.3  | 65.4 |  | 11.0                               | 33.3  | 38.5  | 15.1  | 5.1                                 | 19.9  | 12.3  |
| sov                                              | 77.2                        | 75.3  | 66.0 | 60.7 | 76.8                         | 81.1  | 59.8 |  | 4.8                                | -9.0  | 30.1  | 28.8  | 13.0                                | 22.6  | 32.8  |
| osv                                              | 76.3                        | 66.1  | 72.8 | 62.3 | 78.8                         | 90.6  | 51.2 |  | 9.8                                | 22.3  | 34.6  | 24.2  | 8.8                                 | 5.3   | 26.2  |
| vso                                              | 78.4                        | 80.3  | 66.8 | 52.5 | 74.5                         | 81.1  | 59.8 |  | 12.0                               | 42.0  | 20.4  | 14.0  | 8.6                                 | 39.2  | 11.0  |
| vos                                              | 74.7                        | 65.1  | 65.5 | 60.7 | 74.8                         | 80.3  | 54.3 |  | 6.5                                | 47.7  | -12.9 | -4.9  | 7.9                                 | 43.1  | 3.3   |
| DM +FASTTEXT, trained with PERTURBATION I (100%) |                             |       |      |      |                              |       |      |  |                                    |       |       |       |                                     |       |       |
| svo                                              | 78.2                        | 63.7  | 86.8 | 62.3 | 76.8                         | 70.9  | 80.3 |  | -8.0                               | -25   | -1.6  | -15.6 | -5.8                                | -17.9 | -4.2  |
| ovs                                              | 78.0                        | 83.7  | 74.9 | 55.7 | 72.9                         | 59.8  | 74.8 |  | 6.6                                | 22.0  | 21.5  | 28.3  | 6.2                                 | 14.7  | 25.2  |
| sov                                              | 74.4                        | 55.3  | 76.2 | 54.1 | 72.9                         | 46.5  | 75.6 |  | 0.1                                | -16.8 | 14.2  | 11.5  | 6.0                                 | -3.1  | 11.2  |
| osv                                              | 76.1                        | 72.9  | 76.2 | 50.8 | 69.6                         | 68.5  | 52.8 |  | 8.5                                | 5.6   | 35.2  | 28.3  | 2.7                                 | -9.5  | 26.6  |
| vso                                              | 78.4                        | 75.6  | 72.8 | 57.4 | 67.2                         | 55.1  | 60.6 |  | 10.2                               | 19.0  | 16.7  | 9.9   | 20.6                                | 27.4  | 16.1  |
| vos                                              | 80.1                        | 73.2  | 83.8 | 67.2 | 69.2                         | 55.1  | 60.6 |  | 6.7                                | 30.4  | 5.9   | 5.7   | 19.0                                | 34.5  | 22.4  |

**Table 6.1:** Two left columns: LAS on the dev split of the counterfactual treebanks (no lexeme rotations) for models trained on data with 100% of transitive clauses randomly reordered. Two right columns: LAS gained by training on the perturbed data (left columns minus the results from Table 4.10).

most prominent for the CNN and FASTTEXT-based parsers and remains minor for the BERT-based parsers ( $\leq 2.5$  LAS across all relations in both languages). While it is possible that more elaborate perturbations could close this SVO gap, it is not obvious whether this could be done with perturbations alone.

### 6.1.2.2 Introducing lexical perturbation (PERTURBATION II)

The decrease in the SVO bias can also be observed for PERTURBATION II which alters both word order and lexical semantics.<sup>2</sup> For most models this drop is also accompanied by the drop in reliance on lexical cues. Indeed, morphology is the only signal on which all models *increase* their reliance, as more data is perturbed. And when lexical semantics is accounted for, the models tap into morphology even more than when only word order is perturbed; with 100% of transitive clauses perturbed with PERTURBATION II, the Finnish BERT parser reaches 96 LAS (!) on subject relations and 95 LAS (!) on object relations for the relevant conditions in rotated counterfactual treebanks. For Polish BERT these numbers are similarly high – 90 LAS for subject and 89 LAS for object.

The above numbers prove that the BERT-based parsers do extract the relevant in-

<sup>2</sup>While I do not report these results here, inspection of LAS performance on the counterfactual treebanks reveals that for CNN and FASTTEXT the equalising of performance across word orders comes at a cost of performance drop on the SVO relations, as for PERTURBATION I.

flexional information and *can* use it as a signal to core sentence meaning, but their architecture and/or pre-training predisposes them not to rely on that signal. They only reach a close to 100% dependence on morphology (in unambiguous sentences) when it is the only available linguistic signal. On one hand, this result is very promising – training data alteration is enough to nudge the models in the correct direction, without hurting the overall performance. On the other hand, it highlights that BERT’s RISS (and from what was shown in Section 5.5 likely also RISS of other transformer models) is inadequate for case-marking languages.

**How much perturbation is enough?** For the Polish BERT-based parser *all transitive clauses* need to be perturbed to reach the best morphological generalisation; i.e., the non-morphological correlations with core grammatical function have to be fully broken. The growth of morphological reliance as the proportion of perturbed data increases is largely linear for both subject and object relations. This suggests that the model requires a lot of ‘persuasion’ to rely on morphology as the dominant signal and that even small traces of word order/lexical correlations with core grammatical function in transitive clauses push it to lean into those signals when making their predictions.

For the Finnish BERT, the curve for morphological reliance reveals a different tendency – the model experiences a prominent rise in the LAS on relevant conditions after as little as 15% of transitive clauses perturbed. After that, the gains become more modest. This suggests that the Finnish model is more inclined to tap into morphological signals than its Polish counterpart. Foreshadowing, my experiments from Chapter 7 suggest that this difference is likely to stem from the differences in the morphological systems of the two languages, such as different patterns of case syncretism or different degrees of allomorphy in case marking.

### 6.1.2.3 Altering more types of clauses (PERTURBATION III)

I notice little difference in the general trends for models trained with PERTURBATION III, when compared to PERTURBATION II. Some minor discrepancies include the decreased reliance on lexicosemantic signals of the first group of models and in some cases an earlier plateau on the morphological reliance measure. Both are unsurprising; PERTURBATION III alters the lexemes of more nouns than PERTURBATION II and at each proportion step perturbs more data overall since the proportion is calculated for all sentences and not just those with a transitive verb. As for PERTURBATION II, for BERT-based parsers, the standard evaluation performance remains stable as the proportion of perturbed data increases. This is a very good sign, since it suggests that even though the perturbations are less localised, they successfully change the models’ generalisation strategies, without hurting the overall performance.

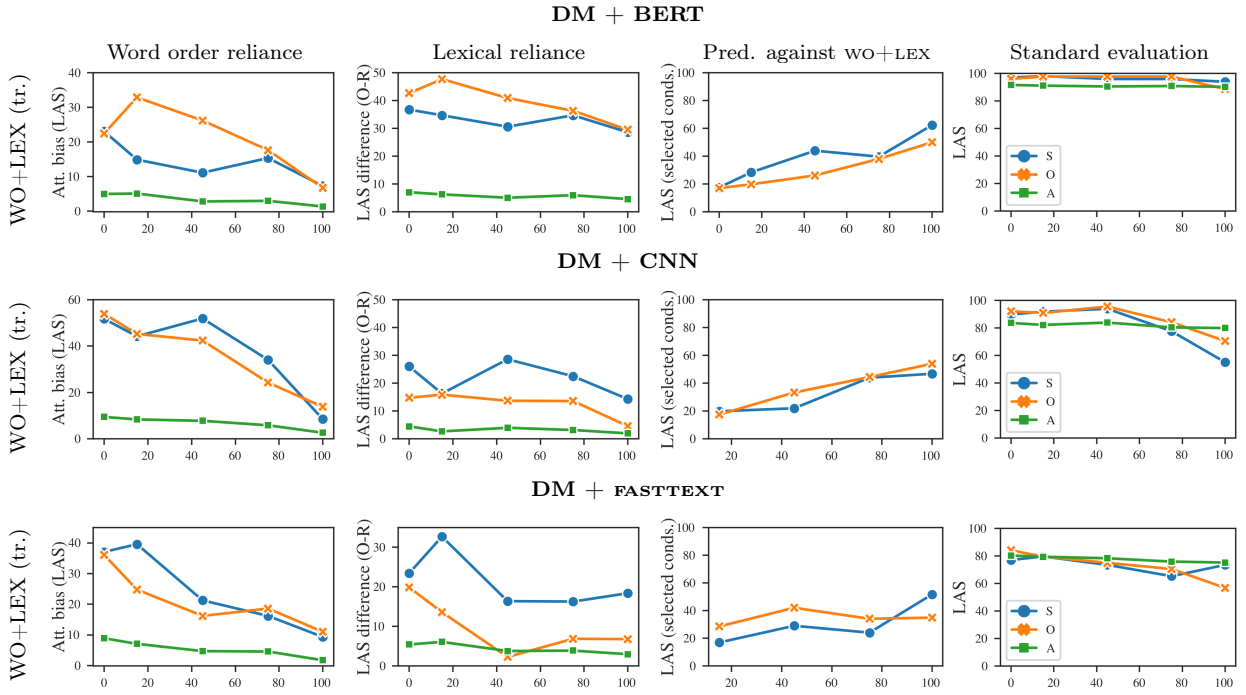


**Figure 6.3:** Evaluation on ambiguous clauses. Changes in Polish models’ generalisations as they are trained on more perturbed data. S, O and A stand for *nsubj*, *obj* and *all*.

### 6.1.3 Results (ambiguous clauses)

The perturbations proposed in Section 6.1.1.1 alter selected sentences regardless of whether the morphology is ambiguous. Although the full clausal ambiguity is a rare occurrence (see Section 4.1.1), this means that, at highest proportions of perturbed clauses, the models are inevitably exposed to ambiguous clauses with no reliable disambiguating signal. In this section I test whether such exposition affects their ability to rely on lexical semantics and word order as disambiguating strategies. I do so by evaluating the Polish models trained on data with various proportions of clauses perturbed (the same models as in Section 6.1.2) using variations of metrics from Section 6.1.1.2 which apply to *ambiguous* clauses.<sup>3</sup> I do not evaluate the Finnish models due to limited amounts of ambiguous clauses remaining after the filtering (see Section 4.5).

Note that given the morphosyntactic ambiguity, metrics from Section 6.1.1.2 have different interpretations to when they are applied to unambiguous clauses. For example, in *word order reliance*, word order no longer stands in opposition to both lexical semantics and morphology, but only to lexical semantics. An equivalent also applies to the *lexical reliance*, meaning that, in principle, when calculated on ambiguous clauses, these metrics measure largely the same thing but in different ways and on different data (word order metric is calculated based on simplified sentences for reasons discussed in Section 4.2.2). Further, the ambiguous equivalent of morphological reliance no longer measures reliance on morphology,

<sup>3</sup>See Section 4.2.1 for more information on how I filter the data to exclude unambiguous sentences.

but how often the models make predictions that go against the collaboration of word order and lexicosemantic signals, which both point towards the *incorrect* interpretation.

The Polish results for the WO+LEX (**tr.**) perturbation (PERTURBATION II), presented in Figure 6.3, indicate that the more clauses are perturbed in the training data, the lesser the models’ tendency to rely on word order and lexical semantics as disambiguators. This is manifested in graphs titled ‘Prediction against WO+LEX’ (column 3 in Figure 6.3). In the absence of morphological markings, the models base their subject/object predictions on (collaborating) word order and lexical semantics only in ~50% of instances, suggesting that their selection of object vs subject label for each argument is largely down to chance (the UAS remains close to 100 in all considered conditions). The perturbations also affect the models’ overall performance on ambiguous sentences as manifested in the ‘standard evaluation’ graphs, with the CNN and FASTTEXT-based parsers most affected.

### 6.1.4 Discussion

The above results demonstrate that without the ability to rely on word order and lexical semantics the models *do* pick up on morphological signals to a much greater extent. This proves that they have a greater capacity to rely on morphology than what they settle for when trained on original UD treebanks. My results also reveal that the BERT-based parsers can change their generalisation strategies with almost no effect on their overall accuracy in standard evaluation. Further, I have shown that an increased reliance on one of the three – word order, lexical semantics, morphology – is closely tied to a *decreased* reliance on the other one/two. This supports my hypothesis that these linguistic signals are the key basis of the models’ solutions and that they are in direct competition (see Section 2.7).

But while training data alteration can successfully improve morphological generalisation of neural parsers, it has a downside: it is not clear how introducing unacceptable sentences at train time would affect the models’ performance down the line. While I have not noticed any big performance drops on *all* relations in standard evaluation, further analysis is required to ensure the models perform as well as those trained on unaltered language. To add to the above, the widely applied alterations appear to affect the models’ ability to rely on word order and lexical semantics as disambiguating strategies (see Section 6.1.3) – an undesirable side-effect which might be challenging to overcome. And finally, this method only compensates for the inadequate RISS prevalent in the models. A more long term solution is to directly target signal preferences by creating models with opposing biases which work well when trained on natural data.<sup>4</sup>

---

<sup>4</sup>Note how training data alterations do not change the signal sensitivities of the models. Rather, they take away the option to rely on the alternative signals (see Section 3.1.4 for a more detailed discussion).

|    | nouns        | verbs                                       | adjectives/numerals/determiners |
|----|--------------|---------------------------------------------|---------------------------------|
| PL | case, number | person, number, gender, animacy, tense/mood | case, number, gender, animacy   |
| RU | case, number | person, number, gender, tense/mood          | case, number, gender, animacy   |
| FI | case, number | person, number, polarity, tense/mood        | case, number                    |
| ET | case, number | person, number, polarity, tense/mood        | case, number                    |

**Table 6.2:** Inflectional features for different POS in Polish, Russian, Finnish and Estonian.

## 6.2 Multi-task training

In this section I test whether introducing a morphosyntactic tagging objective can sway the models to rely on morphology to a greater extent. Here, I expect that such objective will force the models to retrieve the morphosyntactic information which communicates the core relations. However, it is unclear whether this encouragement to pay attention to subword content will be enough to encourage the models to *use* this information as a signal to core sentence meaning.

### 6.2.1 Experimental details

**Models** To isolate the effects of morphosyntactic tagging, I experiment with models which differ from those in Section 4.3 only in the new tagging component and the new objective. The new objective is to predict inflectional, non-lexeme specific features for each token in a sequence (see Table 6.2). Note that this differs from the morphosyntactic tagging objectives employed by other multi-task models for UD parsing, such as UDify (Kondratyuk and Straka, 2019) or UDPipe (Straka, 2018), where *all* UD features are predicted. Here, I decide to include only a subset of all features for two reasons: (i) only inflectional features are relevant for agreement and government, and (ii) the UD annotation includes a lot of features which are not inflectional, such as *VerbForm*, *Foreign* or *NumType* (see appendix Table A.3).

The new tagging component shares the encoder with the dependency parser, but introduces a new softmax layer. The softmax is applied to each word in an input sequence to compute the probability distribution over the possible morphosyntactic tags. I represent each morphosyntactic tag as a separate token in the tag vocabulary; i.e., I do not predict values for individual features independently. This follows Kondratyuk and Straka (2019), who found this approach to produce higher accuracy.

In addition to the above models I also experiment with UDify (Kondratyuk and Straka, 2019) – an existing, well performing multi-task parser, which achieved first place in the SIGMORPHON 2019 Shared Task on morphological analysis (McCarthy et al., 2019b). It is a multilingual model based on the multilingual BERT (Devlin et al., 2019a) and was trained to predict: (i) a POS tag (UPOS), (ii) a full morphosyntactic tag (UFeats), (iii) a

lemma and (iv) a dependency head and label per each token in an input sequence.

**Evaluation** As in Section 6.1, I test the multi-task models’ generalisation strategies by analysing to what extent they rely on each of the plausible linguistic signals in a one vs rest setup. I use the same measures as in Section 6.1.1.2 to quantify a model’s reliance on each signal. Throughout all experiments, I compare the values of those measures for the multi-task models to the values of corresponding single-objective models from Section 4.3.

## 6.2.2 Results

I present the results for my multi-task models in Figure 6.4. From those results it is clear that the incorporation of an additional morphosyntactic tagging objective, like the one proposed here, is *not enough* to increase the models’ reliance on morphology – across the relations and languages the increase in morphological reliance is either minimal or non-existent. CNN-based parsers benefit from the additional objective the most. This suggests that the failure of these models to rely on morphology in the experiments from the earlier chapters partially stems from the dependency parsing objective being insufficient for the model to learn to retrieve the relevant morphosyntactic information.

For CNN-based parsers I notice that in many cases their increase in reliance on morphology goes in hand with increase in reliance on lexical semantics, especially for the object relations (see e.g., the lexical reliance plots for Finnish and Estonian). For FASTTEXT and BERT this trend is reversed – reliance on lexical semantics tends to decrease for the object relations, across all languages.

The results for UDify, presented in Figure 6.5, resemble those for my multi-task models. For this model, trained with a more elaborate multi-task objective, the reliance on morphology is never greater than 75 LAS, across both core relations and all four languages. The model also displays signs of over-reliance on word order and lexical semantics.

### 6.2.2.1 Morphosyntactic tagging accuracy

In this subsection I ask two follow-up questions to get a deeper understanding of the earlier results. The first question is whether the lack of improvement in Figure 6.4 stems from insufficient morphosyntactic tagging accuracy. The second and more interesting question, is whether the accuracy for morphosyntactic tagging itself is affected by the ordering of the core clausal elements and lexical relationships between words in a clause. To answer those questions I analyse the models’ performance on the tagging task for: (i) the original, unaltered development splits of the treebanks the models were trained on, (ii) the development splits of my counterfactual treebanks from Section 4.2 (reordered treebanks) and (iii) the development splits of my rotated counterfactual treebanks from

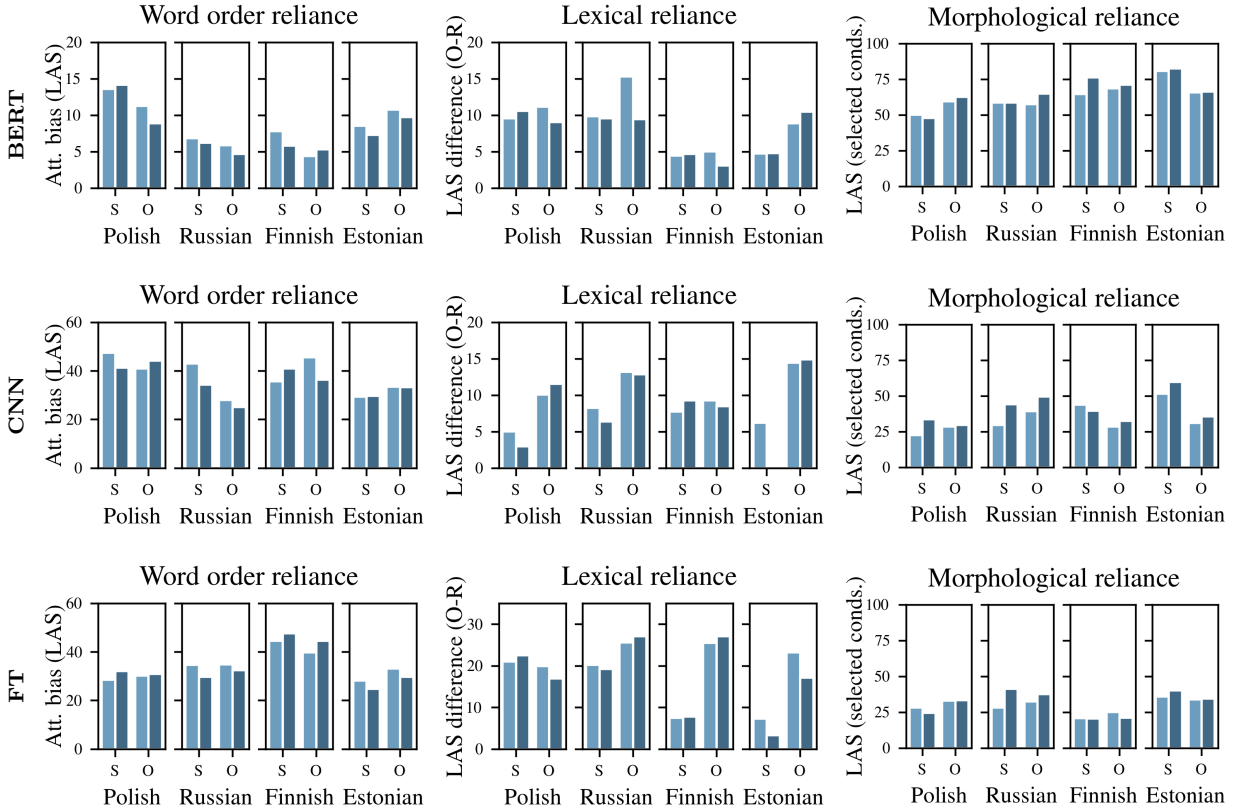


**Figure 6.4:** Reliance on different signals for the models trained with a multi-task objective (dark blue) and the models trained only with the dependency parsing objective (pale blue) for subject (S) and object (O) relations.

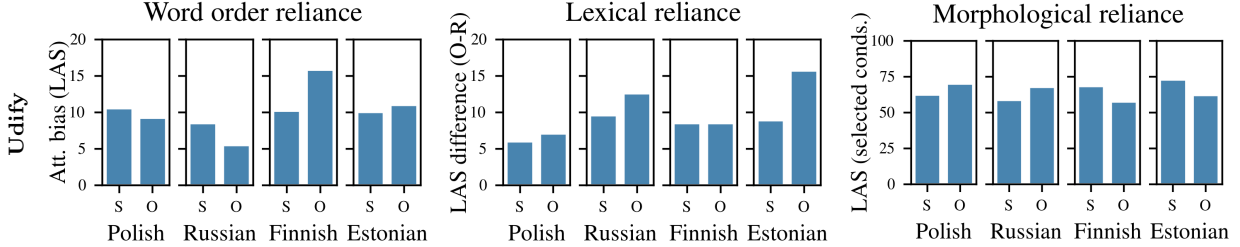


**Figure 6.5:** Reliance on different signals for the multilingual, multi-task UDify model for subject (S) and object (O).

Section 5.4 (reordered treebanks in which core argument lexemes are swapped). I present these results in Table 6.3.

Table 6.3 reveals that the tagging accuracy on the original, unaltered data is high for all considered models. BERT-based parsers emerge as the best performing, with accuracies close to 100% for all four languages. CNN and FASTTEXT-based models perform worse, but their accuracies surpass 90% in most conditions. Out of the four languages, I note the highest CNN/FASTTEXT results for Russian, with the accuracy of  $\sim 96$ . Since Russian is associated with the largest treebank – the SynTagRus treebank is almost three times

|                                                                                | BERT |      |      |      | CNN  |      |      |      | FASTTEXT |      |      |      |
|--------------------------------------------------------------------------------|------|------|------|------|------|------|------|------|----------|------|------|------|
|                                                                                | P    | R    | F    | E    | P    | R    | F    | E    | P        | R    | F    | E    |
| Original development data (standard evaluation)                                |      |      |      |      |      |      |      |      |          |      |      |      |
| Orig                                                                           | 98.1 | 99.0 | 98.6 | 97.7 | 91.7 | 96.0 | 92.7 | 91.4 | 93.1     | 96.4 | 90.1 | 89.0 |
| Counterfactual treebanks (only word order varied)                              |      |      |      |      |      |      |      |      |          |      |      |      |
| svo                                                                            | 98.7 | 99.2 | 99.5 | 97.4 | 88.7 | 93.8 | 90.2 | 87.5 | 91.3     | 97.1 | 88.2 | 87.1 |
| ovs                                                                            | 95.6 | 97.8 | 99.1 | 96.5 | 74.4 | 82.6 | 84.9 | 82.3 | 76.5     | 84.3 | 79.6 | 81.4 |
| sov                                                                            | 96.5 | 98.3 | 98.7 | 95.5 | 81.2 | 85.9 | 84.3 | 80.6 | 83.8     | 88.8 | 83.4 | 81.1 |
| osv                                                                            | 95.9 | 98.1 | 99.1 | 95.8 | 77.2 | 85.5 | 87.8 | 81.3 | 77.9     | 89.2 | 83.4 | 80.4 |
| vso                                                                            | 91.3 | 95.4 | 98.5 | 96.0 | 77.6 | 78.9 | 81.1 | 83.0 | 75.3     | 76.1 | 75.3 | 74.0 |
| vos                                                                            | 94.6 | 98.1 | 97.8 | 95.8 | 77.5 | 86.5 | 83.1 | 82.7 | 77.4     | 79.4 | 76.6 | 74.7 |
| Rotated counterfactual treebanks (word order varied + core argument rotations) |      |      |      |      |      |      |      |      |          |      |      |      |
| svo                                                                            | 92.0 | 91.7 | 98.8 | 96.4 | 83.5 | 87.7 | 88.8 | 86.8 | 76.4     | 85.6 | 84.0 | 82.1 |
| ovs                                                                            | 84.2 | 87.5 | 97.5 | 94.7 | 68.7 | 76.3 | 84.3 | 80.9 | 64.7     | 71.2 | 74.9 | 75.3 |
| sov                                                                            | 83.0 | 82.8 | 97.4 | 94.1 | 76.5 | 77.9 | 84.0 | 81.0 | 70.9     | 77.6 | 79.9 | 76.7 |
| osv                                                                            | 86.0 | 88.1 | 97.4 | 94.5 | 71.4 | 79.1 | 85.9 | 81.0 | 66.2     | 76.0 | 80.1 | 76.1 |
| vso                                                                            | 76.7 | 78.2 | 97.2 | 94.5 | 71.5 | 71.7 | 80.5 | 82.1 | 63.3     | 66.9 | 70.2 | 68.1 |
| vos                                                                            | 83.6 | 87.3 | 97.2 | 93.5 | 72.0 | 79.6 | 81.9 | 81.6 | 64.0     | 68.9 | 69.9 | 67.7 |

**Table 6.3:** The word order and lexicosemantic biases of the tagging component of the multi-task parsers. Reported values are tagging accuracies. P, R, F, E stands for Polish, Russian, Finnish and Estonian, respectively.

larger than the second largest, Estonian EDT<sup>5</sup> – a likely reason for the lower tagging performance of the non-BERT parsers is data sparsity.

The high performance on the unaltered treebanks, however, is not invariant to the changes in the S-V-O ordering and lexeme changes. The tagging accuracy drops acutely for most models when those variables are altered, with the largest drops observed for CNN and FASTTEXT. This means that morphosyntactic tagging suffers from the same over-reliance on those signals, as the dependency parsing task. This is an extremely interesting result, but its deeper investigation falls beyond the scope of this thesis. Future work should test whether this emerges only in a setting in which the tagging task is coupled with the parsing task or whether the neural morphosyntactic taggers (and possibly models for many other tasks based on the same neural architectures) would be prone to arriving to the same ill-suited generalisations, even when trained with the tagging objective alone.

Based on the results Table 6.3, it is possible that the incorporation of the tagging objective does not increase the CNN and FASTTEXT-based parsers’ reliance on morphology because the models are much less capable of retrieving the relevant morphosyntactic information in clauses where morphology is the only signal. Here, an interesting direction for future research would be to replace the context-dependent morphosyntactic tagger with one that is context-independent; i.e., one which predicts the tag based on the input embedding alone, pre-contextualisation. This would prevent the tagging component from relying on word order and lexical cues and could lead to better tagging results in challenging settings and perhaps increased reliance on morphology in the parsing task. On the other

<sup>5</sup>The SynTagRus treebank consists of 87,336 sentences, while EDT of 30,973 sentences.

hand, for BERT-based models, I notice little increase in their reliance on morphology in Figure 6.4 even for the languages for which the models achieve very high tagging accuracy throughout the challenging conditions – Finnish and Estonian. In the light of findings from related research (see Section 3.4) which provide evidence for transformer-models encoding very high quantities of morphosyntactic information, this suggests that the models’ underreliance on morphology stems from their ‘reluctance’ to interpret this information as a key indicator of meaning (the missed connection hypothesis) and that encouraging the models to pay closer attention to word forms through the additional objective is insufficient to construct this missing link. Consequently, it is unlikely that the CNN and FASTTEXT-based parsers would form a causal link between inflectional markers and core sentence meaning, even if they achieved a better tagging accuracy.

## 6.3 Conclusion

In this chapter I investigated whether LSTM-based and transformer-based UD dependency parsers for CMLs can be pushed towards the correct morphology-centered generalisation by simple alterations to their training data (Section 6.1) and training objective (Section 6.2).

Through my experiments in Section 6.2, I have shown that incorporating a morphosyntactic tagging objective is insufficient to increase the models’ reliance on morphology. While such objective might push the models to extract relevant information, it fails to encourage them to use morphology as a signal to meaning. I have also shown that the new tagging component itself suffers from the SVO and lexical over-reliance, akin to that of the parsers. This suggests that the tendencies I reveal throughout this thesis are more wide-spread and could concern more NLP tasks.

Training data perturbations (Section 6.1), on the other hand, can successfully improve generalisation of neural parsers. Without the ability to rely on word order and lexical semantics as signals during training, the models start to pick up on morphological signals to a much greater extent, with minimal drops in their overall performance; BERT-based parsers in particular seem to perform equally well regardless of what types of linguistic signals they rely on. This finding is of key importance because it demonstrates that the models have a much greater capacity to rely on morphology than what they settle for when trained on original UD treebanks. Indeed, for the BERT-based parsers, simple training data perturbations are enough to bring the models close to 100% dependence on morphology over word order and lexical semantics. This proves that the parsers do extract the relevant inflectional information and *can* use it as a signal to core sentence meaning, but their architecture and/or pre-training predisposes them not to rely on that signal.

A natural question that arises here is: why? It is a multi-faceted *why* which encompasses multiple varied questions. For example: what aspects of the models cause them to often

disregard morphology as the primary signal to meaning, even when it is so prevalent in the training data (see Section 4.1.1)? Or: what aspects of a morphological system make it hard for the models to rely on morphology? These two questions are critical because answering them will lead to building better neural models with adequate signal sensitivities, suited to work well when trained on natural data in a given language. And while providing a complete answer to either of them is beyond the scope of this thesis, in the next chapter I discuss some preliminary experiments which aim to shed more light on the latter.



# THE EFFECTS OF INFORMATION ENCODING: EXPERIMENTS WITH SYNTHETIC VERSIONS OF POLISH

---

In the previous chapter I demonstrated that the state-of-the-art neural parsers for case-marking languages have a far greater capacity to rely on morphology as a signal to meaning than what they settle for. In many instances where the models *could* rely on morphology as the primary signal, they show a preference for relying on word order and lexical semantics instead. In this chapter I ask whether this resistance to rely on morphology as a signal to meaning can be tied to specific aspects of the morphological system of the modeled language. Providing an answer to this question is not only interesting from a theoretical standpoint (are some languages harder to model?), but it can also highlight directions towards building better, more morphologically-oriented models for case-marking languages.

To test whether aspects of a morphological system can ‘discourage’ or ‘encourage’ a model to rely on morphology as a signal, I design a set of experiments in which I train and evaluate the neural parsers from Section 4.3 (those employing the full architecture of [Dozat and Manning \(2017\)](#)) on *synthetic languages*, with different case marking mechanisms. I base all of my synthetic languages on Polish – each of them is a result of applying a different type of alteration to the Polish morphology. I ground my morphological alterations in the morphological typology of [Bickel and Nichols \(2007\)](#) (see Section 2.6.2), manipulating the following properties: flexivity, exponence, fusion and nominative-accusative syncretism. Note that my altering of the Polish morphology makes this chapter very different from the earlier chapters – there, I experimented with grammatical sentences in existing languages; here I experiment with synthetic, artificial languages.

The chapter is structured as follows. I begin with a short discussion of related work (Section 7.1), followed by a discussion of my synthetic languages (Section 7.2) and the

experimental design (Section 7.3). Then, in Section 7.4, I present and discuss the results of my experimentation and conclude the chapter in Section 7.5.

## 7.1 Related work

The NLP field has a long history of experimenting with artificial languages for various purposes. Some work aims to test/improve robustness of neural models, by experimenting on versions of English with broken word order (Clouatre et al., 2021, Gupta et al., 2021, Hessel and Schofield, 2021, Sankar et al., 2019, Sinha et al., 2021, White and Cotterell, 2021) or an adapted marking of inflectional features (Tan et al., 2020). Other work proposes to use artificial languages as development, pre-training data for transfer-learning (Ri and Tsuruoka, 2022, Wang and Eisner, 2016). Yet another purpose – one most aligned with that of this chapter – is to test learning capabilities of neural networks. Linzen et al. (2016b) provide an overview of this line of work in their related work section, listing, among others, research which tests whether the RNN-based models can learn context free languages (Gers and Schmidhuber, 2001, Rodriguez, 2001, Rodriguez et al., 1999) or a language which simulates English relative clauses (Elman, 1991, 1993).

Perhaps most similar to my work are the studies of Amrhein and Sennrich (2021) and Ravfogel et al. (2019). The first focuses on machine translation and aims to answer: “how hard is it for a [translation] model to learn a specific morphological phenomenon”. The authors investigate this question by experimenting on various synthetic versions of German in which one of the five phenomena – compounding, circumfixation, infixation, vowel harmony or reduplication – was inserted. The work of Ravfogel et al. (2019), on the other hand, falls within the area of LODNA (see Section 3.4) and seeks to answer: “how do typological properties such as word order and morphological case marking affect the ability of neural sequence models to acquire the syntax of a language”. They do so via experimentation on synthetic versions of English, constructed by manipulating word order, case systems and types of agreement. But while closely aligned with my work, Ravfogel et al.’s work differs from mine in three important respects. First, the authors use English as the base language and only experiment with rigid word order and fully flexible word order settings – i.e., they omit the harder case of a language that has a dominant order but allows for other orderings. Second, they only experiment with language variants with simple nonflexive concatenative morphology in which each morpheme has 3 characters which clearly stand out within the English morphophonology. Third, they only experiment with an agreement task (see Section 3.4), and hence do not test whether varying the means of encoding grammatical information has any effect on what the models *rely on as a signal to meaning*.

## 7.2 Synthetic languages

Throughout this chapter I experiment on five different **synthetic languages**, which differ from the Polish language exclusively in their means of encoding case information. Importantly, this means that the statistical word order and lexical tendencies remain the same as in the original language (see Section 4.1). Each synthetic language is constructed by first removing the existing case markings from the language – i.e., **decasing** the language – and then reinserting the case information, using a type of marking selected for that particular synthetic language. The decasing involves reinflecting all word forms, so that they encode the base, nominative case while keeping all other features intact (e.g., a feminine, singular, accusative form *zieloną* (*green*) is replaced by the feminine, singular, nominative *zielona*). This processing is akin to that applied to one of the baselines in Section 3.3.

My selection of case marking systems was guided by the morphological typology of Bickel and Nichols (2007) (see Section 2.6.2). Specifically, I hypothesise that the following aspects of the Polish morphology are particularly challenging to model (see Section 2.4.4 and Section 2.6.2 for more detailed definitions):

- High degree of case syncretism – nominative and accusative often share the same form.
- High degree of flexivity (prominent allomorphy) – morphemes often have multiple realisations.
- Relatively high degree of exponence – one morph often encodes multiple grammatical features.
- High degree of fusion – most morphs are phonologically bound to the stem and their combination with the stem at times results in phonological stem adjustments.

Accordingly, I propose five *simple* methods to reduce syncretism, flexivity, exponence and/or fusion, by altering how case is encoded within the language. Each of these five methods yields a different synthetic language; I describe them in more detail in the following subsections. Importantly, the five synthetic languages I propose here only scratch the surface of what explorations are possible in this space and results discussed here are only preliminary. However, despite their simplicity and shortcomings the following alterations are enough to reveal very interesting tendencies in models’ generalisation patterns, as I demonstrate in Section 7.4.

## I. Case tag concatenations

The first alteration introduces case marking which resembles the ‘unambiguous case system’ introduced by Ravfogel et al. (2019). After decasing, I re-introduce the case marking by concatenating the case tag to the form, as demonstrated below. I use UniMorph<sup>1</sup> (Kirov et al., 2016, McCarthy et al., 2020) for the case tags.

- (22) a. Zielona<sub>a</sub>                      łąkę                      → Zielona<sub>acc</sub> łąka<sub>acc</sub>  
Green.F.SG.<sub>ACC</sub> meadow(F).SG.<sub>ACC</sub>
- b. Głodne                              dziecko                      → Głodne<sub>acc</sub> dziecko<sub>acc</sub>  
Hungry.F.SG.<sub>NOM/ACC</sub> child(F).SG.<sub>NOM/ACC</sub>

This alteration makes the case marking much simpler and very explicit – even more so because each of the tags is quite long, with 3 characters. The simplicity comes from (i) the reduced exponence for adjectives (previously one morph encoded three categories – gender, number and case; now these three are encoded in two morphs and, importantly, case morphs encode exclusively case information) and (ii) eliminated case syncretism and allomorphy. In addition, the new case system has no case morph ambiguity – each morph is tied to only one morpheme – and the marking of a particular case on nouns is the same as for adjectives.

## II. Case tag concatenations; first character only

The second alteration mirrors the first, but this time, instead of concatenating the entire tag, I only concatenate the *first* character of the tag:

- (23) a. Zielona<sub>a</sub>                      łąkę                      → Zielona<sub>a</sub> łąka<sub>a</sub>  
Green.F.SG.<sub>ACC</sub> meadow(F).SG.<sub>ACC</sub>
- b. Głodne                              dziecko                      → Głodne<sub>a</sub> dziecko<sub>a</sub>  
Hungry.N.SG.<sub>NOM/ACC</sub> child(N).SG.<sub>NOM/ACC</sub>

This has the same range of effects as the above perturbation, but makes the marking a bit less explicit – it may be more challenging to identify one character as a meaningful unit within a language.

## III. Case tag concatenations; first character only + syncretism

In this alteration, as in the one above, I decase the form and concatenate the first character of the case tag, but this time I introduce the ambiguity between the nominative and the accusative case for nouns and adjectives. I model this nominative-accusative case syncretism after the real syncretism patterns in Polish, marking the accusative with an ‘n’

---

<sup>1</sup><https://unimorph.github.io/>

character instead of an ‘a’ whenever an accusative form is the same as the nominative one in Polish. I have discussed the Polish case syncretism patterns for nouns in Section 2.4.4. The patterns for adjectives closely mirror those for nouns, except the singular feminine adjectives which have a unique accusative form (unlike the nouns) (Tokarski, 2001, p. 139).

- (24) a. Zielona<sub>a</sub>      łąk<sub>e</sub>      → Zielona<sub>a</sub> łąka<sub>a</sub>  
 Green.F.SG.ACC meadow(F).SG.ACC
- b. Głodne      dzieck<sub>o</sub>      → Głodne<sub>n</sub> dzieck<sub>o</sub>  
 Hungry.N.SG.NOM/ACC child(N).SG.NOM/ACC

#### IV. Case tag concatenations; first character only + (simple) allomorphy

This alteration resembles the third one, but this time I introduce allomorphy, instead of case syncretism. The allomorphy introduced here is very simple. I mark each case with two different morphs – one for singular nouns and one for plural nouns – keeping each morph unambiguous. Note that keeping each morph unambiguous with as many as 14 different settings (7 cases, 2 numbers), when each morph is represented by a single character is rather unnatural. Nevertheless, I experiment with this setting to test whether having more than one form per each case affects the models’ performance.

- (25) a. Zielona<sub>a</sub>      łąk<sub>e</sub>      → Zielona<sub>a</sub> łąka<sub>a</sub>  
 Green.F.SG.A meadow(F).SG.A
- b. Zielone      łąki      → Zielona<sub>b</sub> łąka<sub>b</sub>  
 Green.F.PL.NOM/ACC meadow(F).PL.NOM/ACC

#### V. Case encoded in a preposition

Since “anything you can do with cases you can also do with adpositions” (Zwicky, 1992, p. 370), in the fifth alteration, after decasing all forms I insert the case tag as a *preposition*. This reduces the degree of fusion within the language, making it more analytic, while also reducing the exponence for adjectives and getting rid of syncretism and allomorphy (as in synthetic language I).

- (26) a. Zielona<sub>a</sub>      łąk<sub>e</sub>      → Acc zielona łąka  
 Green.F.SG.ACC meadow(F).SG.ACC
- b. Głodne      dzieck<sub>o</sub>      → Acc głodne 2  
 dzieck<sub>o</sub>  
 Hungry.N.SG.NOM/ACC child(N).SG.NOM/ACC

## 7.3 Experimental design

**Models** As in the earlier chapters, I experiment on the models from Section 4.3, which I train on the original treebanks for the synthetic languages; i.e., treebanks otherwise unaltered, apart from the changes to the case marking (see the following paragraph for more details on the construction of such treebanks). I evaluate such models within the framework proposed across the earlier chapters, using a different measure to quantify the reliance on word order, lexicosemantic and morphological signals (see Section 6.1.1.2 for more details). As in the earlier chapters, I train four different versions for each model – each version trained with a different random seed – and report the average results.

In addition to the models trained on the synthetic languages from Section 7.2, I also experiment with three **control models**, which are trained (and tested) on the decased Polish treebank and are fed *gold* case information as an additional feature, encoded in a 100-dimensional embedding concatenated to the BERT/CNN/FASTTEXT word representation.

**Synthetic language treebanks** I construct a UD treebank for each synthetic language, using the Polish PDB treebank (Wróblewska, 2018) as a base. As the first step, I remove the case information from all POS that encode this information in Polish; i.e., pronouns, determiners, quantifiers, nouns and adjectives (including verbal nouns and verbal adjectives). To do this conversion I manually decase word forms with a matching POS, to ensure the gold quality of the data.<sup>2</sup> The second step is to reinsert the case information into the language. I do this differently for each synthetic language (see Section 7.2).

In addition to creating a version of the original UD Polish treebank, I also create: 6 counterfactual treebanks, with varied word orders (see Section 4.2), a treebank filtered down to contain transitive sentences with lexemes of core arguments rotated (see Section 5.3) and 6 counterfactual treebanks with varied word orders and with lexemes of core arguments rotated (see Section 5.4). While creating those treebanks I always apply the synthetic language alterations last. For example, to create the counterfactual treebanks for a synthetic language I first filter, simplify and reorder the sentences, as in Section 4.2, and then alter the means of encoding case information.

## 7.4 Results

I present the results for the CNN and BERT-based models trained and evaluated on each synthetic language in Figure 7.1. In the same figure I also present the results for these models processing unaltered Polish (*orig.*). I omit the results for the FASTTEXT-based

---

<sup>2</sup>I do this out-of-context so this does not guarantee 100% accuracy, but the errors should be very infrequent.

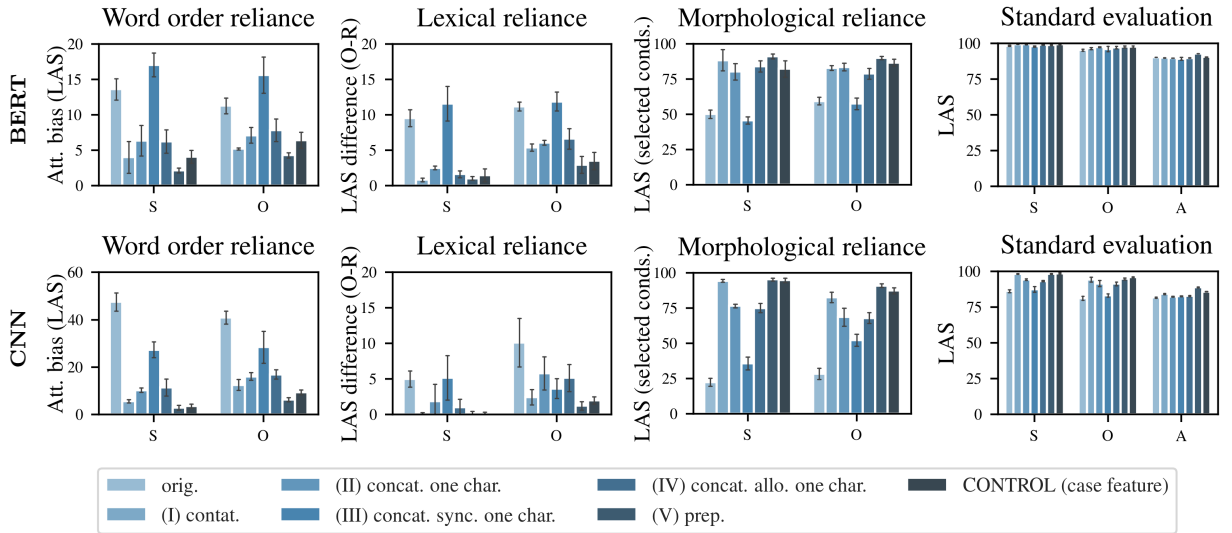


**Figure 7.1:** The generalisation strategies of models trained on synthetic Polish languages, with different means of encoding case information. S, O and A stand for *subject*, *object* and *all*. Grey bars indicate the standard deviation.

parsers from the figure, since these results lead to less interesting insights – the new concatenations of the form + case tag are missing from the embedding vocabulary and the model fails on all concatenative languages.<sup>3</sup>

Note that, in my experiments, the CNN-based parser is the only non pre-trained model – i.e., the only model that learns each synthetic language from scratch. Since BERT has been pre-trained on normal, unaltered Polish, when fine-tuned it has to re-learn the morphology. Further, its tokenisation might be sub-optimal for the new synthetic word forms. All this is likely to have consequences for its performance on the synthetic languages.

**Simple case tag concatenations (I and II)** One result that really stands out in Figure 7.1 is that concatenating the full case tag to the decased form (*concat.*) leads to a conspicuous increase in the reliance on morphology. This applies to both models, but is particularly prominent for the CNN-based parser – on subject relations, its reliance on morphology goes from < 25 LAS to 90+ LAS (!). The morphological reliance of CNN *concat.* even surpasses that of BERT *concat.*, but this could be due to the latter being pre-trained on unaltered Polish. Notably, for both models, the tendency to increase their reliance on morphology goes in hand with a steep decrease in their reliance on word order and lexical semantics, suggesting that the signals are in direct competition.

This result means that, when trained on a language with more explicit and unambiguous case marking, the models change their strategies and learn to rely on morphology over the alternative linguistic cues. This is an extremely interesting result, especially when

<sup>3</sup>The results for FASTTEXT can be found in Appendix G, Table G.11, where I report complete LAS tables for the rotated counterfactual treebanks for all models for the synthetic languages.

viewed in the context of the multi-task experiments from the previous chapter (Section 6.2). In those earlier results, models which achieved very high (90+%) morphological tagging accuracy (also across different word orders) still under-relied on morphology. Hence, the difference between the results for the models trained on unaltered Polish (*orig.*) and the *concat.* results in Figure 7.1 is likely to stem not from the improved retrieval of morphological features but the newly developed tendency to *rely* on those features. In other words, there seems to be something about the simplicity of the *concat.* marking that encourages the models to rely on it over the alternatives.

The reliance on morphology drops slightly when the full case tag is replaced by only a single character (*concat. one char.*), suggesting that the more prominent the morphs, the more likely they are to be picked up on and used as a signal to meaning. But despite this drop, the morphological reliance remains much higher than that of the models trained on unaltered Polish, suggesting that the morph distinguishability is only responsible for a relatively small part of the observed increase in *concat.*

**Introducing syncretism and allomorphy (III and IV)** A very interesting pattern emerges from introducing syncretism back into the equation. The generalisation patterns of the BERT-based models trained on the synthetic language with nominative-accusative syncretism that closely follows the original Polish patterns largely resemble those of models trained on unaltered Polish. I.e., for both BERT *orig.* and BERT *concat. sync. one char.* I observe similar reliance on word order and lexical semantics and similarly low morphological reliance. For the CNN-based models I notice a slight improvement in reliance on morphology over the CNN *orig.* but much smaller than that observed for synthetic languages with no case ambiguity.

In contrast, reinserting a simple version of allomorphy to the *concat. one char.* language does not seem to affect the models’ generalisation strategies. This suggests that in contrast to the many-to-one function-to-form mapping (syncretism, as introduced in *concat. sync. one char.*), the one-to-many mapping is not challenging to the models.

In the light of the above results, it appears that occasional ambiguity of case markings is the major factor pushing the *orig.* models (i.e., those evaluated in Chapters 4, 5, 6) away from relying on morphology. Given the scale of the drop caused by the reintroduction of syncretism, it is likely that every time the models encounter an object in an ambiguous nominative/accusative form they are ‘discouraged’ from relying on subword content as a signal to subject/objecthood, regardless of whether or not verbal agreement markings are ambiguous.

**Case marking vs prepositions and case feature (IV and V)** Comparing the results for the full tag concatenation (*concat.*) to the results for the language in which case information is encoded in a preposition (*prep.*) suggests that the models are only slightly

more likely to rely on case marking if it is more analytic. This is a somewhat surprising result – one would expect this difference to be more prominent. Reinserting the full case tag as a separate token right next to the original form makes the signal *very* explicit and, in humans, analytic grammatical structures are often considered easier to acquire by adult learners (Haspelmath and Michaelis, 2017). Perhaps this small difference is due to my use of the UniMorph tags as case representations. When concatenated to the decased forms, these tags do not necessarily fit within the Polish morphophonology, which might make them particularly easy to spot in the *concat.* language.

Interestingly, the morphological reliance in both *concat.* and *prep.* is higher than that for the *control* model which processes decased language and takes the *gold* case information as input, suggesting that the models ‘prefer’ to get this type of information directly in the text stream. Although I do not present these result in Figure 7.1, I have also experimented with feeding a gold case as an additional feature to models processing *unaltered Polish*. These results largely mirrored those for the model processing the decased language and can be found in Appendix G, Table G.11. A similar experiment was conducted by Vania and Lopez (2018), who have shown that providing a gold or a predicted case tag as an additional feature to the dependency parsers for Czech and Russian subtly improves parsing performance. Unlike Vania and Lopez (2018) I have not experimented with predicted tags. Perhaps, incorporating such tags could improve the models’ generalisations. However, this is an uncertain direction, given that in Section 6.2 the tagging task was subject to the same types of word order and lexical biases as the parsing task.

## 7.5 Conclusion

The most important take-away from this chapter is that neural parsers can rely on subword signals over the word order/lexical signals, with a close to 100% dependency on morphology, even without any encouragement in a form of training data/objective alterations (akin to those explored in Chapter 6). However, this happens only for languages with very simple case marking systems. The more complex the morphology becomes, the more the models are pushed towards relying on alternative linguistic signals. Syncretism, in particular, emerged as one of the most challenging properties. The explicitness of the morphemes also appears to be of relevance – models are more likely to pick up on morphs as a signal if they have 3 characters, compared to 1-character morphs. These results reveal that whether or not the models end up relying on a feature might be related to the difficulty of its extraction (even if the models are capable of extracting it).

All this was revealed via experimentation with very simple synthetic languages that only scratch the surface of what is possible within this space. This highlights this type of exploration as a very promising direction for analysis of neural models’ generalisation

strategies. In the future I hope to investigate other, more elaborate ways of manipulating the morphological system of Polish and other languages. One particularly interesting follow-up would be to account not only for case marking, but also for agreement marking. Given the drop I observed from re-introducing case syncretism in Section 7.4, I wonder whether these results would be any different if agreement was never ambiguous. Investigations into agreement could reveal whether the models analyse the ambiguity of case marking in isolation from the ambiguity of subject-verb agreement marking or whether they develop a concept of *genuine clausal ambiguity* which arises from analysing both types of ambiguity in tandem (i.e., a more abstract, higher-level feature). If the models cannot do the latter, then a better performance might be achieved by making them interpret both case marking and agreement marking together; for example via additional training objectives.

---

## CONCLUSION

---

In this thesis I asked whether the modern neural dependency parsers for case marking languages (CMLs) are morphologically competent; i.e., whether they can use information encoded in word-forms of these languages to identify the core sentence meaning. This question was primarily motivated by a wish to test whether neural state-of-the-art models for CMLs make the right predictions for the right reasons – a direction which receives much attention within the English NLP (Abdou et al., 2022, Clouatre et al., 2021, Gupta et al., 2021, Hessel and Schofield, 2021, McCoy et al., 2019, Papadimitriou et al., 2022, Poliak et al., 2018, Sankar et al., 2019, Sinha et al., 2021, among others), but is under-explored in morphologically-rich languages. But CMLs’ use of morphology to express core sentence meaning is also interesting from a theoretical perspective, since it constitutes a unique and important challenge to language modeling. Models for CMLs need to learn to rely on morphological signals whenever they are available, but revert to relying on word order/lexical semantics when morphology is ambiguous. The challenge here is to recognise the conditionally-determined importance of competing linguistic signals and studying it is relevant to NLP field, as a whole, since it is bound to apply in many other scenarios, spanning different languages and tasks.

To answer the above question, I proposed a new type of evaluation paradigm within which one can test a model’s relative reliance on three different types of plausible linguistic signals: morphology, word order and lexical semantics. This paradigm is centered on evaluating the models on carefully crafted, naturalistic datasets that allow one to answer what types of signals the models *use* when making their predictions. Through evaluation on such datasets I revealed that modern state-of-the-art neural dependency parsers for Polish, Russian, Finnish and Estonian often fail to recognise morphology as the primary indicator of syntax, learning to over-rely on word order and lexical semantics instead. This appears to hold regardless of the particulars of a model’s architecture or its training objective, although parsers based on pre-trained transformer models appear to rely on

morphological signals to a far larger extent than their counterparts based on FASTTEXT embeddings and non-pretrained CNNs.

Importantly, these results are not due to a shortage of morphological signals in the training data. Quite the opposite, they arise despite the models: (i) being constantly exposed to unambiguous morphological markings (see Section 4.1), and (ii) being capable of extracting much of this morphosyntactic information, as evidenced by research discussed in Section 3.4. What is more, as I revealed in Chapter 6, the models *can* rely on morphology to a greater extent if they are trained on data in which morphological signals are the only available linguistic signals. This applies to BERT-based parsers in particular, which can reach close to absolute dependence on morphology as a signal to subject/objecthood, while maintaining the same overall performance. All of the above suggests that the modern NLP architectures are in some way biased against relying on morphology. And although this tendency can be overridden with the right training data, with word order and lexical semantics controlled for, such methods only compensate for the inadequate relative input signal sensitivities (RISS, as defined in Section 3.1) prevalent in the models. A more long term solution is to directly target those ‘preferences’ by creating models with opposing biases which work well when trained on natural data.

The above findings have important theoretical and practical implications. Despite some correct tendencies, none of the models develops the linguistic generalisation pre-requisite to learning the grammar of a language, when they are trained on naturally occurring, unaltered data. This has direct consequences for the robustness of these models – while morphology remains consistent across different genres and domains, word order distribution and lexical relationships in the data vary. Consequently, relying on the latter two over morphology can lead to severe performance degradation in data that does not follow the word order distribution of the original training data or has different lexical characteristics. I also want to stress that the above consequences go beyond the specific languages I consider and, based on what I observed in Section 6.2, likely beyond the parsing task itself. As such, they point to a bigger issue within the NLP field which requires further study in future research. I hope that the evaluation paradigm proposed in this thesis, along with the data and code which accompany this submission, will prove helpful in such future explorations.

## 8.1 What next?

With this thesis I only take the first step within an important, under-explored area within NLP. While I have revealed a lot of important tendencies, this work leaves many questions unanswered. In the light of my findings, I identify two promising directions for future research: the first involves further explorations into the facets of neural models’ relative

sensitivities to different types of input signals (i.e., their RISS); the second involves building models that exhibit different sensitivities, better suited to neural modeling of CMLs. I briefly describe those directions in the following subsections.

### 8.1.1 Further explorations of RISS

Deeper analysis of the models' inherent signal preferences will give a better understanding of what should be changed within a model to make it more sensitive to morphological signals. Some unanswered, yet important questions within this stream include:

- How do the models' learning strategies change as they learn? Note that this could be investigated within the evaluation paradigm proposed in this thesis which enables straightforward tracking of changes in a model's generalisation strategies. Work in this scope could allow us to answer whether the models over-rely on non-morphological signals because they get stuck in local minima, which in turn could offer a more direct solutions to the underlying problem.
- Do the (i) amount of the training data and (ii) the paradigm coverage in the training data affect the strategies of the models?
- Do the models have a notion of morphosyntactic ambiguity of a clause and, if so, is it handled in a principled manner? Similar work has already been done for English (Aina and Linzen, 2021), but ambiguity is yet to be studied in neural NLP track for CMLs.
- Are the models' subject/object predictions affected by *attractors* in a form of non-argument nouns in structural cases (e.g., nominative/accusative)? This is an important question – while we should strive to build models which make use of morphology, they should also recognise the hierarchical structure of a sentence.
- Is the word-order and lexical bias already present in the pre-trained (masked) language models? For example, would they consider an ungrammatical Polish SVO clause more likely than a grammatical VOS clause?

Further, much more work could be carried out in the scope of the framework proposed in Chapter 7, to shed more light on how different aspects of a morphological system affect what input stream signals form the basis of neural models' solutions. This may include, for example, experimentation with stem changes, verbal agreement, and more elaborate patterns of allomorphy which introduce morph ambiguity.

**Beyond graph-based dependency parsing** Future research into RISS could also expand on the evaluated architectures (e.g., by including shift-reduce parsers) and explore more tasks, including those in which the models are not directly trained to predict the core relations, but identifying them is key to performing the task. Such tasks include question answering (Calijorne Soares and Parreiras, 2020, Rajpurkar et al., 2016, Reddy et al., 2019), natural language inference (Bowman et al., 2015, Chen et al., 2017) or, in a multi-modal setting, visual question answering (Antol et al., 2015) or image captioning (Bernardi et al., 2016, Hossain et al., 2019, Sharma et al., 2018). Finally, an interesting research direction would involve expanding my proposed framework to other types of languages, adapting the expectations for the relative ranking of word order, lexicosemantic and morphological signals to suit the inherent ranking of these cues in human language processing of a chosen language.

### 8.1.2 Towards more morphologically competent models

Another direction for future research is to build neural models better suited to CMLs by introducing changes to existing neural architectures and/or improving existing training paradigms. Here, one might explore how different tokenisation approaches affect models' generalisations (see Appendix D for an overview of different approaches to tokenisation in neural NLP). While my results reveal that the word-piece tokenisation is sufficient for the models to extract relevant morphosyntactic information (at least in Polish, Russian, Finnish and Estonian), perhaps different tokenisation approaches might make the models more prone to rely on morphology as a signal to meaning. Beyond tokenisation, it is also worth exploring the role of positional encodings within the transformer models, as well as the effects of a model's size. Finally, augmentation of neural approaches with symbolic methods could also prove to be a fruitful direction.

---

## REFERENCES

---

- Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. Word Order Does Matter and Shuffled Language Models Know It. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.476. URL <https://aclanthology.org/2022.acl-long.476>.
- Judit Ács, Dániel Lévai, and Andras Kornai. Evaluating Transferability of BERT Models on Uralic Languages. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 8–17, Syktyvkar, Russia (Online), 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.iwclul-1.2>.
- Gustavo Aguilar, Bryan McCann, Tong Niu, Nazneen Rajani, Nitish Shirish Keskar, and Tamar Solorio. Char2Subword: Extending the Subword Embedding Space Using Robust Character Compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1640–1651, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.141. URL <https://aclanthology.org/2021.findings-emnlp.141>.
- Laura Aina and Tal Linzen. The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty Through Generation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 42–57, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.4. URL <https://aclanthology.org/2021.blackboxnlp-1.4>.
- Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level Language Modeling with Deeper Self-Attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166, 2019. URL <https://ojs.aaai.org//index.php/AAAI/article/view/4182>.

- Andrei Alexandrescu and Katrin Kirchhoff. Factored Neural Language Models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4, New York City, USA, 2006. Association for Computational Linguistics. URL <https://aclanthology.org/N06-2001>.
- Chantal Amrhein and Rico Sennrich. How Suitable Are Subword Segmentation Strategies for Translating Non-Concatenative Morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.findings-emnlp.60>.
- Avery D. Andrews. The major functions of the noun phrase. In *Language Typology and Syntactic Description*, volume 1, page 132–223. Cambridge University Press, 2 edition, 2007. doi: 10.1017/CBO9780511619427.003. URL <https://www.cambridge.org/core/books/language-typology-and-syntactic-description/F4250D98B63306E13932BED28C651D4F>.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. URL [https://openaccess.thecvf.com/content\\_iccv\\_2015/papers/Antol\\_VQA\\_Visual\\_Question\\_ICCV\\_2015\\_paper.pdf](https://openaccess.thecvf.com/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf).
- Raimo Anttila. *An Introduction to Historical and Comparative Linguistics*. Macmillan New York, 1972.
- Peter Arkadiev, Marian Klammer, and Masini F Audring J. Morphological theory, language description and typology. In *The Oxford Handbook of Morphological Theory*, pages 435–454. Oxford University Press, 2018.
- Peter M Arkadiev. Morphology in typology: Historical retrospect, state of the art, and prospects. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2020.
- Mikel Artetxe and Holger Schwenk. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl\_a\_00288. URL <https://aclanthology.org/Q19-1038>.
- Duygu Ataman and Marcello Federico. Compositional Representation of Morphologically-Rich Input for Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

- 305–311, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2049. URL <https://aclanthology.org/P18-2049>.
- Oded Avraham and Yoav Goldberg. The Interplay of Semantics and Morphology in Word Embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 422–426, Valencia, Spain, April 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-2067>.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL <https://arxiv.org/abs/1607.06450>.
- Geoff Bacon and Terry Regier. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*, 2019. URL <https://arxiv.org/abs/1908.09892>.
- Matthew Baerman. Case syncretism. In *The Oxford Handbook of Case*. Oxford University Press, 2009.
- Matthew Baerman and Dunstan Brown. Case Syncretism. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <https://wals.info/chapter/28>.
- Matthew Baerman, Dunstan Brown, and Greville G. Corbett. Introduction. In *The Syntax-Morphology Interface: A Study of Syncretism*, Cambridge Studies in Linguistics, page 1–12. Cambridge University Press, 2005a. doi: 10.1017/CBO9780511486234.002.
- Matthew Baerman, Dunstan Brown, and Greville G. Corbett. Cross-linguistic typology of features. In *The Syntax-Morphology Interface: A Study of Syncretism*, Cambridge Studies in Linguistics, page 37–125. Cambridge University Press, 2005b. doi: 10.1017/CBO9780511486234.004.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.0473>.
- Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1041. URL <https://aclanthology.org/D15-1041>.

- Nikolay Banar, Walter Daelemans, and Mike Kestemont. Character-level transformer-based neural machine translation. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 149–156, 2020.
- Marco Baroni. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint arXiv:2106.08694*, 2021. URL <https://arxiv.org/abs/2106.08694>.
- Marco Baroni and Roberto Zamparelli. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA, 2010. Association for Computational Linguistics. URL <https://aclanthology.org/D10-1115>.
- Elizabeth Bates and Brian MacWhinney. Second-language acquisition from a functionalist perspective: Pragmatic, semantic, and perceptual strategies. *Annals of the New York Academy of Sciences*, 1981.
- Elizabeth Bates and Brian MacWhinney. Functionalist approaches to grammar. In *Language acquisition: The state of the art*. Cambridge University Press, 1982.
- Peter Battaglia, Jessica Blake Chandler Hamrick, Victor Bapst, Alvaro Sanchez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andy Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Jayne Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. URL <https://arxiv.org/pdf/1806.01261.pdf>.
- Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13, 2021. ISSN 0891-2017. doi: 10.1162/coli\_a\_00422. URL [https://doi.org/10.1162/coli\\_a\\_00422](https://doi.org/10.1162/coli_a_00422).
- Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. doi: 10.1162/tacl\_a\_00254. URL <https://aclanthology.org/Q19-1004>.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. On the Linguistic Representational Power of Neural Machine Translation Models. *Computational Linguistics*, 46(1):1–52, 2020. doi: 10.1162/coli\_a\_00367. URL <https://aclanthology.org/2020.cl-1.1>.

- Emily M. Bender. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers, 2013. ISBN 1627050116, 9781627050111.
- Yoshua Bengio, Paolo Frasconi, and Patrice Simard. The problem of learning long-term dependencies in recurrent networks. In *IEEE International Conference on Neural Networks*, pages 1183–1188 vol.3, 1993. doi: 10.1109/ICNN.1993.298725.
- Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994. doi: 10.1109/72.279181.
- Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.
- Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. Morphological Priors for Probabilistic Neural Word Embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 490–500, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1047. URL <https://aclanthology.org/D16-1047>.
- Balthasar Bickel and Johanna Nichols. Inflectional morphology. In *Language Typology and Syntactic Description*, volume 3, page 169–240. Cambridge University Press, 2 edition, 2007. doi: 10.1017/CBO9780511618437.003.
- Balthasar Bickel and Johanna Nichols. Fusion of Selected Inflectional Formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013a. URL <https://wals.info/chapter/20>.
- Balthasar Bickel and Johanna Nichols. Exponence of Selected Inflectional Formatives. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013b. URL <https://wals.info/chapter/21>.
- Balthasar Bickel and Johanna Nichols. Inflectional Synthesis of the Verb. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013c. URL <https://wals.info/chapter/22>.
- Balthasar Bickel and Johanna Nichols. Sampling Case and Tense Formatives. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013d. URL <https://wals.info/chapter/s5>.

- Robert Bielecki. On the Nature of the Accusative in Finnish. *Lingua Posnaniensis*, 51: 19–38, 2009. doi: 10.2478/v10122-009-0002-1.
- Jeff A. Bilmes and Katrin Kirchhoff. Factored Language Models and Generalized Parallel Backoff. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 4–6, 2003. URL <https://aclanthology.org/N03-2002>.
- Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier. Labeled Grammar Induction with Minimal Supervision. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–876, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2143. URL <https://aclanthology.org/P15-2143>.
- James P. Blevins. Word-based morphology. *Journal of Linguistics*, 42(3):531–573, 2006. doi: 10.1017/S0022226706004191.
- James P Blevins, Farrell Ackerman, and Malouf. Word and paradigm morphology. In J. Audring and F. Masini, editors, *Oxford Handbooks in Linguistics*, chapter 13. Oxford University Press, 2019.
- Douglas Bloom. Case syncretism and word order freezing in the Russian language. Master’s thesis, Stanford University, 1999.
- Kerstin Blume. A contrastive analysis of interaction verbs with dative complements. *Linguistics*, 36(2):253–280, 1998. URL <https://doi.org/10.1515/ling.1998.36.2.253>.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017a. doi: 10.1162/tacl\_a\_00051. URL <https://aclanthology.org/Q17-1010>.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017b. doi: 10.1162/tacl\_a\_00051. URL <https://aclanthology.org/Q17-1010>.
- Geert Booij. Against split morphology. In *Yearbook of morphology 1993*, pages 27–49. Springer, 1993.
- Geert Booij. *Inherent versus contextual inflection and the split morphology hypothesis*, pages 1–16. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-017-3716-6. doi: 10.1007/978-94-017-3716-6\_1. URL [https://doi.org/10.1007/978-94-017-3716-6\\_1](https://doi.org/10.1007/978-94-017-3716-6_1).

- Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500, San Francisco USA, 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317593. URL <https://dl.acm.org/doi/10.1145/3308560.3317593>.
- Ina Bornkessel-Schlesewsky and Matthias Schlewsky. The Role of Prominence Information in the Real-Time Comprehension of Transitive Constructions: A Cross-Linguistic Approach. *Language and Linguistics Compass*, 3(1):19–58, 2009. doi: <https://doi.org/10.1111/j.1749-818X.2008.00099.x>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2008.00099.x>.
- Kaj Bostrom and Greg Durrett. Byte Pair Encoding is Suboptimal for Language Model Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.414. URL <https://aclanthology.org/2020.findings-emnlp.414>.
- Jan A. Botha and Phil Blunsom. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1899–1907. JMLR.org, 2014. URL <http://proceedings.mlr.press/v32/botha14.html>.
- Gosse Bouma, Djamé Seddah, and Daniel Zeman. Overview of the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 151–161, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwpt-1.16. URL <https://aclanthology.org/2020.iwpt-1.16>.
- Gosse Bouma, Djamé Seddah, and Daniel Zeman. From Raw Text to Enhanced Universal Dependencies: The Parsing Shared Task at IWPT 2021. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 146–157, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwpt-1.15. URL <https://aclanthology.org/2021.iwpt-1.15>.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642,

- Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL <https://aclanthology.org/D15-1075>.
- James Bradbury and Richard Socher. MetaMind Neural Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 264–267, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2308. URL <https://aclanthology.org/W16-2308>.
- Dustan Brown. Morphological Typology. In *The Oxford Handbook of Linguistic Typology*, Oxford Handbooks in Linguistics. Oxford University Press, 2010.
- Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020. ISSN 1319-1578. doi: <https://doi.org/10.1016/j.jksuci.2018.08.005>. URL <https://www.sciencedirect.com/science/article/pii/S131915781830082X>.
- Kris Cao and Marek Rei. A Joint Model for Word Embedding and Word Morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1603. URL <https://aclanthology.org/W16-1603>.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.
- Angel Chan, Elena Lieven, and Michael Tomasello. Children’s Understanding of the Agent-Patient Relations in the Transitive Construction: Cross-Linguistic Comparisons Between Cantonese, German, and English. *Cognitive Linguistics*, 20(2), 2009. doi: 10.1515/cogl.2009.015.
- Branden Chan, Stefan Schweter, and Timo Möller. German’s Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.598. URL <https://aclanthology.org/2020.coling-main.598>.
- Rui P. Chaves and Stephanie N. Richter. Look at that! BERT can be easily distracted from paying attention to morphosyntax. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 28–38, Online, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.scil-1.3>.

- Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2005. URL <https://aclanthology.org/K18-2005>.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *Proceedings of the 2014 Annual Conference of International Speech Communication Association (INTERSPEECH)*, pages 2635–2639, 2014. doi: 10.21437/Interspeech.2014-564. URL [https://www.isca-speech.org/archive/pdfs/interspeech\\_2014/chelba14\\_interspeech.pdf](https://www.isca-speech.org/archive/pdfs/interspeech_2014/chelba14_interspeech.pdf).
- Huadong Chen, Shujian Huang, David Chiang, Xinyu Dai, and Jiajun Chen. Combining Character and Word Information in Neural Machine Translation Using a Multi-Level Attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1284–1293, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1116. URL <https://aclanthology.org/N18-1116>.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1152. URL <https://aclanthology.org/P17-1152>.
- Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. Revisiting Character-Based Neural Machine Translation with Capacity and Compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4295–4305, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1461. URL <https://aclanthology.org/D18-1461>.
- Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.
- Noam Chomsky. *Language and problems of knowledge: The Managua lectures*. MIT press, 1987.
- Grzegorz Chrupała and Afra Alishahi. Correlating Neural and Symbolic Representations of Language. In *Proceedings of the 57th Annual Meeting of the Association for Computa-*

- tional Linguistics*, pages 2952–2962, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1283. URL <https://aclanthology.org/P19-1283>.
- Yoeng-Jin Chu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14: 1396–1400, 1965.
- Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1160. URL <https://aclanthology.org/P16-1160>.
- Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical Multiscale Recurrent Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=S1di0sfgl>.
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00448. URL [https://doi.org/10.1162/tacl\\_a\\_00448](https://doi.org/10.1162/tacl_a_00448).
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT’s Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL <https://aclanthology.org/W19-4828>.
- Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. Demystifying Neural Language Models’ Insensitivity to Word-Order. *arXiv preprint arXiv:2107.13955*, 2021. URL <https://arxiv.org/abs/2107.13955>.
- Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Lambek Festschrift Linguistic Analysis*, pages 345–384, 2010.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12(null):2493–2537, 2011. ISSN 1532-4435.
- Bernard Comrie. *Language universals and linguistic typology*. The University of Chicago Press, 1981.

- Bernard Comrie. *Tense*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1985. doi: 10.1017/CBO9781139165815.
- Bernard Comrie. Alignment of Case Marking of Full Noun Phrases. In *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <https://wals.info/chapter/98>.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single  $\&\!#\ast$  vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL <https://aclanthology.org/P18-1198>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Ryan Cotterell and Hinrich Schütze. Morphological Word-Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1140. URL <https://aclanthology.org/N15-1140>.
- Ryan Cotterell and Hinrich Schütze. Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association for Computational Linguistics*, 6: 33–48, 2018. doi: 10.1162/tacl\_a\_00003. URL <https://aclanthology.org/Q18-1003>.
- Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. Labeled Morphological Segmentation with Semi-Markov Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1017. URL <https://aclanthology.org/K15-1017>.
- Ryan Cotterell, Hinrich Schütze, and Jason Eisner. Morphological Smoothing and Extrapolation of Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1156. URL <https://aclanthology.org/P16-1156>.

- Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00425. URL [https://doi.org/10.1162/tacl\\_a\\_00425](https://doi.org/10.1162/tacl_a_00425).
- Forrest Davis and Marten van Schijndel. Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1159–1171, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.93. URL <https://aclanthology.org/2021.acl-long.93>.
- Marie-Catherine de Marneffe and Christopher D. Manning. The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, 2008. Coling 2008 Organizing Committee. URL <https://aclanthology.org/W08-1301>.
- Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. *Technical report, Stanford University*, 2008. URL [https://downloads.cs.stanford.edu/nlp/software/dependencies\\_manual.pdf](https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf).
- Marie-Catherine de Marneffe and Joakim Nivre. Dependency Grammar. *Annual Review of Linguistics*, 5(1):197–218, 2019. doi: 10.1146/annurev-linguistics-011718-011842. URL <https://www.annualreviews.org/doi/epdf/10.1146/annurev-linguistics-011718-011842>.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021. ISSN 0891-2017. doi: 10.1162/coli\_a\_00402. URL [https://doi.org/10.1162/coli\\_a\\_00402](https://doi.org/10.1162/coli_a_00402).
- Peter de Swart. Case markedness. In *Case, valency and transitivity*, chapter 11. John Benjamins Publishing, 2006.
- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*, 2019. URL <https://arxiv.org/abs/1912.09582>.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL <http://www.jstor.org/stable/2984875>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the*

- 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland, 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6620>.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, pages 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL <https://doi.org/10.1145/3278721.3278729>.
- Robert M. W. Dixon. *A Grammar of Boumaa Fijian*. University of Chicago Press, Chicago, 1988.
- Cícero dos Santos and Victor Guimarães. Boosting Named Entity Recognition with Neural Character Embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3904. URL <https://aclanthology.org/W15-3904>.
- Cícero Nogueira dos Santos and Bianca Zadrozny. Learning Character-level Representations for Part-of-Speech Tagging. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1818–1826. JMLR.org, 2014. URL <http://proceedings.mlr.press/v32/santos14.html>.
- Allen Downey. There is still only one test. <http://alldowney.blogspot.com/2016/06/there-is-still-only-one-test.html>, 2016. Accessed: 05.05.2022.
- David Dowty. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619, 1991. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/415037>.

- Timothy Dozat and Christopher D. Manning. Deep Biaffine Attention for Neural Dependency Parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=Hk95PK9le>.
- Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, number 155, pages 53–66, 2018. URL <https://ep.liu.se/ecp/155/007/ecp18155007.pdf>.
- Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL <https://wals.info/>.
- Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. One Size Does Not Fit All: Comparing NMT Representations of Different Granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1504–1516, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1154. URL <https://aclanthology.org/N19-1154>.
- Marina Dyakonova. *A phase-based approach to Russian free word order*. Netherlands Graduate School of Linguistics, 2009. URL [https://www.lotpublications.nl/Documents/230\\_fulltext.pdf](https://www.lotpublications.nl/Documents/230_fulltext.pdf).
- Daniel Edmiston. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv preprint arXiv:2004.03032*, 2020. URL <https://arxiv.org/abs/2004.03032>.
- Jack Edmonds. Optimum Branchings. *Journal of Research of The National Bureau of Standards B*, 71(4):233–240, 1967. URL [https://nvlpubs.nist.gov/nistpubs/jres/71b/jresv71bn4p233\\_a1b.pdf](https://nvlpubs.nist.gov/nistpubs/jres/71b/jresv71bn4p233_a1b.pdf).
- Martin Ehala. The word order of Estonian: implications to universal language. *Journal of Universal Language*, 7(1):49–89, 2006.
- Martin Ehala. Linguistic strategies and markedness in Estonian morphology. *Language Typology and Universals*, 62(1-2):29–48, 2009.
- Max Eichler, Gözde Gül Şahin, and Iryna Gurevych. LINSPECTOR WEB: A Multilingual Probing Suite for Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*,

- pages 127–132, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3022. URL <https://aclanthology.org/D19-3022>.
- Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun’ichi Tsujii. CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.609. URL <https://aclanthology.org/2020.coling-main.609>.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl\_a\_00359. URL <https://aclanthology.org/2021.tacl-1.10>.
- M. Elbeze and A. M. Derouault. A morphological model for large vocabulary speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 577–580 vol.1, 1990.
- Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. doi: [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E). URL <https://www.sciencedirect.com/science/article/pii/036402139090002E>.
- Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2):195–225, 1991.
- Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.
- Mati Ereht. *Estonian language*. Estonian academy publishers, Tallinn, 2007.
- Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL <https://aclanthology.org/N15-1184>.
- Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. Morphological Inflection Generation Using Character Sequence to Sequence Learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California,

2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1077. URL <https://aclanthology.org/N16-1077>.
- Maryia Fedzechkina, Elissa L Newport, and T Florian Jaeger. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive science*, 41(2):416–446, 2017.
- J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32, 1957. URL <https://www.bibsonomy.org/bibtex/25e3d6c72cdd123a638f71886d78f3c1e/brightbyte>.
- Richard Futrell and Roger P. Levy. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 50–59, 2019. doi: 10.7275/jb34-9986. URL <https://aclanthology.org/W19-0106>.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1502134112. URL <https://www.pnas.org/content/112/33/10336>.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1004. URL <https://aclanthology.org/N19-1004>.
- Philip Gage. A New Algorithm for Data Compression. *C Users Journal*, 12(2):23–38, 1994. ISSN 0898-9788.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501. URL <https://aclanthology.org/W18-2501>.
- Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual Fairness in Text Classification through Robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’19, pages 219–226,

- New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL <https://doi.org/10.1145/3306618.3317950>.
- Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.265. URL <https://aclanthology.org/2020.acl-main.265>.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.10. URL <https://aclanthology.org/2020.acl-demos.10>.
- F.A. Gers and J. Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194 vol.3, 2000. doi: 10.1109/IJCNN.2000.861302.
- Felix A Gers and E Schmidhuber. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340, 2001.
- Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. Language Modeling for Morphologically Rich Languages: Character-Aware Modeling for Word-Level Prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465, 2018. doi: 10.1162/tac1\_a\_00032. URL <https://aclanthology.org/Q18-1032>.
- Petra Geutner. Using morphology towards better large-vocabulary speech recognition systems. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 445–448. IEEE, 1995.
- Edward Gibson. On the Adequacy of the Competition Model. *Language*, 68(4):812–830, 1992. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/416855>.
- Edward Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1): 1–76, 1998. ISSN 0010-0277. doi: [https://doi.org/10.1016/S0010-0277\(98\)00034-1](https://doi.org/10.1016/S0010-0277(98)00034-1). URL <https://www.sciencedirect.com/science/article/pii/S0010027798000341>.

- Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019. ISSN 1364-6613. doi: <https://doi.org/10.1016/j.tics.2019.02.003>. URL <https://www.sciencedirect.com/science/article/pii/S1364661319300580>.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5426. URL <https://aclanthology.org/W18-5426>.
- Goran Glavaš and Ivan Vulić. Is Supervised Syntactic Parsing Beneficial for Language Understanding Tasks? An Empirical Investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.eacl-main.270>.
- Yoav Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017.
- Yoav Goldberg. Assessing BERT’s syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019. URL <https://arxiv.org/abs/1901.05287>.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.
- Stefan Grünewald, Frederik Tobias Oertel, and Annemarie Friedrich. RobertNLP at the IWPT 2021 Shared Task: Simple Enhanced UD Parsing for 17 Languages. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 196–203, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwpt-1.21. URL <https://aclanthology.org/2021.iwpt-1.21>.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1108. URL <https://aclanthology.org/N18-1108>.
- Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. BERT & Family Eat Word Salad: Experiments with Text Understanding. *arXiv preprint arXiv:2101.03453*, 2021. URL <https://arxiv.org/abs/2101.03453>.
- Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gall  . Character-based NMT with Transformer. *arXiv preprint arXiv:1911.04997*, 2019. URL <http://arxiv.org/abs/1911.04997>.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL <https://aclanthology.org/N18-2017>.
- Michael Hahn and Marco Baroni. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484, 2019. doi: 10.1162/tacl\_a\_00283. URL <https://aclanthology.org/Q19-1033>.
- Rowan Hall Maudslay and Ryan Cotterell. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.11. URL <https://aclanthology.org/2021.naacl-main.11>.
- Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
- M. Haspelmath and A.D. Sims. *Understanding Morphology*. Understanding language series. Hodder Education, 2010. ISBN 9780340950012.
- Martin Haspelmath. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80, 2011. doi: 10.1515/flin.2011.002. URL <https://doi.org/10.1515/flin.2011.002>.
- Martin Haspelmath. The morph as a minimal linguistic form. *Morphology*, 30(2): 117–134, 2020. doi: 10.1007/s11525-020-09355-5. URL <https://doi.org/10.1007/s11525-020-09355-5>.

- Martin Haspelmath and Susanne Maria Michaelis. Analytic and synthetic. In *Language Variation-European Perspectives VI: Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), Leipzig 2015*. John Benjamins Publishing Company, 2017.
- Martin Haspelmath et al. Non-canonical marking of core arguments in European languages. *Typological Studies in Language*, 46:53–84, 2001.
- David Haussler. Quantifying inductive bias: AI learning algorithms and Valiant’s learning framework. *Artificial Intelligence*, 36(2):177–221, 1988. ISSN 0004-3702. doi: [https://doi.org/10.1016/0004-3702\(88\)90002-1](https://doi.org/10.1016/0004-3702(88)90002-1). URL <https://www.sciencedirect.com/science/article/pii/0004370288900021>.
- Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3):493–531, 2014. URL <https://link.springer.com/article/10.1007/s10579-013-9244-1>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.
- Johannes Heinecke. Hybrid Enhanced Universal Dependencies Parsing. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 174–180, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwpt-1.18. URL <https://aclanthology.org/2020.iwpt-1.18>.
- David P Helmbold and Philip M Long. On the inductive bias of dropout. *The Journal of Machine Learning Research*, 16(1):3403–3454, 2015.
- Jack Hessel and Alexandra Schofield. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.27. URL <https://aclanthology.org/2021.acl-short.27>.

- John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL <https://aclanthology.org/D19-1275>.
- John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL <https://aclanthology.org/N19-1419>.
- Katrin Hiietam. Hiietam, Katrin. 2004. Case marking in Estonian grammatical relations. *Leeds Working Papers in Linguistics and Phonetics*, 10, 2004.
- Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. *Diploma, Technische Universität München*, 91(1), 1991.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Charles F. Hockett. Problems of Morphemic Analysis. *Language*, 23(4):321–343, 1947. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/410295>.
- Paul J. Hopper and Sandra A. Thompson. Transitivity in Grammar and Discourse. *Language*, 56(2):251–299, 1980. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/413757>.
- MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*, 2019. URL <https://arxiv.org/abs/1911.12246>.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.158. URL <https://aclanthology.org/2020.acl-main.158>.

- Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL <https://aclanthology.org/2020.findings-emnlp.7>.
- Matthias Huck, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser. Producing Unseen Morphological Variants in Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 369–375, Valencia, Spain, 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-2059>.
- Paul Ibbotson and Michael Tomasello. Prototype constructions in early language acquisition. *Language and Cognition*, 1, 2009. doi: 10.1515/LANGCOG.2009.004.
- Osmo Ikola. *Nykysuomen Käsikirja [Handbook of modern Finnish]*. Weilin & Göös, Helsinki, 1971.
- R. Jakobson. *Word and Language*. Number v. 2. De Gruyter Mouton, 1971. ISBN 9789027917669. URL <https://books.google.co.uk/books?id=qcrX0JD3UGcC>.
- R.O. Jakobson. Contribution to the general theory of case: general meanings of the Russian cases. In *Roman Jakobson. Russian and Slavic grammar: Studies 1931-1981*, pages 59–103. Mouton de Gruyter, Berlin, 1936/84. translation of Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutung der russischen Kasus, originally in TCLP 6.
- L. Janda and U. Hill. Concepts of case and time in Slavic. *Glossos*, 2002.
- Bibi Janssen and Natalia Meir. Production, comprehension and repetition of accusative case by monolingual Russian and bilingual Russian-Dutch and Russian-Hebrew-speaking children. *Linguistic Approaches to Bilingualism*, 9(4-5):736–765, 2019.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL <https://aclanthology.org/P19-1356>.
- Mark Johnson. How the Statistical Revolution Changes (Computational) Linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11, Athens, Greece,

- March 2009. Association for Computational Linguistics. URL <https://aclanthology.org/W09-0103>.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl\_a\_00065. URL <https://aclanthology.org/Q17-1024>.
- Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. URL <https://arxiv.org/abs/1602.02410>.
- Daniel Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. draft). <https://web.stanford.edu/~jurafsky/slp3/>, 2017. Accessed: 2017-05-06.
- Olga Kagan. *The Semantics of Case*. Cambridge University Press, 2020.
- Elsi Kaiser. The discourse functions and syntax of OSV word order in Finnish. In *36th Regional Meeting of Chicago Linguistic Society*, pages 179–194, 2000.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1062. URL <https://aclanthology.org/P14-1062>.
- Jenna Kanerva, Filip Ginter, and Sampo Pyysalo. Turku Enhanced Parser Pipeline: From Raw Text to Enhanced Graphs in the IWPT 2020 Shared Task. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 162–173, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwpt-1.17. URL <https://aclanthology.org/2020.iwpt-1.17>.
- Katharina Kann, Ryan Cotterell, and Hinrich Schütze. Neural Morphological Analysis: Encoding-Decoding Canonical Segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1097. URL <https://aclanthology.org/D16-1097>.

- F. Karlsson and A. Chesterman. *Finnish: An Essential Grammar*. Essential grammar. Routledge, 2008. ISBN 9780415439145. URL <https://books.google.co.uk/books?id=ftM1bN5TebkC>.
- Lauri Karttunen. Numbers and Finnish numerals. *SKY Journal of Linguistics*, 19:407–421, 2006.
- Jungo Kasai and Robert Frank. Jabberwocky Parsing: Dependency Parsing with Lexical Noise. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 113–123, 2019. doi: 10.7275/h12q-k754. URL <https://aclanthology.org/W19-0112>.
- Jerrold J. Katz and Jerry A. Fodor. The Structure of a Semantic Theory. *Language*, 39(2):170–210, 1963. ISSN 00978507, 15350665. URL <http://www.jstor.org/stable/411200>.
- Kazuya Kawakami, Chris Dyer, and Phil Blunsom. Learning to Create and Reuse Words in Open-Vocabulary Neural Language Modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1492–1502, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1137. URL <https://aclanthology.org/P17-1137>.
- Vera Kempe and Brian MacWhinney. The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition*, 20(4):543–587, 1998. doi: 10.1017/S0272263198004045.
- Eugene Kharitonov, Marco Baroni, and Dieuwke Hupkes. How BPE affects memorization in transformers. *arXiv preprint arXiv:2110.02782*, 2021. URL <https://arxiv.org/pdf/2110.02782.pdf>.
- Yoon Kim, Yacine Jernite, David A. Sontag, and Alexander M. Rush. Character-Aware Neural Language Models. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2741–2749. AAAI Press, 2016. URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12489>.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- Paul Kiparsky. The Projection of Arguments: Lexical and Compositional Factors. In *Projecting from the lexicon*. CSLI Publications, 1998.

- Paul Kiparsky. Structural case in Finnish. *Lingua*, 111(4-7):315–376, 2001.
- Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language*, 20(4):589–608, 2006.
- Svetlana Kiritchenko and Saif Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL <https://aclanthology.org/S18-2005>.
- Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3121–3126, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL <https://aclanthology.org/L16-1498>.
- Olesya Kisselev. Word Order Patterns in the Writing of Heritage and Second Language Learners of Russian. *Russian Language Journal*, 69:149–174, 2019. ISSN 00360252. URL <https://www.jstor.org/stable/26877881>.
- Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. Obtaining a Better Understanding of Distributional Models of German Derivational Morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63, London, UK, 2015. Association for Computational Linguistics. URL <https://aclanthology.org/W15-0108>.
- Seppo Kittilä. Case and the typology of transitivity. In *The Oxford Handbook of Case*. Oxford University Press, 2009.
- Stav Klein and Reut Tsarfaty. Getting the ##life out of living: How Adequate Are Word-Pieces for Modelling Complex Morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.24. URL <https://aclanthology.org/2020.sigmorphon-1.24>.
- Mateusz Klimaszewski and Alina Wróblewska. COMBO: A New Module for EUD Parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 158–166, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwpt-1.16. URL <https://aclanthology.org/2021.iwpt-1.16>.

- Arne Köhn. What’s in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1246. URL <https://aclanthology.org/D15-1246>.
- Dan Kondratyuk and Milan Straka. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1279. URL <https://aclanthology.org/D19-1279>.
- Päivi Koskinen. Subject-verb agreement and covert raising to subject in Finnish. *Toronto Working Papers in Linguistics*, 17, 1999.
- Grzegorz Krajewski and Elena Lieven. Competing cues in early syntactic development. In *Competing motivations in grammar and usage*, pages 163–177. Oxford Univ. Press, 2014.
- Katarzyna Krasnowska-Kieraś and Alina Wróblewska. Empirical Linguistic Study of Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1573. URL <https://aclanthology.org/P19-1573>.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. URL <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>.
- Paul R. Kroeger. *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511801679.
- Jiří Krámský. *The word as a linguistic unit*. De Gruyter Mouton, 1969. ISBN 9783110998351.
- Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL <https://aclanthology.org/P18-1007>.
- Alexander Kuhnle. Evaluating visually grounded language capabilities using microworlds. Technical Report UCAM-CL-TR-942, University of Cambridge, Computer Laboratory, 2020. URL <https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-942.pdf>.
- Yuri Kuratov and M. Arkhipov. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *arXiv preprint arXiv:1905.07213*, 2019. URL <https://arxiv.org/pdf/1905.07213.pdf>.
- Dariusz Kłeczek. PolBERT: Attacking Polish NLP Tasks with Transformers. In Maciej Ogrodniczuk and Łukasz Kobylński, editors, *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, 2020. URL <http://poleval.pl/files/poleval2020.pdf>.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1030. URL <https://aclanthology.org/N16-1030>.
- Karim Lasri, Alessandro Lenci, and Thierry Poibeau. Does BERT really agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.181. URL <https://aclanthology.org/2022.findings-acl.181>.
- Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. Compositionally Derived Representations of Morphologically Complex Words in Distributional Semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL <https://aclanthology.org/P13-1149>.
- Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989. URL <https://direct.mit.edu/neco/article-abstract/1/4/541/5515/Backpropagation-Applied-to-Handwritten-Zip-Code?redirectedFrom=PDF>.

- Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten Digit Recognition with a Back-propagation Network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990. URL <http://yann.lecun.com/exdb/publis/pdf/lecun-90c.pdf>.
- Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017a. doi: 10.1162/tacl\_a\_00067. URL <https://aclanthology.org/Q17-1026>.
- Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017b. doi: 10.1162/tacl\_a\_00067. URL <https://aclanthology.org/Q17-1026>.
- Aet Lees. The case of the object in early Estonian and Finnish texts. In *Proceedings of the 2004 Conference of the Australian Linguistics Society*, 2005.
- Joël Legrand and Ronan Collobert. Deep Neural Networks for Syntactic Parsing of Morphologically Rich Languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 573–578, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2093. URL <https://aclanthology.org/P16-2093>.
- Henna Pauliina Lemetyinen. *Investigating Finnish-speaking children’s noun morphology: How do young children acquire case marking?* PhD thesis, The University of Manchester, 2016.
- Jiahuan Li, Yutong Shen, Shujian Huang, Xinyu Dai, and Jiajun Chen. When is Char Better Than Subword: A Systematic Study of Segmentation Algorithms for Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 543–549, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.69. URL <https://aclanthology.org/2021.acl-short.69>.
- Jindřich Libovický and Alexander Fraser. Towards Reasonably-Sized Character-Level Transformer NMT by Finetuning Subword Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2572–2579, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.203. URL <https://aclanthology.org/2020.emnlp-main.203>.

- Jindrich Libovický, Helmut Schmid, and Alexander Fraser. Why don't people use character-level machine translation? *arXiv preprint arXiv:2110.08191*, 2021. URL <https://arxiv.org/abs/2110.08191>.
- Yongjie Lin, Yi Chern Tan, and Robert Frank. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4825. URL <https://aclanthology.org/W19-4825>.
- Liina Lindström. Partitive subjects in Estonian dialects. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 8(2):191–231, 2017.
- Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fernandez, Silvio Amir, Luís Marujo, and Tiago Luís. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1176. URL <https://aclanthology.org/D15-1176>.
- Tal Linzen. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.465. URL <https://aclanthology.org/2020.acl-main.465>.
- Tal Linzen and Marco Baroni. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212, 2021.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016a. doi: 10.1162/tacl\_a\_00115. URL <https://aclanthology.org/Q16-1037>.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016b. doi: 10.1162/tacl\_a\_00115. URL <https://aclanthology.org/Q16-1037>.
- Adam Liška, Germán Kruszewski, and Marco Baroni. Memorize or generalize? searching for a compositional rnn in a haystack. *arXiv preprint arXiv:1802.06467*, 2018. URL <https://arxiv.org/abs/1802.06467>.

- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1112. URL <https://aclanthology.org/N19-1112>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b. URL <https://arxiv.org/abs/1907.11692>.
- Debra L Long and Chantel S Prat. Individual differences in syntactic ambiguity resolution: Readers vary in their use of plausibility information. *Memory & Cognition*, 36(2): 375–391, 2008.
- Jiri Lukavsky and Filip Smolik. Word order and case inflection in Czech: On-line sentence comprehension in children and adults. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, number 31, 2009.
- Minh-Thang Luong and Christopher D. Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1100. URL <https://aclanthology.org/P16-1100>.
- Thang Luong, Richard Socher, and Christopher Manning. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL <https://aclanthology.org/W13-3512>.
- Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1002. URL <https://aclanthology.org/P15-1002>.
- Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. CharBERT: Character-aware Pre-trained Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online),

- December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.4. URL <https://aclanthology.org/2020.coling-main.4>.
- Brian MacWhinney. *The competition model: the input, the context, and the brain*, page 69–90. Cambridge Applied Linguistics. Cambridge University Press, 2001. doi: 10.1017/CBO9781139524780.005.
- Brian MacWhinney. New directions in the competition model. In *Beyond nature-nurture*, pages 131–160. Psychology Press, 2004.
- Brian MacWhinney and Elizabeth Bates. Functionalism and the competition model. *The crosslinguistic study of sentence processing*, pages 3–73, 1989.
- Brian MacWhinney, Elizabeth Bates, and Reinhold Kliegl. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23(2):127–150, 1984. ISSN 0022-5371. doi: [https://doi.org/10.1016/S0022-5371\(84\)90093-8](https://doi.org/10.1016/S0022-5371(84)90093-8). URL <https://www.sciencedirect.com/science/article/pii/S0022537184900938>.
- Ilya Makarchuk. SOV in Russian: A corpus study. 16th Conference on Typology and Grammar for Young Scholars in Institute for Linguistic Studies, 2019. URL <https://www.youngconfspb.com/application/files/4015/7340/3040/Abstracts-2019.pdf#page=176>.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993. URL <https://aclanthology.org/J93-2004>.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL <https://aclanthology.org/2020.acl-main.645>.
- Rebecca Marvin and Tal Linzen. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1151. URL <https://aclanthology.org/D18-1151>.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. The SIGMORPHON 2019 Shared

- Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/W19-4226. URL <https://aclanthology.org/W19-4226>.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-4226. URL <https://aclanthology.org/W19-4226>.
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangel'skiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France, may 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://www.aclweb.org/anthology/2020.lrec-1.483>.
- R. Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*, 2018. URL <https://arxiv.org/abs/1802.09091>.
- R. Thomas McCoy, Robert Frank, and Tal Linzen. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, 2020a. doi: 10.1162/tacl\_a\_00304. URL <https://aclanthology.org/2020.tacl-1.9>.
- R. Thomas McCoy, Junghyun Min, and Tal Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.21. URL <https://aclanthology.org/2020.blackboxnlp-1.21>.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence,

- Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL <https://aclanthology.org/P19-1334>.
- Ryan McDonald, Koby Crammer, and Fernando Pereira. Online Large-Margin Training of Dependency Parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219852. URL <https://aclanthology.org/P05-1012>.
- Ken McRae, Todd R. Ferretti, and Liane Amyote. Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes*, 12(2-3):137–176, 1997. doi: 10.1080/016909697386835. URL <https://doi.org/10.1080/016909697386835>.
- Igor Mel'čuk. Dependency in natural language. *Dependency in linguistic description*, 111: 1, 2009.
- Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What Happens To BERT Embeddings During Fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.4. URL <https://aclanthology.org/2020.blackboxnlp-1.4>.
- Matti Miestamo. A typological perspective on negation in Finnish dialects. *Nordic Journal of Linguistics*, 34(2):83–104, 2011. doi: 10.1017/S0332586511000126.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop*, 2013a.
- Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013b. URL <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.
- Merilin Miljan. *Grammatical case in Estonian*. PhD thesis, The University of Edinburgh, 2009.
- Merilin Miljan and Ronnie Cann. Rethinking case marking and case alternation in Estonian. *Nordic Journal of Linguistics*, 36(3):333–379, 2013. doi: 10.1017/S0332586513000309.

- Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.212. URL <https://aclanthology.org/2020.acl-main.212>.
- Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.
- Yasumasa Miyamoto and Kyunghyun Cho. Gated Word-Character Recurrent Language Model. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1992–1997, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1209. URL <https://aclanthology.org/D16-1209>.
- Mahdi Mohseni and Amirhossein Tebbifakhr. MorphoBERT: a Persian NER System with BERT and Morphological Analysis. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 23–30, Trento, Italy, 11–12 September 2019. Association for Computational Linguistics. URL <https://aclanthology.org/2019.nsurl-1.4>.
- G.D. Morley. *Syntax in Functional Grammar: An Introduction to Lexicogrammar in Systemic Linguistics*. Continuum, 2000. ISBN 9780826447340. URL <https://books.google.co.uk/books?id=UPlsAAAAIAAJ>.
- Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.490. URL <https://aclanthology.org/2020.acl-main.490>.
- Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särge. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291, 2014. URL <http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf>.

- Kadri Muischnek, Kaili Müürisep, and Dage Dage Särg. CG roots of UD treebank of Estonian web language. In *Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar-Methods, Tools and Applications, 30 September 2019, Turku, Finland*, number 168, pages 23–26. Linköping University Electronic Press, 2019.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. URL <https://aclanthology.org/C18-1198>.
- Ad Neeleman and Fred Weerman. Syntactic Effects of Morphological Case. In *The Oxford Handbook of Case*. Oxford University Press, 2009. URL <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199206476.001.0001/oxfordhb-9780199206476-e-19>.
- Adriaan Neeleman and Johannes Van de Koot. Word Order and Information Structure. In *The Oxford Handbook of Information Structure*. Oxford University Press, 2016. URL [https://discovery.ucl.ac.uk/id/eprint/1462462/1/Word%20Order%20and%20Information%20Structure%20\(repository\).pdf](https://discovery.ucl.ac.uk/id/eprint/1462462/1/Word%20Order%20and%20Information%20Structure%20(repository).pdf).
- Timothy Niven and Hung-Yu Kao. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL <https://aclanthology.org/P19-1459>.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). URL <https://aclanthology.org/L16-1262>.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.497>.

- Antoine Nzeyimana and Andre Niyongabo Rubungo. KinyaBERT: a Morphology-aware Kinyarwanda Language Model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.367. URL <https://aclanthology.org/2022.acl-long.367>.
- Christopher Olah. Conv nets: A modular perspective, Jul 2014. URL <https://colah.github.io/posts/2014-07-Conv-Nets-Modular/>.
- Christopher Olah. Understanding lstm networks, Aug 2015. URL <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- Carmel O’Shannessy. Competition between word order and case-marking in interpreting grammatical relations: A case study in multilingual acquisition. *Journal of Child Language*, 38:763–92, 2010. URL [https://www.researchgate.net/publication/47698908\\_Competition\\_between\\_word\\_order\\_and\\_case-marking\\_in\\_interpreting\\_grammatical\\_relations\\_A\\_case\\_study\\_in\\_multilingual\\_acquisition](https://www.researchgate.net/publication/47698908_Competition_between_word_order_and_case-marking_in_interpreting_grammatical_relations_A_case_study_in_multilingual_acquisition).
- Sebastian Padó, Aurélie Herbelot, Max Kisselew, and Jan Šnajder. Predictability of Distributional Semantics in Derivational Word Formation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-1122>.
- Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. When classifying grammatical role, BERT doesn’t care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.71. URL <https://aclanthology.org/2022.acl-short.71>.
- Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China, December 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.aacl-main.17>.
- Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of

- JMLR Workshop and Conference Proceedings*, pages 1310–1318. JMLR.org, 2013. URL <http://proceedings.mlr.press/v28/pascanu13.html>.
- Agnieszka Patejuk and Adam Przepiórkowski. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2018. URL <http://nlp.ipipan.waw.pl/Bib/pat:prz:18:book.pdf>.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL <https://aclanthology.org/N18-1202>.
- Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2089–2096, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2012/pdf/274\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf).
- Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.98. URL <https://aclanthology.org/2021.findings-acl.98>.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL <https://aclanthology.org/2020.acl-main.420>.
- Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking Word Embeddings using Subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1010. URL <https://aclanthology.org/D17-1010>.
- Frank Plank. Split morphology: How agglutination and flexion mix. *Linguistic Typology*, 3:279–340, 1999. doi: 10.1515/lity.1999.3.3.279.

- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL <https://aclanthology.org/S18-2023>.
- Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-Dropout: Simple and Effective Subword Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.170. URL <https://aclanthology.org/2020.acl-main.170>.
- Peng Qian, Xipeng Qiu, and Xuanjing Huang. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1140. URL <https://aclanthology.org/P16-1140>.
- Siyu Qiu, Qing Cui, Jiang Bian, Bin Gao, and Tie-Yan Liu. Co-learning of Word Representations and Morpheme Representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 141–150, Dublin, Ireland, 2014. Dublin City University and Association for Computational Linguistics. URL <https://aclanthology.org/C14-1015>.
- R. Quirk, Q. Randolph, R. Quirk, S. Greenbaum, Pearson Longman, G.N. Leech, and J. Svartvik. *A Grammar of Contemporary English*. Longman, 1972. ISBN 9780582524446. URL <https://books.google.co.uk/books?id=YEBBAQAAIAAJ>.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264>.

- Angela Ralli. The role of morphology in gender determination: evidence from modern greek. *Linguistics*, 40(3):519–551, 2002. doi: doi:10.1515/ling.2002.022. URL <https://doi.org/10.1515/ling.2002.022>.
- Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. Can LSTM Learn to Capture Agreement? The Case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5412. URL <https://aclanthology.org/W18-5412>.
- Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1356. URL <https://aclanthology.org/N19-1356>.
- Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.15. URL <https://aclanthology.org/2021.conll-1.15>.
- Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.eacl-main.295>.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl\_a\_00266. URL <https://aclanthology.org/Q19-1016>.
- Marek Rei, Gamal Crichton, and Sampo Pyysalo. Attending to Characters in Neural Sequence Labeling Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-1030>.
- Ryokan Ri and Yoshimasa Tsuruoka. Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models. In *Proceedings of the 60th Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7302–7315, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.504. URL <https://aclanthology.org/2022.acl-long.504>.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL <https://aclanthology.org/2020.acl-main.442>.
- Paul Rodriguez. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural computation*, 13(9):2093–2118, 2001.
- Paul Rodriguez, Janet Wiles, and Jeffrey L Elman. A recurrent neural network that learns to count. *Connection Science*, 11(1):5–40, 1999.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl\_a\_00349. URL <https://aclanthology.org/2020.tacl-1.54>.
- Patricia L. Rounds and Ruth Kanagy. Acquiring Linguistic Cues To Identity AGENT: Evidence from Children Learning Japanese as a Second Language. *Studies in Second Language Acquisition*, 20(4):509–542, 1998. ISSN 02722631, 14701545. URL <http://www.jstor.org/stable/44486899>.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL <https://aclanthology.org/2021.acl-long.243>.
- Tatyana Ruzsics, Olga Sozinova, Ximena Gutierrez-Vasques, and Tanja Samardzic. Interpretability for Morphological Inflection: from Character-level Predictions to Subword-level Rules. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3189–3201, Online, 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.eacl-main.278>.

- Gözde Gül Şahin and Mark Steedman. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1545. URL <https://aclanthology.org/D18-1545>.
- Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics*, 46(2): 335–385, 2020. doi: 10.1162/coli\_a\_00376. URL <https://aclanthology.org/2020.cl-2.4>.
- Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. Optimizing Segmentation Granularity for Neural Machine Translation. *Machine Translation*, 34(1):41–59, apr 2020. ISSN 0922-6567. doi: 10.1007/s10590-019-09243-8. URL <https://doi.org/10.1007/s10590-019-09243-8>.
- Kristina Sands and Lyle Campbell. Non-canonical subjects and objects in Finnish. In *Non-canonical Marking of Subjects and Objects*, Typological Studies in Language, pages 251–307. John Benjamins Publishing Company, 2001.
- Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1004. URL <https://aclanthology.org/P19-1004>.
- Friedrich Schlegel. *Über die Sprache und die Weisheit der Indier*. Verlag Mohr und Zimmer, 1808.
- Mike Schuster and Kaisuke Nakajima. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012. doi: 10.1109/ICASSP.2012.6289079.
- Rico Sennrich. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain, 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-2060>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin,

- Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL <https://aclanthology.org/P16-1162>.
- Pamela Shapiro and Kevin Duh. BPE and CharCNNs for Translation of Morphology: A Cross-Lingual Comparison and Analysis, 2018. URL <https://arxiv.org/pdf/1809.01301.pdf>.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- Tatiana Shavrina and Olga Shapovalova. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser. *Proceedings of the Corpora*, pages 78–84, 2017.
- Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, and Chris Dyer. The Role of Context in Neural Morphological Disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 181–191, Osaka, Japan, 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-1018>.
- Anna Siewierska. Syntactic Weight vs Information Structure and Word Order Variation in Polish. *Journal of Linguistics*, 29(2):233–265, 1993. ISSN 00222267, 14697742. URL <http://www.jstor.org/stable/4176234>.
- Anna Siewierska. *Variation in major constituent order; a global and a European perspective*, pages 475–552. De Gruyter Mouton, 1998. URL <https://doi.org/10.1515/9783110812206.475>.
- Anna Siewierska and Dik Bakker. Case and alternative strategies: Word order and agreement marking. In *The Oxford Handbook of Case*, Oxford Handbooks in Linguistics, chapter 19. Oxford University Press, 2008. URL <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199206476.001.0001/oxfordhb-9780199206476-e-20>.
- Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.230. URL <https://aclanthology.org/2021.emnlp-main.230>.

- Kaius Sinnemäki. A typological perspective on Differential Object Marking. *Linguistics*, 52(2):281–313, 2014. doi: doi:10.1515/ling-2013-0063. URL <https://doi.org/10.1515/ling-2013-0063>.
- Dan I. Slobin and Thomas G. Bever. Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3):229–265, 1982. ISSN 0010-0277. doi: [https://doi.org/10.1016/0010-0277\(82\)90033-6](https://doi.org/10.1016/0010-0277(82)90033-6). URL <https://www.sciencedirect.com/science/article/pii/0010027782900336>.
- Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden, 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-2006. URL <https://aclanthology.org/E14-2006>.
- Richard Socher and Christopher D. Manning. Deep Learning for NLP (without Magic). In *NAACL HLT 2013 Tutorial Abstracts*, pages 1–3, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL <https://aclanthology.org/N13-4001>.
- Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In James Bailey, Alistair Moffat, Charu C. Aggarwal, Maarten de Rijke, Ravi Kumar, Vanessa Murdock, Timos K. Sellis, and Jeffrey Xu Yu, editors, *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 553–562. ACM, 2015. doi: 10.1145/2806416.2806493. URL <https://doi.org/10.1145/2806416.2806493>.
- Alexey Sorokin, Tatiana Shavrina, Olga Lyashevskaya, Victor Bocharov, Svetlana Alexeeva, Kira Droganova, Alena Fenogenova, and Dmitry Granovsky. MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian. In *Proceedings of the International Conference Dialogue 2017*, 2017.
- Andrew Spencer. Gender as an Inflectional Category. *Journal of Linguistics*, 38(2):279–312, 2002. ISSN 00222267, 14697742. URL <http://www.jstor.org/stable/4176736>.
- Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. *Advances in neural information processing systems*, 28, 2015a.
- Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *Deep Learning Workshop (ICML 2015)*, 2015b.

- Malgorzata Staroń and Michèle Kail. Off-line sentence comprehension in Polish monolinguals: Developmental study based on the Competition Model. *Psychology of Language and Communication*, 8(1):63–70, 2004.
- Milan Straka. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2020. URL <https://aclanthology.org/K18-2020>.
- Gregory Stump. Inflection classes. In *The Oxford Handbook of Inflection*. Oxford University Press, 2015.
- H. Sulkala and M. Karjalainen. *Finnish*. Croom Helm descriptive grammars series. Routledge, 1992. ISBN 9780415026437. URL <https://books.google.co.uk/books?id=rlbnz0oY3LgC>.
- Urmas Sutrop and Jaagup Roomet. *Estonian language, 5th edition*. Eesti Instituut, 2015.
- Ilya Sutskever, James Martens, and Geoffrey E. Hinton. Generating Text with Recurrent Neural Networks. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1017–1024. Omnipress, 2011. URL [https://icml.cc/2011/papers/524\\_icmlpaper.pdf](https://icml.cc/2011/papers/524_icmlpaper.pdf).
- Patrizia Tabossi, Michael J Spivey-Knowlton, Ken McRae, and Michael K Tanenhaus. Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. *Attention and performance 15: Conscious and nonconscious information processing*, pages 589–615, 1994. URL <https://psycnet.apa.org/record/1994-98352-024>.
- Aarne Talman and Stergios Chatzikyriakidis. Testing the Generalization Power of Neural Network Models across NLI Benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4810. URL <https://aclanthology.org/W19-4810>.
- Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.455. URL <https://aclanthology.org/2020.emnlp-main.455>.

- Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. EstBERT: A Pretrained Language-Specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden. URL <https://aclanthology.org/2021.nodalida-main.2>.
- Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast Character Transformers via Gradient-based Subword Tokenization. In *International Conference on Learning Representations (ICLR)*, 2022. URL <https://openreview.net/forum?id=JtBRnr1OEFN>.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovered the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL <https://aclanthology.org/P19-1452>.
- Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019b. URL <https://openreview.net/forum?id=SJzSgnRcKX>.
- Lucien Tesnière. *Éléments de syntaxe structurale*. C. Klincksieck, Paris, 1959.
- Lucien Tesnière. *Elements of Structural Syntax*. John Benjamins Publishing Company, 2015. doi: 10.1075/z.185.
- Harry Tily. *The Role of Processing Complexity in Word Order Variation and Change*. PhD thesis, Stanford University, 2010. URL <https://ezp.lib.cam.ac.uk/login?url=https://www.proquest.com/dissertations-theses/role-processing-complexity-word-order-variation/docview/2449144929/se-2?accountid=9851>.
- Alan Timberlake. *The Nominative Object in Slavic, Baltic, and West Finnic*. Peter Lang International Academic Publishers, 1974.
- Alan Timberlake. The nominative object in Finnish. *Lingua*, 35(3):201–230, 1975. ISSN 0024-3841. doi: [https://doi.org/10.1016/0024-3841\(75\)90059-5](https://doi.org/10.1016/0024-3841(75)90059-5). URL <https://www.sciencedirect.com/science/article/pii/0024384175900595>.

- Alan Timberlake. *A Reference Grammar of Russian*. Cambridge University Press, 2004. ISBN 9780521772921. URL <https://books.google.co.uk/books?id=-VFNWqXxRoMC>.
- Jan Tokarski. *Fleksja polska, wydanie III*. Wydawnictwo Naukowe PWN, Warszawa, 2001.
- Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and Whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12, Los Angeles, CA, USA, 2010. Association for Computational Linguistics. URL <https://aclanthology.org/W10-1401>.
- Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics*, 39(1):15–22, 2013. doi: 10.1162/COLI\_a\_00133. URL <https://aclanthology.org/J13-1003>.
- Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.660. URL <https://aclanthology.org/2020.acl-main.660>.
- Mycal Tucker, Peng Qian, and Roger Levy. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.76. URL <https://aclanthology.org/2021.findings-acl.76>.
- Marten van Schijndel, Aaron Mueller, and Tal Linzen. Quantity doesn’t buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1592. URL <https://aclanthology.org/D19-1592>.
- Clara Vania and Adam Lopez. From Characters to Words to in Between: Do We Capture Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1184. URL <https://aclanthology.org/P17-1184>.

- Clara Vania and Adam Lopez. Explicitly modeling case improves neural dependency parsing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 356–358, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5447. URL <https://aclanthology.org/W18-5447>.
- Clara Vania, Andreas Grivas, and Adam Lopez. What do character-level models learn about morphology? The case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2583, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1278. URL <https://aclanthology.org/D18-1278>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- David Vilar and Marcello Federico. A Statistical Extension of Byte-Pair Encoding. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 263–275, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwslt-1.31. URL <https://aclanthology.org/2021.iwslt-1.31>.
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*, 2019. URL <https://arxiv.org/abs/1912.07076>.
- Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL <https://aclanthology.org/2020.emnlp-main.14>.
- Wilhelm von Humboldt. *Über die Verschiedenheit des menschlichen Sprachbaues: und ihren Einfluss auf die geistige Entwicklung des Menschengeschlechts*. Druckerei der Königlichen Akademie der Wissenschaften, 1836. URL <https://books.google.co.uk/books?id=dV4SAAAAIAAJ>.

- August Wilhelm von Schlegel. *Observations sur la langue et la littérature provençales*. Librairie grecque-latine-allemande, 1818.
- Atro Voutilainen. FinnTreeBank: Creating a research resource and service for language researchers with Constraint Grammar. *Constraint Grammar Applications*, page 41, 2011.
- Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 56–68, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1006. URL <https://aclanthology.org/P17-1006>.
- Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4115. URL <https://aclanthology.org/W17-4115>.
- Terence Wade and David Gillespie. *A Comprehensive Russian Grammar, 3rd Edition*. Wiley-Blackwell, 2011.
- Dingquan Wang and Jason Eisner. The Galactic Dependencies Treebanks: Getting More Data by Synthesizing New Languages. *Transactions of the Association for Computational Linguistics*, 4:491–505, 2016. doi: 10.1162/tacl\_a\_00113. URL <https://aclanthology.org/Q16-1035>.
- Alex Warstadt and Samuel R Bowman. Linguistic analysis of pretrained sentence encoders with acceptability judgments. *arXiv preprint arXiv:1901.03438*, 2019. URL <https://arxiv.org/abs/1901.03438>.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl\_a\_00290. URL <https://aclanthology.org/Q19-1040>.
- Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.16. URL <https://aclanthology.org/2020.emnlp-main.16>.

- Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency Effects on Syntactic Rule Learning in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.72. URL <https://aclanthology.org/2021.emnlp-main.72>.
- Richard M. Weist. The word order myth. *Journal of Child Language*, 10(1):97–106, 1983. doi: 10.1017/S030500090000516X.
- Jennifer C. White and Ryan Cotterell. Examining the Inductive Bias of Neural Language Models with Artificial Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.38. URL <https://aclanthology.org/2021.acl-long.38>.
- Edward WD Whittaker and Philip C Woodland. Particle-based language modelling. In *Sixth International Conference on Spoken Language Processing*, 2000.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Charagram: Embedding Words and Sentences via Character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1157. URL <https://aclanthology.org/D16-1157>.
- Ethan Wilcox, Peng Qian, Richard Futrell, Ryosuke Kohita, Roger Levy, and Miguel Ballesteros. Structural supervision improves few-shot learning and syntactic generalization in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4640–4652, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.375. URL <https://aclanthology.org/2020.emnlp-main.375>.
- Angelika Wittek and Michael Tomasello. German-speaking children’s productivity with syntactic constructions and case morphology: Local cues act locally. *First Language*, 25(1):103–125, 2005. doi: 10.1177/0142723705049120. URL <https://doi.org/10.1177/0142723705049120>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. URL <https://arxiv.org/abs/1910.03771>.

- Alina Wróblewska. Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6020. URL <https://aclanthology.org/W18-6020>.
- Shijie Wu and Ryan Cotterell. Exact Hard Monotonic Attention for Character-Level Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1148. URL <https://aclanthology.org/P19-1148>.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016. URL <https://arxiv.org/abs/1609.08144>.
- Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.190. URL <https://aclanthology.org/2022.acl-long.190>.
- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10: 291–306, 2022. doi: 10.1162/tacl\_a\_00461. URL <https://aclanthology.org/2022.tacl-1.17>.
- Zhilin Yang, Bhuwan Dhingra, Ye Yuan, Junjie Hu, William W. Cohen, and Ruslan Salakhutdinov. Words or Characters? Fine-grained Gating for Reading Comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=B1hdzd5lg>.
- Yuki Yoshimura and Brian MacWhinney. The use of pronominal case in English sentence interpretation. *Applied Psycholinguistics*, 31(4):619–633, 2010. doi: 10.1017/S0142716410000160.
- Fabio Massimo Zanzotto, Ioannis Korkontzelos, Francesca Fallucchi, and Suresh Manandhar. Estimating Linear Models for Compositional Distributional Semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*,

- pages 1263–1271, Beijing, China, 2010. Coling 2010 Organizing Committee. URL <https://aclanthology.org/C10-1142>.
- Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *Proceedings of ICLR*, 2015. URL <https://arxiv.org/pdf/1409.2329.pdf>.
- Daniel Zeman. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2008/pdf/66\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf).
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, oct 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2001. URL <https://aclanthology.org/K18-2001>.
- Kelly Zhang and Samuel Bowman. Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5448. URL <https://aclanthology.org/W18-5448>.
- Yi Zhu, Ivan Vulić, and Anna Korhonen. A Systematic Study of Leveraging Subword Information for Learning Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 912–932, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1097. URL <https://aclanthology.org/N19-1097>.
- Arnold M Zwicky. Jottings on adpositions, case inflections, government, and agreement. *The joy of grammar: A Festschrift in honor of James D. McCawley*, pages 369–383, 1992.



---

# UNIVERSAL DEPENDENCIES TAXONOMY

---

## A.1 Dependency relations

| Relation | Definition                                                                                                              |
|----------|-------------------------------------------------------------------------------------------------------------------------|
| acl      | adnominal clause; finite or non-finite clause modifying a nominal                                                       |
| advcl    | adverbial clause modifying a predicate or modifier word                                                                 |
| advmod   | adverb or adverbial phrase modifying a predicate or modifier word                                                       |
| amod     | adjectival modifier of a nominal                                                                                        |
| appos    | appositional modifier; a nominal used to define, name, or describe the referent of a preceding nominal                  |
| aux      | auxiliary; links a function word expressing tense, mood, aspect, voice, or evidentiality to a predicate                 |
| case     | links a case-marking element (preposition, postposition, or clitic) to a nominal                                        |
| cc       | links a coordinating conjunction to the following conjunct                                                              |
| ccomp    | clausal complement of a verb or adjective without an obligatorily controlled subject                                    |
| clf      | (numeral) classifier; a word reflecting a conceptual classification of nouns linked to a numeric modifier or determiner |
| compound | any kind of word-level compounding (noun compound, serial verb, phrasal verb)                                           |
| conj     | conjunct; links two elements which are conjoined                                                                        |
| cop      | copula; links a function word used to connect a subject and a nonverbal predicate to the nonverbal predicate            |
| csubj    | clausal syntactic subject of a predicate                                                                                |
| dep      | unspecified dependency, used when a more precise relation cannot be determined                                          |

|            |                                                                                                                                                    |
|------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| det        | determiner (article, demonstrative, etc.) in a nominal                                                                                             |
| discourse  | discourse element (interjection, filler, or non-adverbial discourse marker)                                                                        |
| dislocated | a peripheral (initial or final) nominal in a clause that does not fill a regular role of the predicate but has roles such as topic or afterthought |
| expl       | expletive; links a pronominal form in a core argument position but not assigned any semantic role to a predicate                                   |
| fixed      | fixed multiword expression; links elements of grammaticalised expressions that behave as function words or short adverbials                        |
| flat       | flat multiword expression; links elements of headless semi-fixed multiword expressions like names                                                  |
| goeswith   | links parts of a word that are separated but should go together according to standard orthography or linguistic wordhood                           |
| iobj       | indirect object; nominal core argument of a verb that is not its subject or (direct) object                                                        |
| list       | links elements of comparable items interpreted as a list                                                                                           |
| mark       | marker; links a function word marking a clause as subordinate to the predicate of the clause                                                       |
| nmod       | nominal modifier; a nominal modifying another nominal                                                                                              |
| nummod     | numeric modifier; numeral in a nominal                                                                                                             |
| nsubj      | nominal subject; nominal core argument which is the syntactic subject (or pivot) of a predicate                                                    |
| obj        | object; the core argument nominal which is the most basic core argument that is not the subject, typically the most directly affected participant  |
| obl        | oblique; a nominal functioning as a non-core (oblique) modifier of a predicate                                                                     |
| orphan     | links orphaned dependents of an elided predicate                                                                                                   |
| parataxis  | links constituents placed side by side with no explicit coordination or subordination                                                              |
| punct      | punctuation attached to the head of its clause or phrase                                                                                           |
| reparandum | repair of a (normally spoken language) disfluency                                                                                                  |
| root       | root of the sentence                                                                                                                               |
| vocative   | nominal directed to an addressee                                                                                                                   |
| xcomp      | clausal complement of a verb or adjective with an obligatorily controlled subject                                                                  |

**Table A.1:** List of universal dependency relation types. Table from [de Marneffe et al. \(2021\)](#).

## A.2 Part-of-speech

| Traditional<br>POS | UPOS  | Category                                                 |
|--------------------|-------|----------------------------------------------------------|
| noun               | noun  | common noun                                              |
|                    | propn | proper noun                                              |
| verb               | verb  | main verb                                                |
|                    | aux   | auxiliary verb or other tense, aspect, or mood particle  |
| adjective          | adj   | adjective                                                |
|                    | det   | determiner (including article)                           |
|                    | num   | numeral (cardinal)                                       |
| adverb             | adv   | adverb                                                   |
| pronoun            | pron  | pronoun                                                  |
| preposition        | adp   | adposition (preposition/postposition)                    |
| conjunction        | cconj | coordinating conjunction                                 |
|                    | sconj | subordinating conjunction                                |
| interjection       | intj  | interjection                                             |
| –                  | part  | particle (special single word markers in some languages) |
| –                  | x     | other (e.g., words in foreign language expressions)      |
| –                  | sym   | non-punctuation symbol (e.g., a hash (#) or emoji)       |
| –                  | punct | punctuation                                              |

**Table A.2:** List of universal part-of-speech. Table from [de Marneffe et al. \(2021\)](#).

### A.3 Morphological features

|                 | Feature   | Values                                                                                                                                              |
|-----------------|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| pronominal type | PronType  | Art Dem Emp Exc Ind Int Neg Prs Rcp Rel Tot                                                                                                         |
| numeral type    | NumType   | Card Dist Frac Mult Ord Range Sets                                                                                                                  |
| possessive      | Poss      | Yes                                                                                                                                                 |
| reflexive       | Reflex    | Yes                                                                                                                                                 |
| foreign word    | Foreign   | Yes                                                                                                                                                 |
| abbreviation    | Abbr      | Yes                                                                                                                                                 |
| wrong spelling  | Typo      | Yes                                                                                                                                                 |
| gender          | Gender    | Com Fem Masc Neut                                                                                                                                   |
| animacy         | Animacy   | Anim Hum Inan Nhum                                                                                                                                  |
| noun class      | NounClass | Bantu1-23 Wol1-12 ...                                                                                                                               |
| number          | Number    | Coll Count Dual Grpa Grpl Inv Pauc Plur Ptan Sing Tri                                                                                               |
| case            | Case      | Abs Acc Erg Nom<br>Abe Ben Cau Cmp Cns Com Dat Dis Equ Gen Ins Par<br>Tem Tra Voc<br>Abl Add Ade All Del Ela Ess Ill Ine Lat Loc Per Sub<br>Sup Ter |
| definiteness    | Definite  | Com Cons Def Ind Spec                                                                                                                               |
| comparison      | Degree    | Abs Cmp Equ Pos Sup                                                                                                                                 |
| verbal form     | VerbForm  | Conv Fin Gdv Ger Inf Part Sup Vnoun                                                                                                                 |
| mood            | Mood      | Adm Cnd Des Imp Ind Irr Jus Nec Opt Pot Prp Qot<br>Sub                                                                                              |
| tense           | Tense     | Fut Imp Nfut Past Pqp Pres                                                                                                                          |
| aspect          | Aspect    | Hab Imp Iter Perf Prog Prosp                                                                                                                        |
| voice           | Voice     | Act Antip Bfoc Cau Dir Inv Lfoc Mid Pass Rcp                                                                                                        |
| evidentiality   | Evident   | Fh Nfh                                                                                                                                              |
| polarity        | Polarity  | Neg Pos                                                                                                                                             |
| person          | Person    | 0 1 2 3 4                                                                                                                                           |
| politeness      | Polite    | Elev Form Humb Infm                                                                                                                                 |
| clusivity       | Clusivity | In Ex                                                                                                                                               |

**Table A.3:** List of universal morphological features. Table from [de Marneffe et al. \(2021\)](#).

---

## LEIPZIG GLOSSING ABBREVIATIONS

---

Most of the following abbreviations come from the Leipzig list (<https://www.eva.mpg.de/lingua/resources/glossing-rules.php>), but a few were added to cover all phenomena described in this thesis.

|       |                                                  |
|-------|--------------------------------------------------|
| 1     | first person                                     |
| 2     | second person                                    |
| 3     | third person                                     |
| A     | agent-like argument of canonical transitive verb |
| ABL   | ablative                                         |
| ABS   | absolutive                                       |
| ACC   | accusative                                       |
| ADJ   | adjective                                        |
| ADV   | adverb(ial)                                      |
| AGR   | agreement                                        |
| ALL   | allative                                         |
| ANTIP | antipassive                                      |
| APPL  | applicative                                      |
| ART   | article                                          |
| AUX   | auxiliary                                        |
| BEN   | benefactive                                      |
| CAUS  | causative                                        |
| CLF   | classifier                                       |
| COM   | comitative                                       |
| COMP  | complementizer                                   |
| COMPL | completive                                       |
| COND  | conditional                                      |
| F COP | copula                                           |

|       |                                                    |
|-------|----------------------------------------------------|
| CVB   | converb                                            |
| DAT   | dative                                             |
| DECL  | declarative                                        |
| DEF   | definite                                           |
| DEM   | demonstrative                                      |
| DET   | determiner                                         |
| DIST  | distal                                             |
| DISTR | distributive                                       |
| DU    | dual                                               |
| DUR   | durative                                           |
| ERG   | ergative                                           |
| EXCL  | exclusive                                          |
| F     | feminine                                           |
| FOC   | focus                                              |
| FUT   | future                                             |
| GEN   | genitive                                           |
| IMP   | imperative                                         |
| INCL  | inclusive                                          |
| IND   | indicative                                         |
| INDF  | indefinite                                         |
| INF   | infinitive                                         |
| INS   | instrumental                                       |
| INTR  | intransitive                                       |
| IPFV  | imperfective                                       |
| IRR   | irrealis                                           |
| LOC   | locative                                           |
| M     | masculine                                          |
| N     | neuter                                             |
| N-    | non- (e.g. NSG nonsingular, NPST nonpast)          |
| NEG   | negation, negative                                 |
| NMLZ  | nominalizer/nominalization                         |
| NOM   | nominative                                         |
| OBJ   | object                                             |
| OBL   | oblique                                            |
| P     | patient-like argument of canonical transitive verb |
| PASS  | passive                                            |
| PFV   | perfective                                         |
| PL    | plural                                             |

|      |                                                |
|------|------------------------------------------------|
| POSS | possessive                                     |
| PRED | predicative                                    |
| PRF  | perfect                                        |
| PRS  | present                                        |
| PROG | progressive                                    |
| PROH | prohibitive                                    |
| PROX | proximal/proximate                             |
| PST  | past                                           |
| PTCP | participle                                     |
| PURP | purposive                                      |
| Q    | question particle/marker                       |
| QUOT | quotative                                      |
| RECP | reciprocal                                     |
| REFL | reflexive                                      |
| REL  | relative                                       |
| RES  | resultative                                    |
| S    | single argument of canonical intransitive verb |
| SBJ  | subject                                        |
| SBJV | subjunctive                                    |
| SG   | singular                                       |
| TOP  | topic                                          |
| TR   | transitive                                     |
| VOC  | vocative                                       |

---

|      |           |
|------|-----------|
| INES | inessive  |
| PRT  | partitive |



## OVERVIEW OF THE RELEVANT NEURAL ARCHITECTURES

---

This section provides a brief overview of (i) the SKIP-GRAM model (Appendix C.1), (ii) the convolutional neural networks (CNNs) (Appendix C.2), (iii) the long short-term memory networks (LSTMs) (Appendix C.3), and (iv) the transformer models (Appendix C.4). The descriptions of neural architectures are based on Goldberg (2017), Olah (2015), and Olah (2014) for CNNs and LSTMs, and on Vaswani et al. (2017) for the transformer.

### C.1 The skip-gram model

SKIP-GRAM (Mikolov et al., 2013a,b) is a model of lexical semantics, based on the **distributional hypothesis** which states that words appearing in similar contexts also share similar meanings (Firth, 1957, Harris, 1954). Its main application is to construct pre-trained word embeddings which can be fed as input to neural models for specific NLP tasks. The model is trained to predict context-words of a given target-word, where the contexts are the immediate neighbours of the target and are retrieved using a window of an arbitrary size  $n$  (by capturing  $n$  words to the left and  $n$  words to the right of the target).

The SKIP-GRAM training data consists of word pairs  $(V_t, V_c)$ , where  $V$  is the vocabulary and  $t, c \in \{1, \dots, |V|\}$  are indices of a target-word and one of its context-words. The model is trained with a negative-sampling objective<sup>1</sup> to differentiate between the correct examples, retrieved from the corpus and the incorrect, randomly generated pairs. For each correct example the model draws  $m$  negative ones, with  $m$  being a hyper-parameter. These incorrect samples hold the same  $V_c$  as the original, while their  $V_t$  is drawn from an arbitrary noise distribution, typically the unigram distribution raised to the power

---

<sup>1</sup>The negative-sampling objective was proposed in Mikolov et al. (2013b) as a faster and more efficient alternative to the hierarchical softmax objective from Mikolov et al. (2013a).

0.75 (Mikolov et al., 2013b). Let  $D$  be the set of all correct pairs,  $D'$  denote a set of all negatively sampled  $|D| \times m$  pairs and  $P(C = 1|V_t, V_c)$  be the probability of  $(V_t, V_c)$  being a correct pair, originating from the corpus. The last is calculated using the sigmoid function:

$$P(C = 1|V_t, V_c) = \sigma(\mathbf{e}_t \cdot \mathbf{o}_c) = \frac{1}{1 + e^{-\mathbf{e}_t \cdot \mathbf{o}_c}} \quad (\text{C.1})$$

Here,  $\mathbf{e}_t \in \mathbb{R}^d$  is an **input-embedding** – the  $t$ -th row of a matrix  $\mathbf{E} \in \mathbb{R}^{|V| \times d}$ , holding representations of target-words, while  $\mathbf{o}_c \in \mathbb{R}^d$  is an **output-embedding** – the  $c$ -th row of a matrix  $\mathbf{O} \in \mathbb{R}^{|V| \times d}$ , holding context-word representations. Given this setting, the negative-sampling objective is defined as maximising:

$$\sum_{(V_t, V_c) \in D} \log \sigma(\mathbf{e}_t \cdot \mathbf{o}_c) + \sum_{(V_t, V_c) \in D'} \log \sigma(-\mathbf{e}_t \cdot \mathbf{o}_c) \quad (\text{C.2})$$

The model is trained using stochastic gradient ascent, with the learning rate changing throughout the training process and being proportional to the number of remaining examples.

## C.2 Convolutional neural networks

Convolutional neural networks (CNNs) (LeCun et al., 1989, 1990) “identify indicative local predictors in a large structure” (Goldberg, 2017) and are often integrated into larger models, as automatic extractors of features. CNNs first gained popularity in the vision community, proving very successful for object detection (Krizhevsky et al., 2012). They were first applied to text by Collobert et al. (2011).

The simplest CNN architecture for NLP consists of two layers: a convolution layer and a pooling layer. Both are described in the following paragraphs.

**Convolution** The convolution layer applies a non-linear **filter** function over instantiations of a  $k$ -sized window, sliding across the sentence representation. The filter can be thought of as a specialised local feature extractor, looking for the presence of a particular feature in each  $k$ -sized chunk. The sentence might be represented either as a vector, obtained through the concatenation of word/character embeddings, or as a matrix, resulting from stacking the embeddings on top of one another (see Figure C.1). The chosen representation affects the mathematical description of the filter, but the two approaches are equivalent and give the same results; see Goldberg (2017, p. 154). For the purpose of the following description, I assume the *sentence-as-a-matrix* view. Given such sentence representation, the filter takes  $k \times d$  matrices as inputs, where  $d$  is the dimensionality of

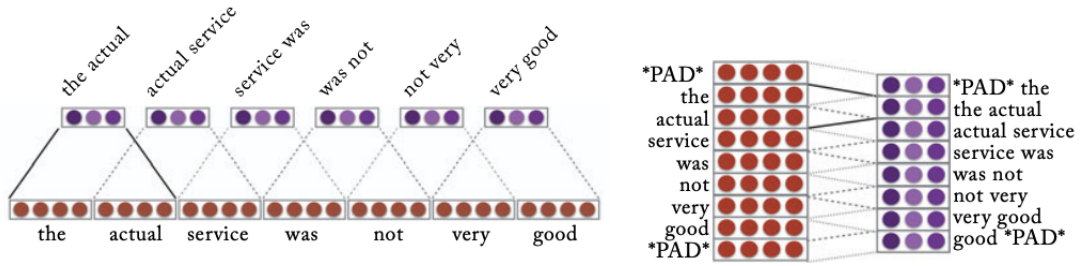


**Figure C.1:** Examples of inputs and outputs of a convolution layer with 3 filters with  $k = 2$  (marked with 3 different shades of purple) applied to a sentence represented as a vector (left) and as a matrix (right) (Goldberg, 2017).

the embeddings. It can be formalised as:

$$f(\mathbf{X}) = g(\text{sum}(\mathbf{W} \odot \mathbf{X}) + b) \quad (\text{C.3})$$

where  $\odot$  is the Hadamard product<sup>2</sup>,  $g$  is some non-linear function, e.g., the Rectified Linear Unit (ReLU)<sup>3</sup>, while  $\mathbf{W}$  and  $b$  are the trainable parameters. Note that the output of the filter is a *scalar*, so its application across all  $k \times d$  windows within the sentence yields a vector.<sup>4</sup> In practice, convolution layers often comprise multiple filters, each associated with a different set of weights, in order to detect different patterns in the input. In such instance, the output of a convolution layer applied to text is a matrix (see the right hand side of Figure C.1).

The filter can be applied at each possible instantiation of a window, or some windows can be skipped as the filter ‘slides across’ the sentence. This is controlled by a hyper-parameter called **stride**. For example, with stride of 2, only windows starting at indexes divisible by 2 are considered. Note that with a filter of stride 1 and the sentence of length  $l$ , the representation resulting from the full filter application (across the whole sentence) is of size  $l - k + 1$ . This is called a **narrow convolution**. An alternative is a **wide convolution** where the sentence is padded with extra  $k - 1$  words to each side, resulting in  $l + k + 1$  dimensional vector. A benefit of wide convolution lies in “all weights in the filter reaching the entire sentence, including the words at the margins” (Kalchbrenner et al., 2014).

**Pooling** The pooling layers reduce the dimensionality of the convolution layer’s outputs, while maintaining their original structure. This is done by applying a ‘summarising’

<sup>2</sup>Element-wise multiplication of two matrices:  $(A \odot B)_{ij} = (A)_{ij}(B)_{ij}$

<sup>3</sup> $f(x) = \max(0, x)$

<sup>4</sup>For text, the width of the filter window is set to the dimensionality of the embeddings and spans the full width of the sentence matrix. Note that this is often not the case for CNNs applied to different types of 2D inputs. For instance, filter windows of CNNs in vision research rarely cover the full width of the image (see the right hand side of Figure C.2); consequently, convolution layers with a single filter output matrices rather than vectors.

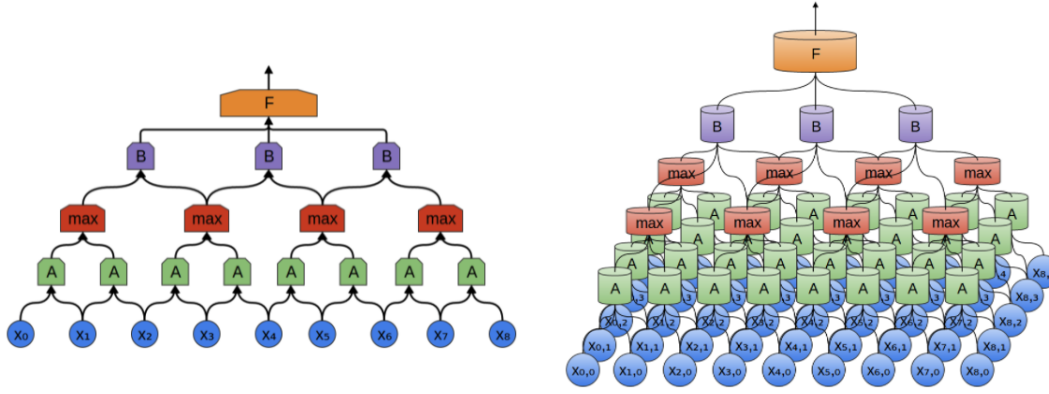


**Figure C.2:** A visualisation of a CNN with 1D filters (left) and 2D filters (right) (Olah, 2014). A and B represent two different convolution layers, while *max* represents a max-pooling layer.

operation over  $n$ -sized blocks of the input;  $n$  is a hyper-parameter which controls the locality of the summarising – pooling can summarise the whole sentence, if the location of the feature is irrelevant, or can work on patches of the input (see the right hand side of Figure C.2). A popular pooling layer is **max-pooling**, which returns the maximum value across the outputs of each filter; i.e., one value is ‘pooled’ for each filter. The representation resulting from the pooling is fed as input into further layers of the model.

**Stacked CNN** Notably convolution+pooling operations can be stacked, by feeding the output of the pooling layer into a new convolution layer. In such a network, the further convolution layers have a potential to encode higher-level, more abstract features. Further, neural architectures can have multiple convolution layers applied in parallel, each with different hyper-parameters, and concatenate their outputs for further processing.

### C.3 Long short-term memory networks

Long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) is a type of a recurrent neural network (RNN) (Elman, 1990). RNNs specialise in sequences – they encode arbitrarily sized sequence inputs as fixed sized vectors. Similarly to CNNs, they are often used as components within larger networks.

**RNN** The key aspect of an RNN is that it keeps a memory of *all* previous inputs, maintained in a **state** vector. The network is defined recursively – the state at a time step  $t$ ,  $\mathbf{s}_t$ , depends on the previous state  $\mathbf{s}_{t-1}$  and the new input  $\mathbf{x}_t$ . The output at time  $t$  is  $\mathbf{s}_t$ :

$$\begin{aligned}\mathbf{s}_t &= g(\mathbf{s}_{t-1}\mathbf{W}^s + \mathbf{x}_t\mathbf{W}^x + \mathbf{b}) \\ \mathbf{y}_t &= \mathbf{s}_t\end{aligned}\tag{C.4}$$

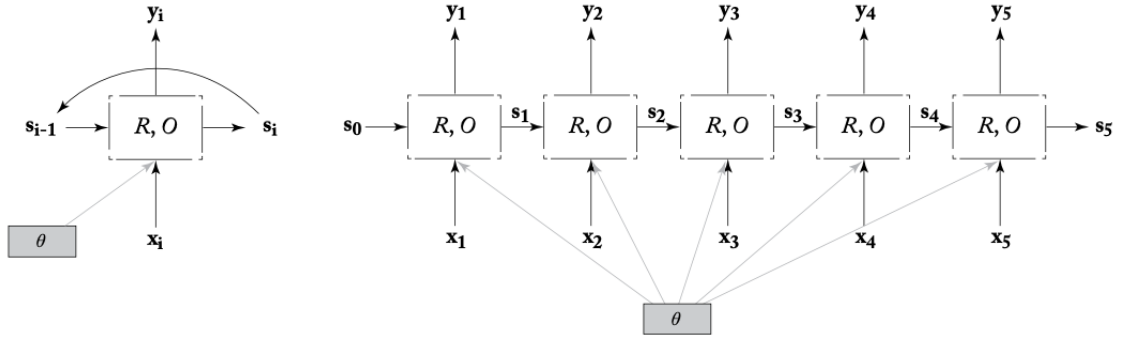


**Figure C.3:** Graphical representation of an RNN – rolled (left) and unrolled RNN (right) (Goldberg, 2017).

where  $\mathbf{W}^s$  and  $\mathbf{W}^x$  are the parameters of the model and  $g$  is a non-linearity. The base of the recursion is an initial step  $s_0$  (see Figure C.3). Note how calculating the state  $s_t$  involves repeated multiplication by the matrix  $\mathbf{W}^s$ . This causes a serious problem for effective training of RNNs – gradients in the later steps diminish quickly during the back-propagation, making it difficult for the network to capture long range dependencies. In the words of Goodfellow et al. (2016, p. 404):

The gradient of a long term interaction has exponentially smaller magnitude than the gradient of a short term interaction. It does not mean that it is impossible to learn, but that it might take a very long time to learn long-term dependencies, because the signal about these dependencies will tend to be hidden by the smallest fluctuations arising from short-term dependencies.

This is often referred to as the **vanishing gradient problem** and has been thoroughly discussed in the literature (Bengio et al., 1993, 1994, Hochreiter, 1991, Pascanu et al., 2013).

**LSTM** is a variant of an RNN designed to avoid the vanishing gradient problem. It does so through an elaborate *gating mechanism*, which controls how much information from the previous state is kept and how much information from the input is introduced at each time step. The state of an LSTM is separated into two separate vectors: (i) the **core memory cell**,  $\mathbf{c}_t$ , and (ii) the **working memory**. At each step, the network makes use of three different gates:

- (i) **i** – controlling how much of the new update,  $\mathbf{z}$ , is kept in the new memory  $\mathbf{c}_t$ ,
- (ii) **f** – controlling how much of the old memory  $\mathbf{c}_{t-1}$  is forgotten, and
- (iii) **o** – controlling what information from the new memory cell  $\mathbf{c}_t$  is kept in the working state  $\mathbf{h}_t$ , which also acts as the output.

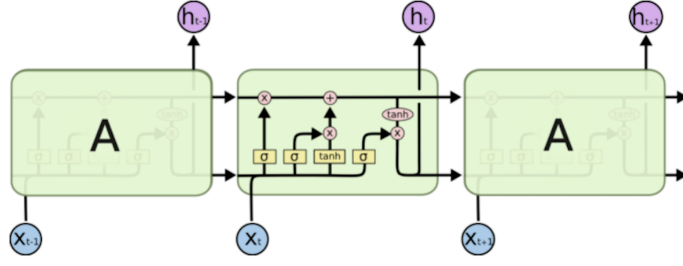


**Figure C.4:** Graphical representation of an LSTM (unrolled) (Olah, 2015). The top horizontal line represents the memory cell  $\mathbf{c}_t$  (see Appendix C.3), the merging of lines represents concatenation.

Each gate is a vector with values in the range  $[0, 1]$ , computed based on the current input  $\mathbf{x}_t$  and the previous working state  $\mathbf{h}_{t-1}$ . Finally, instead of incorporating the input  $\mathbf{x}_t$  directly into the state, like in the RNN, LSTM first computes a candidate update vector  $\mathbf{z}$  based on  $\mathbf{x}_t$  and  $\mathbf{h}_{t-1}$ , which is then *partially* incorporated into  $\mathbf{c}_t$  (how much is kept is controlled by the input gate  $\mathbf{i}$ ). The full formulation of the LSTM is as follows:

$$\begin{aligned}
 \mathbf{c}_t &= \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{z} & \mathbf{i} &= \sigma([\mathbf{x}_t, \mathbf{h}_{t-1}] \mathbf{W}^i + \mathbf{b}^i) \\
 \mathbf{h}_t &= \mathbf{o} \odot \tanh(\mathbf{c}_t) & \mathbf{f} &= \sigma([\mathbf{x}_t, \mathbf{h}_{t-1}] \mathbf{W}^f + \mathbf{b}^f) \\
 \mathbf{y}_t &= \mathbf{h}_t & \mathbf{o} &= \sigma([\mathbf{x}_t, \mathbf{h}_{t-1}] \mathbf{W}^o + \mathbf{b}^o) \\
 & & \mathbf{z} &= \tanh([\mathbf{x}_t, \mathbf{h}_{t-1}] \mathbf{W}^z + \mathbf{b}^z)
 \end{aligned} \tag{C.5}$$

where  $[\mathbf{x}_t, \mathbf{h}_{t-1}]$  is the concatenation of vectors  $\mathbf{x}_t$  and  $\mathbf{h}_{t-1}$ ,  $\odot$  is the Hadamard product and  $\mathbf{W}^*$  and  $\mathbf{b}^*$  are the trainable parameters of the network. See Figure C.4 for a graphical representation of an LSTM cell.

Notably, in the literature one can find many variants on the LSTM architecture presented above. For example, some introduce *peephole connections*, allowing the gates to look inside the memory cell (Gers and Schmidhuber, 2000), while others couple together the forget and input gates. See Greff et al. (2016) for an overview of such alternative architectures.

**Bidirectional LSTM** (biLSTM) is an LSTM variant popular in the NLP literature. It maintains two separate states at each time step  $t$ : (i) the **forward state**,  $(\mathbf{c}_t^f, \mathbf{h}_t^f)$ , based on all *past* states and (ii) the **backward state**,  $(\mathbf{c}_t^b, \mathbf{h}_t^b)$ , based on all the *upcoming* states. Each is generated by a different LSTM, with its own set of weights; the LSTM with the forward state is fed the input sequence in order, while the one with the backward state is fed the sequence in reverse. The output of a biLSTM at time  $t$  is the concatenation of the output vectors from both LSTMs.

## C.4 Transformer models

The transformer architecture, introduced by Vaswani et al. (2017), was motivated by the need to improve on the poor parallelisation of RNNs. Unlike RNNs, the transformer relies on *attention mechanisms* rather than recurrence, to draw global dependencies. This allows for much of the computation to happen at the same time, making it feasible to train larger models. Since its introduction in 2017, the transformer architecture has become extremely popular in NLP, leading to state-of-the-art results for many datasets, across various tasks.

The remainder of this section is structured as follows. First, I provide an overview of the type of attention used in transformer architectures (Appendix C.4.1). Next, I describe the model of Vaswani et al. (2017) (Appendix C.4.2), which forms the basis of most transformer-based models used throughout NLP.

### C.4.1 Attention

Attention mechanisms entered the scene of neural NLP in 2015, through Bahdanau et al.’s (2015) paper on sequence-to-sequence machine translation. At that time, neural machine translation evolved around RNN-based **encoder-decoder** architectures which (i) read in a sequence of source language tokens, (ii) mapped them into a sequence of continuous representations, via an RNN called *the encoder*, and (iii) fed the last output from the encoder (a single vector) into *the decoder* RNN to generate an output sequence in the target language, one token at a time. Bahdanau et al. (2015) proposed a variant on that architecture in which the decoder is exposed to *all* vector outputs from the encoder, across all time steps, instead of just the final one. This was done via feeding the decoder a weighted average of the encoder’s outputs, with weights computed via a bi-variate function of the decoder’s hidden state and the encoder’s outputs, parametrised with a feed-forward neural layer. This mechanism, termed **attention**, allowed the decoder to search for the parts of the source sequence most relevant for predicting the next target token and led to notable improvements first for translation and later for many other NLP tasks.

**Scaled dot-product attention** Similar attention mechanism lies at the core of transformer architectures, where it is used for two purposes: (i) to construct a weighted representation of the encoder’s output for each decoder time step, as in Bahdanau et al. (2015), and (ii) to construct contextualised representations of the tokens from the input sequence. Attention mechanism applied for the latter purpose is called **self-attention** and it is to transformers what recurrent connections are to RNNs.

The attention mechanism of a transformer can be defined in terms of a function taking three parameters:

- (i) a **query** vector –  $\mathbf{q} \in \mathbb{R}^{d_k}$

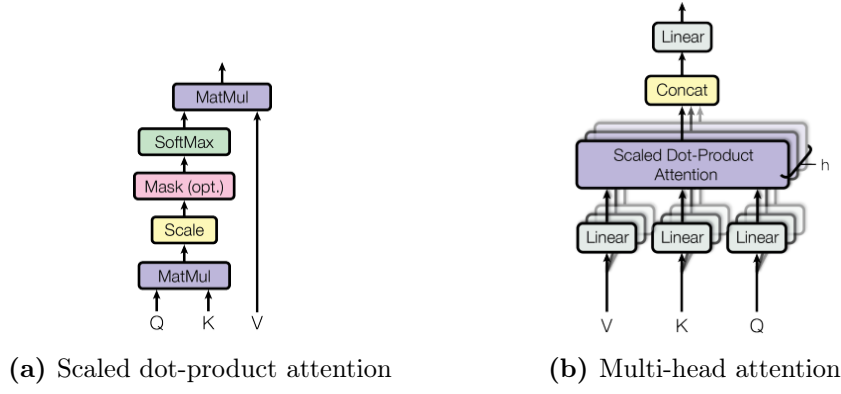


**Figure C.5:** Attention mechanisms used in transformer models (Vaswani et al., 2017).

- (ii) a matrix holding **key** vectors –  $\mathbf{K} \in \mathbb{R}^{n \times d_k}$
- (iii) a matrix holding **value** vectors –  $\mathbf{V} \in \mathbb{R}^{n \times d_v}$

The output of the function is a linear combination of the *value* vectors, with weights computed based on the dot products between the *query* and the *key* vectors as in the following equation:

$$\text{Att}^*(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left( \frac{\mathbf{q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (\text{C.6})$$

Note that there is a one-to-one correspondence between the matrices  $\mathbf{K}$  and  $\mathbf{V}$ ; in practice, they are usually set to the same matrix. The scaling by  $\sqrt{d_k}$  – the square root of the dimensionality of the query and the key vectors – is introduced to prevent large dot products from pushing the softmax function into regions with small gradients (Vaswani et al., 2017).

Importantly, the computation of the attention for a single query vector is independent of that for alternative query vectors. This means that a new sequence of representations, based on multiple queries (e.g., one per each token in a sentence), can be computed in parallel, as in Equation (C.7). See Figure C.5a for a graphical visualisation.

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (\text{C.7})$$

The input matrices of the *Att* function can be outputs from various components of the model, depending on the purpose of the attention layer. For self-attention layers, all  $\mathbf{Q}$ ,  $\mathbf{K}$  and  $\mathbf{V}$  are the same matrix – the input embeddings or the outputs from the previous layer. For decoder’s attention over encoder’s outputs, vectors in  $\mathbf{Q}$  come from the previous layer of the decoder, while vectors in  $\mathbf{K}$  and  $\mathbf{V}$  are the encoder’s outputs across all time steps.

**Multi-head attention** In practice, to allow the model to pick up on different types of dependencies between words, transformers employ a number of scaled dot-product attention components, called **heads**, in each attention layer. The input to the multi-head

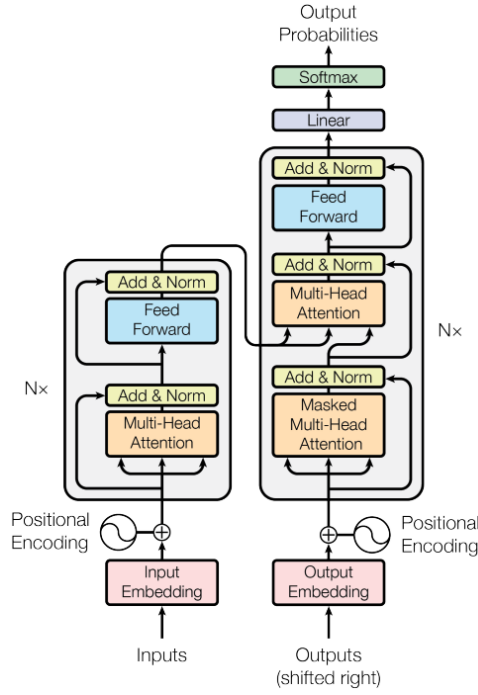


**Figure C.6:** The original transformer architecture (Vaswani et al., 2017).  $N$  is a hyper-parameter which determines the number of encoder/decoder blocks.

attention layer remains the same as before – the query, key and value matrices. These form the basis for inputs to the attention heads – each of the  $h$  heads takes as input linearly projected  $\mathbf{Q}$ ,  $\mathbf{K}$  and  $\mathbf{V}$ . The output of the multi-head attention is the projected concatenation of outputs from all heads, as shown in Figure C.5b and the following equation:

$$\begin{aligned} \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = & \text{Concat}(\text{Att}(\mathbf{Q}\mathbf{W}^{Q1}, \mathbf{K}\mathbf{W}^{K1}, \mathbf{V}\mathbf{W}^{V1}), \\ & \text{Att}(\mathbf{Q}\mathbf{W}^{Q2}, \mathbf{K}\mathbf{W}^{K2}, \mathbf{V}\mathbf{W}^{V2}), \\ & \dots, \\ & \text{Att}(\mathbf{Q}\mathbf{W}^{Qh}, \mathbf{K}\mathbf{W}^{Kh}, \mathbf{V}\mathbf{W}^{Vh})) \mathbf{W}^O \end{aligned} \quad (\text{C.8})$$

where  $\mathbf{W}^{Q*} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $\mathbf{W}^{K*} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $\mathbf{W}^{V*} \in \mathbb{R}^{d_{\text{model}} \times d_v}$  and  $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$  are trainable parameters of a model.

#### C.4.2 The original transformer (Vaswani et al., 2017)

Vaswani et al. (2017) proposed the transformer for the task of machine translation. Consequently, their model is a transducer with an encoder-decoder structure. Both encoder and decoder are constructed from blocks which use multi-head attention and fully connected feed-forward layers, as presented in Figure C.6. These blocks can be stacked on top of one another to increase the depth of the model; Vaswani et al. (2017) use  $N = 6$  blocks for both the encoder and the decoder.

**The encoder** blocks consist of a layer of multi-head self-attention, followed by a fully connected feed-forward layer. The latter is formalised as follows:

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}^1 + \mathbf{b}^1)\mathbf{W}^2 + \mathbf{b}^2 \quad (\text{C.9})$$

It is a *position-wise* layer, which means that the function is applied to each vector in a sequence “separately and identically” (Vaswani et al., 2017).

Both layers in a block make use of residual connections (He et al., 2016) (the input of the layer is added to the layer’s output), followed by layer normalisation (Ba et al., 2016). For example, the output of the self-attention layer is  $\text{LayerNorm}(\mathbf{X} + \text{MultiHead}(\mathbf{X}, \mathbf{X}, \mathbf{X}))$ . These techniques are employed to address problems associated with training very deep neural models (see cited papers for more details). Note that because of the residual connections, the input of the layers has to have the same dimensionality as their output.

**The decoder** takes two types of inputs: (i) (sequential) output from the encoder and (ii) embeddings of the previously generated symbols (the model is auto-regressive). Each decoder block has three layers: multi-head self-attention, multi-head attention over the encoder’s outputs and the position-wise feed-forward network (see Equation (C.9)). Importantly, the decoder’s self-attention inputs are masked, so that at time step  $t$  the model cannot attend to positions  $t + n, n \geq 0$  – the model’s prediction depends only on the known outputs.<sup>5</sup> As in the encoder, each of the layers makes use of residual connections and layer normalization.

**Input Embeddings** The input to the encoder is the sum of learnable input embeddings and *positional encodings*, both of which share the same dimensionality  $d_{\text{model}}$ . The latter are used to inject information about the order of tokens within a sequence – unlike the recurrent connections in RNNs, the attention mechanism does not inherently encode that information. Let  $\mathbf{q}^p \in \mathbb{R}^{d_{\text{model}}}$  be the positional encoding for a sequence position  $p$  and  $i \in$  be one of its dimensions. The encoding  $\mathbf{q}^p$  is constructed from the outputs of  $d_{\text{model}}$  using sinusoid functions – one function per each dimension  $i$ , varying in wavelengths from  $2\pi$  to  $10000 \times 2\pi$ :

$$\mathbf{q}_i^p = \begin{cases} \sin\left(\frac{p}{10000^{i/d_{\text{model}}}}\right), & \text{if } i \bmod 2 = 0 \\ \cos\left(\frac{p}{10000^{(i-1)/d_{\text{model}}}}\right), & \text{otherwise} \end{cases} \quad (\text{C.10})$$

The decoder’s input is constructed in a similar way, through the summation of the output embeddings and the positional encodings. Importantly, the weight matrices for all

---

<sup>5</sup>This is done in the softmax from Equation (C.7) – all input values that correspond to illegal connections are set to  $-\infty$ .

three: the input embeddings, the output embeddings and the pre-softmax projection layer are shared.

**Beyond the transduction** The transformer architecture of [Vaswani et al. \(2017\)](#) can be easily adapted for tasks which do not naturally lend themselves to transduction (e.g., classification or language modeling) by abandoning either the encoder or the decoder. One of the most influential pre-training models of recent years – the Bidirectional Encoder Representations from Transformers (BERT) model ([Devlin et al., 2019a](#)) – is based on [Vaswani et al. \(2017\)](#)’s encoder.



## ALLOWING FOR MORPHOLOGY IN NEURAL NLP

---

What follows is a review of approaches proposed to allow for morphology in neural NLP. Its main purpose is to suitably situate the models and techniques used throughout this thesis, providing broader context. Approaches to allowing for morphology in neural NLP can be categorised based on whether they operate on the level of: (i) morphs, (ii) characters, (iii) character n-grams/word-pieces or (iv) morphosyntactic tags. In the remainder of this section I provide a more detailed overview of research within these categories, with one section devoted to each category.

The relevant literature is vast and spans many NLP tasks, including POS tagging (dos Santos and Zadrozny, 2014, Ling et al., 2015, Rei et al., 2016), named entity recognition (dos Santos and Guimarães, 2015, Lample et al., 2016, Rei et al., 2016), parsing (Ballesteros et al., 2015, Legrand and Collobert, 2016, Tsarfaty et al., 2010, 2013), reading comprehension (Yang et al., 2017) and machine translation (both on the input and the output side) (Bradbury and Socher, 2016, Chen et al., 2018, Chung et al., 2016, Huck et al., 2017, Lee et al., 2017a, Luong and Manning, 2016, Luong et al., 2015, Sennrich et al., 2016, Wu et al., 2016). Given the breadth of relevant work, instead of providing an exhaustive survey, I only review selected works that are representative of the general research landscape. Most of the works I select fit within the literature on static (i.e., non-contextualised) word embeddings, language modeling/contextualised word embeddings or machine translation.

Incorporating subword-level information into a language model is not a novel idea. It has been widely investigated in the past and shown to yield improvements over the word-only models (Alexandrescu and Kirchhoff, 2006, Bilmes and Kirchhoff, 2003, Elbeze and Derouault, 1990, Geutner, 1995, Kirchhoff et al., 2006, Whittaker and Woodland, 2000). In the following review, I focus on the more recent lines of research which integrate subword knowledge into deep RNN/transformer-based neural networks. Most work I

review relates to encoding, rather than generating text sequences, since the latter is not directly relevant for this thesis.

## D.1 Composition of morphs

One line of research into allowing for morphology explores methods of deriving them from the distributional representations of their morphs (see Section 2.2). Such methods give a model direct access to word chunks identified by linguists as meaningful within a language. The morphs can be obtained either from a morphologically annotated dictionary or an external morphological analyser, such as Morfessor (Smit et al., 2014) or CHIPMUNK (Cotterell et al., 2015). Note how such approaches can work well for concatenative morphology, but they are not well suited to other morphological patterns (see Section 2.2.4).

**Overview of approaches** Lazaridou et al. (2013) were one of the first to explore morph-based inputs in the context of deep neural models. They utilised compositional distributional semantic models (CDSMs), originally introduced to build representations of phrases by combining meanings of individual words (Baroni and Zamparelli, 2010, Coecke et al., 2010, Mitchell and Lapata, 2010). Lazaridou et al. (2013) investigated the usefulness of a range of different CDSMs for constructing static embeddings of English word derivations from two morphs: the stem and an affix. All models were evaluated based on their ability to capture human semantic similarity judgements and to recreate existing full-word embeddings, encoding word co-occurrence statistics. The overall best CDSM was the *full additive model* (Zanzotto et al., 2010) which adds the embeddings of the morphs after pre-multiplying them by weight matrices. Other well performing models included weighted addition of the stem and affix embeddings and a *lexical functional model* (LEXFUNC) which treats the affixes as functions (matrices) over stems (vectors) – an approach similar to that in Baroni and Zamparelli (2010).

Kisselew et al. (2015) and Padó et al. (2016) expanded on Lazaridou et al.’s (2013) work, focusing on German derivational morphology. To perform an in-depth analysis of the models’ performance, the authors constructed regression models that give insight into which linguistic features and derivational patterns are the most challenging for the models. They found that the performance strongly depends on the type of a derivational pattern and that within-POS derivations and those creating new argument structures were the hardest to model. Interestingly, both studies also revealed that, for German, the LEXFUNC model, which performed well for English, was worse than the baseline that simply returned the stem’s embedding.

Due to its simplicity and good performance, the additive composition of morph vectors

has been used in a number of works that followed [Lazaridou et al. \(2013\)](#), including [Qiu et al.’s \(2014\)](#) adaptation of the Continuous Bag of Words model (CBOW) ([Mikolov et al., 2013a](#)) (a sister model of SKIP-GRAM from Appendix C.1) and the language model embeddings of [Botha and Blunsom \(2014\)](#). [Bhatia et al. \(2016\)](#) and [Cotterell et al. \(2016\)](#) also employ an additive model of morph semantics in their graphical models for static word embedding construction. [Bhatia et al. \(2016\)](#) construct word embeddings ‘from scratch’ and introduce morphological knowledge to the model as a prior. The word-form similarity assumptions captured in this prior are overridden by distributional statistics: the more frequent the word, the more distributional information influences its representation. [Cotterell et al. \(2016\)](#), on the other hand, utilise a morphological lexicon to smooth existing, pre-trained word representations, in the spirit of *retrofitting* of [Faruqui et al. \(2015\)](#). Their graphical model also allows for extrapolation of vectors for unseen inflected forms.

An alternative morph composition approach, complementary to those from [Lazaridou et al. \(2013\)](#) is to use a neural network. Among such approaches is the work of [Luong et al. \(2013\)](#), who proposed two models for construction of morphologically aware (static) word embeddings, based on recursive neural networks – generalisations of RNNs (Appendix C.3) to tree structures. Their first model learns the embeddings as features in a language model, while the second is trained to reconstruct existing, pre-trained embeddings. In a related line of research, [Cotterell and Schütze \(2018\)](#) found RNNs to be the best performing model of morph composition for their joint model of words’ orthography, morphological segmentation and the composition of morph vectors. Their RNN method outperformed alternative composition methods, including additive composition, on [Lazaridou et al.’s \(2013\)](#) task of reconstructing existing vectors of complex words.

In most recent line of research, morphs have also been used as inputs to transformer-based architectures. The specifics of the taken approach vary across the papers, with some works feeding the morphs directly into the transformer-based encoders ([Mohseni and Tebbifakhr, 2019](#), [Park et al., 2020](#)) and others first processing them locally, at the word-token level, and later feeding them into a larger sentence-level encoder ([Nzeyimana and Niyongabo Rubungo, 2022](#)).

## D.2 Composition of characters

Almost all of the approaches outlined above presuppose that the words have been segmented into morphs, which requires an external morphological analyser or a morphologically annotated dictionary.<sup>1</sup> Further, they are only suitable for concatenative morphological patterns (see Section 2.2.4). An alternative, free of such constraints, is to build models that

---

<sup>1</sup>The model of [Cotterell and Schütze \(2018\)](#) is an exception since it is able to predict a word’s vector given its surface form through marginalising out the segmentation and the transduction probabilities.

learn the underlying morphology automatically, by directly processing characters. Building models that operate directly on characters has many benefits. With direct access to the most basic input signal, the models can automatically derive features most predictive of a given task, in a fully end-to-end fashion. Character-level models are also well predisposed to be robust against misspelling errors and good at handling spelling variations. What is more, they are said to be more parameter efficient (Lee et al., 2017a). However, they are also associated with a few drawbacks. First, character-level models are often more costly to train/run, since character sequences are much longer than word/morph sequences and, in the words of Xue et al. (2022) “computational costs of machine learning models tend to scale with sequence length” (Banar et al., 2020, Clark et al., 2022, Xue et al., 2022). Further, they have no direct means of memorising the meaning of a word/word chunk; such meaning always has to be computed via a function over a word’s characters. While, in theory, this does not prevent the models from memorising the relevant lexicosemantic/morphosyntactic information, it may require deeper architectures and/or more training data/iterations. This is especially so, given that the models also need to identify which character chunks are worth memorising in the first place.

**Overview of approaches** In this line of research, Cao and Rei (2016) extended SKIP-GRAM (see Appendix C.1) by building representations of target-words with a biLSTM (see Appendix C.3) over the sequence of its characters, followed by an attention layer. Apart from learning word embeddings, their model also learns morph boundaries and performs well as a morphological analyser. However, it performs poorly on word similarity and semantic word analogy evaluation. Pinter et al. (2017) also used a biLSTM over characters to create (static) word representations for out-of-vocabulary word forms by training a neural model to reconstruct pre-built embeddings. They called this method MIMICK for its ability to mimic words’ embeddings based purely on their spellings.

Sutskever et al. (2011) were one of the first to adapt RNN-based language models to operate *fully* on character-level – i.e., on non-pre-tokenised, raw text sequence, predicting each word one character at the time. Although their English language model was able to learn the vocabulary rather well, the sentences it generated were meaningless – the model struggled to encode semantics and keep track of long distance dependencies between words. Chung et al. (2017) and Lee et al. (2017b) likewise proposed fully character-level language encoders. Chung et al. (2017) did so in the context of language modeling and employed a *hierarchical multiscale RNN*, designed to automatically detect boundaries between words/phrases and the hierarchical structure of data. Lee et al. (2017b)’s model was proposed for machine translation and is a composition of a CNN layer, multi-layer highway network (Srivastava et al., 2015a) and a single-layer RNN. In a related line of research, Kawakami et al. (2017), Ling et al. (2015), Miyamoto and Cho (2016) and Kim

et al. (2016) proposed models which work on pre-tokenised text, encoding a text sequence by first constructing word embeddings based on characters and then processing these representations at the sentence-level. Kawakami et al.’s (2017) language model is based on a hierarchical LSTM (Sordoni et al., 2015) with two encoders – one operating on character-level and one on word-level – and a character-level decoder. Their model also utilises a caching mechanism to allow for reusing of previously generated words. Ling et al. (2015) used a similar model but employed a biLSTM over characters – the final states of the forward and backward passes were multiplied by weight matrices and summed to create the final representation. Miyamoto and Cho (2016) followed Ling et al. (2015) in using a biLSTM, but they additionally made use of word-level embeddings; their model takes as input a mixture of word and character-level representations and uses a gating mechanism to control how much information comes from each source. Kim et al. (2016), on the other hand, built character-based word representations with a single-layer character-level CNN (CHARCNN) (see Appendix C.2) followed by a two-layer highway network (Srivastava et al., 2015b). They fed such representations into an LSTM-based language model. The resulting model achieved results comparable to the (word-level) state of the art at the time (Zaremba et al., 2015) on the Penn Treebank (Marcus et al., 1993), despite having 60% fewer parameters. Their CNN over characters approach also outperformed the additive composition of morphs from Botha and Blunsom (2014).

The strengths of Kim et al.’s (2016) method were later confirmed by Józefowicz et al. (2016) who tested RNN language models with various architectures and hyper-parameter settings. Their best performing model, which established the new state of the art on the One Billion Word Benchmark (Chelba et al., 2014), used a CNN over characters at the input level. Józefowicz et al. (2016) also experimented with two approaches of incorporating character-level representations at the output level: the CNN SOFTMAX and the CHAR LSTM, predicting a target-word one character at a time, akin to the model of Sutskever et al. (2011). Both methods visibly decreased the number of parameters, but also reduced perplexity. The models of Józefowicz et al. (2016) later served as the basis for the contextualised ELMO embeddings (Peters et al., 2018), which are also constructed with a biLSTM network, taking CNN-constructed inputs, akin to those of Kim et al. (2016). Notably, Peters et al.’s (2018) work was one of the first within the newly emerging paradigm in the NLP field, which can be characterised as moving away from static word embeddings and towards contextualised embeddings, obtained from large pre-trained language models.

Gerz et al. (2018) further addressed the problem of injecting character-level knowledge into the *output* side of a language model, by *retrofitting* (Faruqui et al., 2016) the word-level prediction parameters of Kim et al.’s (2016) language model to be closer to the input, CNN-built embeddings. This solution has the benefit of injecting morphological knowledge,

while maintaining the word-specific semantic information necessary for successful next word prediction. Gerz et al.’s (2018) approach led to improved performance on 46 out of 50 languages studied in the paper, with the strongest effect for languages with high type-to-token ratios, including Polish, Russian, Finnish and Estonian.

In recent years, many works have also proposed character-level versions of transformer (masked) language models. Among these works are (i) El Boukkouri et al. (2020) and Aguilar et al. (2021), which propose models operating on pre-tokenised text, as well as (ii) Al-Rfou et al. (2019), Libovický and Fraser (2020), Banar et al. (2020), Ma et al. (2020), Xue et al. (2022) Clark et al. (2022) and Tay et al. (2022), which introduce models operating directly on raw sequences of characters/bytes. In general, while over the past years the field has been dominated by the chunk-level tokenisation (see the following section), the focus appears to be shifting towards character-based approaches.

### D.3 Composition of character n-grams and word-pieces

In between the morph and character-based approaches lie those based on character n-grams or automatically inferred word segments. These kinds of tokenisations direct a model towards the potentially informative word chunks and supply it with more direct means of memorising chunk-related information. However, they do so at the cost of stripping it from access to the underlying characters. This not only prevents a model from capturing non-concatenative morphological patterns, but it can also be sub-optimal for concatenative morphology – depending on how they are induced, the chunks might not correspond to actual morphs, meaning that a morph can be split into multiple chunks or it may be subsumed within representations of longer character sequences.<sup>2</sup>

Notably, tokenisation into automatically inferred word segments is currently the most popular means of injecting subword-level knowledge into NLP models – most pre-trained transformer-based models take such word segments as inputs. The most popular techniques include: (i) BPE (Gage, 1994, Sennrich et al., 2016), (ii) WORD-PIECE (Schuster and Nakajima, 2012) and UNIGRAM LM (Kudo, 2018). All three are described later in this section.

**Overview of approaches** In the word embedding literature, n-grams have been famously explored by Bojanowski et al. (2017a) in their SKIP-GRAM extension. Similarly to Cao and Rei (2016), Bojanowski et al. (2017a) proposed a (more) morphologically aware variant of the model, but instead of using a biLSTM to represent a target-word, they simply summed the vectors of all its character n-grams beyond a certain length. In word

---

<sup>2</sup>For more discussion on shortcomings of such tokenisation approaches see e.g., Klein and Tsarfaty (2020) or Bostrom and Durrett (2020).

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Algorithm 1</b> Byte-pair encoding (Sennrich et al., 2016; Gage, 1994)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <b>Algorithm 2</b> Unigram LM (Kudo, 2018)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | <b>Algorithm 1:</b> BPE-dropout                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <pre> 1: Input: set of strings <math>D</math>, target vocab size <math>k</math> 2: <b>procedure</b> BPE(<math>D, k</math>) 3:   <math>V \leftarrow</math> all unique characters in <math>D</math> 4:   (about 4,000 in English Wikipedia) 5:   <b>while</b> <math> V  &lt; k</math> <b>do</b> <math>\triangleright</math> Merge tokens 6:     <math>t_L, t_R \leftarrow</math> Most frequent bigram in <math>D</math> 7:     <math>t_{\text{NEW}} \leftarrow t_L + t_R</math> <math>\triangleright</math> Make new token 8:     <math>V \leftarrow V + [t_{\text{NEW}}]</math> 9:     Replace each occurrence of <math>t_L, t_R</math> in 10:    <math>D</math> with <math>t_{\text{NEW}}</math> 11:   <b>end while</b> 12:   <b>return</b> <math>V</math> 13: <b>end procedure</b> </pre> | <pre> 1: Input: set of strings <math>D</math>, target vocab size <math>k</math> 2: <b>procedure</b> UNIGRAMLM(<math>D, k</math>) 3:   <math>V \leftarrow</math> all substrings occurring more than 4:   once in <math>D</math> (not crossing words) 5:   <b>while</b> <math> V  &gt; k</math> <b>do</b> <math>\triangleright</math> Prune tokens 6:     Fit unigram LM <math>\theta</math> to <math>D</math> 7:     <b>for</b> <math>t \in V</math> <b>do</b> <math>\triangleright</math> Estimate token ‘loss’ 8:       <math>L_t \leftarrow p_\theta(D) - p_{\theta'}(D)</math> 9:       where <math>\theta'</math> is the LM without token <math>t</math> 10:    <b>end for</b> 11:    Remove <math>\min( V  - k, \lfloor \alpha  V  \rfloor)</math> of the 12:    tokens <math>t</math> with highest <math>L_t</math> from <math>V</math>, 13:    where <math>\alpha \in [0, 1]</math> is a hyperparameter 14:  <b>end while</b> 15:  Fit final unigram LM <math>\theta</math> to <math>D</math> 16:  <b>return</b> <math>V, \theta</math> 17: <b>end procedure</b> </pre> | <pre> <math>current\_split \leftarrow</math> characters from input word; <b>do</b>   <math>merges \leftarrow</math> all possible merges<sup>1</sup> of tokens   from <math>current\_split</math>;   <b>for</b> <math>merge</math> from <math>merges</math> <b>do</b>     /* The only difference     from BPE */     remove <math>merge</math> from <math>merges</math> with the     probability <math>p</math>;   <b>end</b>   <b>if</b> <math>merges</math> is not empty <b>then</b>     <math>merge \leftarrow</math> select the merge with the     highest priority from <math>merges</math>;     apply <math>merge</math> to <math>current\_split</math>;   <b>end</b> <b>while</b> <math>merges</math> is not empty; <b>return</b> <math>current\_split</math>; </pre> |
| (a) BPE (Gage, 1994, Sennrich et al., 2016)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | (b) UNIGRAM LM (Kudo, 2018)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | (c) BPE-DROPOUT (Provilkov et al., 2020)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |

**Figure D.1:** Popular algorithms for word segmentation. Figures (a) and (b) from Bostrom and Durrett (2020). Figure (c) from Provilkov et al. (2020).

similarity evaluation, their model – FASTTEXT – outperformed the CBOW and SKIP-GRAM baselines across a range of languages, including Russian. It was also the leading model in the syntactic word analogy tasks (Mikolov et al., 2013a), but did not perform as well for semantic analogy. Similarly to Bojanowski et al. (2017a), Wieting et al. (2016) and Ataman and Federico (2018) also use summation of n-gram embeddings in their work – the first do so while constructing static embeddings, the second do so in the context of machine translation.

Another chunk-based approach is to construct words from an automatically learned vocabulary of informative word segments. Sennrich et al. (2016) build such vocabulary with the byte pair encoding (BPE) algorithm (Gage, 1994) (see Figure D.1a). Starting with a vocabulary of individual characters, BPE iteratively replaces the most common pair of symbols with a single, unused symbol and adds it to the vocabulary. This causes frequent character sequences to be merged into a single symbol. The algorithm runs for  $d$  iterations;  $d$  defines the size of the final vocabulary:  $k = |V| + d$ , where  $|V|$  is the size of the initial character vocabulary. The choice of an optimal value for  $d$  largely depends on the architecture of the model and language(s) being modeled (Ding et al., 2019). Optimising the granularity of the BPE vocabulary is an active area of research (Ding et al., 2019, Kharitonov et al., 2021, Salesky et al., 2020, Vilar and Federico, 2021).

Another iterative technique, very similar to BPE, was proposed a few years earlier by Schuster and Nakajima (2012). The only difference between their WORD-PIECE algorithm and BPE lies in that, in each recursive pass, the first does not replace the *most common* pair of symbols, but instead builds a new language model and replaces a pair which leads to the greatest increase in *the likelihood of the training data*. In 2016, Google made use of WORD-PIECES to develop their neural machine translation technology (Johnson et al.,

2017, Wu et al., 2016) (which was state of the art at the time) and, later, to develop BERT (Devlin et al., 2019b) – a widely used pre-trained masked language model.

Kudo (2018) proposed another technique for automatic induction of subword vocabulary. In contrast to the deterministic WORD-PIECE and BPE algorithms, their unigram language model based segmentation algorithm (UNIGRAM LM) can account for multiple segmentations of a single word. This allows for exposing a downstream model to various possible segmentations of words during training, making it more robust to errors in segmentation. It can also enhance a model’s ability to learn morphology. UNIGRAM LM starts with a (heuristically defined) superset of the final vocabulary and progressively prunes it until the desired size  $t$  is reached (see Figure D.1b). At each recursive step, the algorithm fits a unigram language model to the data with the EM algorithm (Dempster et al., 1977) and prunes symbols whose removal results in the lowest increase in the model’s loss. How many symbols are removed is a hyper-parameter, often set to 10% or 20%. Notably, the base characters are never removed. As a last step, the final language model is fit to the data and the probabilities of all symbols are saved to be used for future segmentations.

In related work, Provilkov et al. (2020) proposed another method of regularising subword segmentation during training of a downstream model, called BPE-DROPOUT. As the algorithm segments a word, merging pairs of subsequent symbols according to the table expressing merge rules and their priorities, some of its previous merge operations are undone – each with a probability  $p$  (see Figure D.1c). Provilkov et al. (2020) use  $p > 0$  for downstream training, which allows for the same word to be associated with many segmentation variants. During inference  $p$  is set to 0 (the original BPE is used). The method is simpler than UNIGRAM LM, does not require the heuristic pre-definition of initial vocabulary and in Provilkov et al. (2020) led to better results for a number of translation tasks. Subsequent work has also confirmed that BPE-DROPOUT outperforms BPE in the context of machine translation (Amrhein and Sennrich, 2021).

## D.4 Paradigm and tag-based approaches

The final group of approaches differs from the rest in that it does not compose full-word representations based on subword units. Instead, it directly utilises a word’s morphosyntactic description. For instance, Vulić et al. (2017) incorporated morphological information at the post-processing stage, by using the ATTRACT-REPEL model (Mrkšić et al., 2017) to pull embeddings of inflectional variants of a word towards one another and push derivational antonyms apart. The linguistic constraints guiding this process were generated automatically, based on hand-crafted, language-specific rules. Notably, such rules can be hard to create for languages with sophisticated morphology, and in particular for those with high degree of allomorphy (see Section 2.2.3). The authors themselves point out

that their rules occasionally generate incorrect linguistic constraints, for instance the pair (press, impress) is recognised as antonymous.

In a different line of work, [Avraham and Goldberg \(2017\)](#) proposed a generalisation of FASTTEXT ([Bojanowski et al., 2017a](#)) in which words are represented not as sets of n-grams, but rather sets of linguistic properties which include the surface form, lemma and the morphosyntactic tag. [Cotterell and Schütze \(2015\)](#), on the other hand, focused on explicitly learning German embeddings that exhibit *morphological* similarity via incorporating morphological analysis as an additional training objective. The embeddings were learned using a log-bilinear language model trained to predict not only the following word but also its morphological tag, given the *word-only* context. Using a new metric, termed *MorphoSim*, they demonstrate that in the semantic space learned by their model, the representations lying close to each other have similar morphological tags. In contrast, SKIP-GRAM ([Mikolov et al., 2013a,b](#)) embeddings and embeddings learned by a standard, solely word-predicting language model have neighbours which are further apart in terms of morphology.

## D.5 Comparison of approaches

[Vania and Lopez \(2017\)](#) were one of the first to conduct an empirical comparison of different types of subword-based inputs to neural models. They focused on language modeling across a range of typologically different languages and limited their analysis to four types of subword units – (i) characters, (ii) character tri-grams, (iii) Morfessor’s morphs ([Smit et al., 2014](#)) and (iv) subword sequences learned using BPE – and three different composition methods – (i) biLSTM, (ii) CNN and (iii) addition. Notably, for characters, they employed only CNN and biLSTM, while for n-grams and morphs – biLSTM and addition. They found that all types of subword features improve performance over the word-level features, unless the subword units are composed using addition. The overall best embeddings, across different languages, were those created using biLSTM over character tri-grams. Contrary to [Józefowicz et al. \(2016\)](#), they showed that biLSTMs are, in most cases, a better character composition method than CNNs. This discrepancy may be due to the differing method of composing the final states of the forward and backward LSTMs – [Józefowicz et al. \(2016\)](#) use concatenation, while [Vania and Lopez \(2017\)](#) follow [Ling et al. \(2015\)](#) in using weighted addition. Importantly, [Vania and Lopez \(2017\)](#) found that character sequences obtained using BPE give better results than morphs generated by Morfessor and that none of the subword approaches was better than biLSTM combining morphs from hand-annotated morphological analyses.

[Zhu et al. \(2019\)](#) expanded on [Vania and Lopez’s \(2017\)](#) work by considering more tasks: (i) word similarity prediction, (ii) dependency parsing, and (iii) fine-grained entity

typing, but they considered less languages.<sup>3</sup> They investigated three subword techniques – (i) morphs from CHIPMUNK (Cotterell et al., 2015), (ii) morphs from Morfessor (Smit et al., 2014) and BPE segments – and three composition functions – (i) addition, (ii) single-head attention and (iii) multi-head self-attention (see Appendix C.4.1). Based on their experimentation, they conclude that the choice of most suitable setting depends on the task at hand and the languages involved – there was no universally best configuration. Most relevant for this thesis, well-performing configurations for Finnish and Turkish dependency parsing involved summation of CHIPMUNK/BPE inputs. For English and German it was the summation of CHIPMUNK/BPE inputs, as well as Morfessor inputs composed with single headed attention.<sup>4</sup>

In a closely related study, Vania et al. (2018) test if subword-level models learn morphology by comparing the dependency parsing performance of three subword-level parsers – (i) one which takes character inputs and processes them with an LSTM (CHARLSTM), (ii) one taking CHARCNN inputs, and (iii) one which processes character tri-grams with an LSTM – to the performance of an oracle model which takes gold inflectional tags as input. None of the character-level models matched the performance of the oracle, but among them CHARLSTM emerged as the best model.

More recently, Bostrom and Durrett (2020) compared BPE and UNIGRAM LM based on (i) their influence on the performance of a masked language model (they used RoBERTa (Liu et al., 2019b)), and (ii) a comparison of the resulting tokens to reference morphological segmentation, showing that UNIGRAM LM is superior to BPE on both tasks.

Vylomova et al. (2017), Sennrich (2017), Shapiro and Duh (2018), Cherry et al. (2018), Durrani et al. (2019), Gupta et al. (2019), Belinkov et al. (2020), Amrhein and Sennrich (2021), Li et al. (2021) and Libovický et al. (2021) conducted similar comparisons for neural machine translation (NMT). These works compared various subsets of: BPE, BPE-DROPOUT, character inputs fed directly to the LSTM/transformer-based translation model, word representations created with CHARCNN/CHARLSTM, and Morfessor. The picture that emerges signals that models with more fine-grained inputs (characters or small word-pieces): (i) are better suited for capturing non-concatenative morphology, (ii) are better at handling unseen or very rare tokens, and (iii) can perform on par or better than coarse-grained inputs. However, the latter comes at the expense of higher computation costs, and may not be achievable on some datasets (Libovický et al., 2021). The results of Durrani et al. (2019) and Li et al. (2021) also suggest that character-level inputs lead to better capturing of morphology. Next in line are the Morfessor word chunks, which are in turn followed by

---

<sup>3</sup>Vania and Lopez (2017) considered 10 languages, Zhu et al. (2019) considered 5.

<sup>4</sup>These insights are taken from Figure 2 in Zhu et al. (2019). Notably, the figure also suggests that there are other well performing configurations involving BPE and CHIPMUNK (bpe.ww, sms.mp, bpe.mp), but it is not clear which composition function they employed (ww stands for insertion of a word token, mp stands for an inclusion of an additional position embedding, incorporated via a multiplicative function).

BPE. [Belinkov et al. \(2020\)](#) also arrived to the same ranking when probing representations from NMT models via the morphological tagging task. Finally, experiments of [Shapiro and Duh \(2018\)](#) suggest that both BPE and character-based approaches are complementary – their best translation model employed both subword techniques.



---

## SENTENCE SIMPLIFICATION DECISIONS

---

The decisions regarding which relations to keep during the sentence truncation (see counterfactual treebank construction; Section 4.2.2) are guided by the following three constraints:

1. The sentences should be grammatical and acceptable to native speakers.
2. The sentences should be simple enough so that they remain acceptable after the reordering of their core arguments; i.e., the reordering should not result in long dependencies.
3. The sentences should be complex enough, so that the task is not too trivial and so that they are as natural as possible to native speakers.

The simplified sentences are constructed based on a filtered Universal Dependencies treebank (see Section 4.2.1). To create the simplified test set, ready for reordering, I follow the following bottom-up procedure for each sentence. First, the transitive verb in the original sentence is set to be the simplified sentence’s root. Next, I create a simplified version of the sub-tree rooted by the verb by only attaching dependents linked through selected relation types – this is done recursively; i.e., sub-trees of all attached dependents are also truncated. I primarily exclude relations which are likely to introduce big sub-trees. I keep the relations which introduce core arguments, as well as those linking to non-core arguments and modifiers with typically small sub-trees.

The following sections describe which relations are excluded and which are kept. These decisions were made based on my analysis of the following treebanks: Polish PDB, Russian SynTagRus, Finnish TDT and Estonian EDT.

## E.1 Accepted/rejected relations overview

### E.1.1 Core clausal relations

*nsubj, iobj, obj, csubj, ccomp, xcomp*

All are accepted. However, note that *csubj*, *ccomp* and *xcomp* relations will not be present for the new root. This is because the verbs are selected to have noun subjects and not to include *ccomp* and *xcomp* relations (see Section 4.2.1). The latter is done to avoid problems that arise when sentences with *ccomp* and *xcomp* relations are simplified – skipping those relations can result in unacceptable sentences, while keeping them can result in long dependencies when the words are reordered.

The *xcomp* and *ccomp* relations can appear, however, e.g., if they link dependents of noun participles which are arguments of the verb/noun modifiers of other nouns. In such cases they will be included for the sake of acceptability, although such cases are very unlikely to arise.

### E.1.2 Non-core clausal relations

*advcl, advmod, aux, cop, discourse, dislocated, expl, mark, obl, vocative*

#### Accepted:

- *aux* – Required for grammaticality.
- *expl:pv* – Links Polish reflexive clitics.
- *obl:arg* – In Polish *obl:arg* is used to link verb arguments.

#### Conditionally accepted:

- *advmod* – The relation is kept in Polish and Russian when it links to the negative particle ‘nie’/‘ne’. It is also kept if the size of the dependent’s subtree is  $\leq 3$  – this heuristic choice introduces some variety to the sentences, but prevents the NPs from growing in length.
- *obl* – In general *obl* relations are likely to introduce large sub-trees, which should be avoided. On the other hand, Russian, Finnish and Estonian do not make a distinction between the oblique arguments of verbs, required for grammaticality, and adjuncts which can be dropped. Given the lack of required annotation, I make

a heuristic choice to keep obl relations if the dependent’s subtree is  $\leq 3$ , based on an observation that obl linking to arguments are typically associated with small sub-trees.

- mark – Kept if the head is not the (new) root verb.

**Rejected:** advcl, cop, discourse, dislocated, vocative, expl – those relations either introduce content that is unlikely to increase acceptability (e.g., vocative or discourse) or are likely to introduce prominent sub-trees (e.g., advcl).

### E.1.3 Nominal relations

*acl, amod, appos, case, clf, det, nmod, nummod*

**Accepted:**

- case – Required for grammaticality.
- det – Including it increases acceptability (and grammaticality, as in Polish ‘więcej problemów’).
- nmod:arg – In Polish nmod:arg is used to mark nominal dependents required by nouns.
- nmod:gobj, nmod:gsubj, nmod:poss – Finnish relations used for objects and subjects of noun participles and genitive modifiers (nmod:poss) – all are required for grammaticality and acceptability.
- nummod – Added to increase acceptability, as such modifiers often govern the case in the considered languages.

**Conditionally accepted:**

- amod – Kept if the dependent’s subtree is of size 1. This is done to add substance to sentences, but avoid more complex sub-trees that would decrease grammaticality post-reordering.
- nmod – In general nmod relations are likely to introduce large sub-trees, so they are kept only if the size of the sub-tree is  $\leq 3$ .

**Rejected:** acl, appos, clf (not relevant for the examined languages)

### E.1.4 Linking relations

*cc, conj, list, parataxis*

**Accepted:** *conj* – Accepted only for the subject argument of the verb; required for grammaticality, since it maintains the subject-verb agreement.

**Conditionally accepted:** *cc* – Kept if the head is not the new root verb.

**Rejected:** *list, parataxis*

### E.1.5 MWE relations

*compound, fixed, flat*

All are accepted, since they are unlikely to introduce large sub-trees.

### E.1.6 Special relations

*dep, goeswith, orphan, punct, reparandum, root*

**Accepted:** *dep, root*

**Conditionally accepted:** *punct* – Kept if the head is linked to its head through the *conj* relation or the dependent is a quotation mark; this is done to maintain acceptability.

**Rejected:** *goeswith, orphan, reparandum*

# LEXICAL CATEGORISATION OF UD DEPENDENCY RELATIONS

---

In this appendix I list the rationale behind the categorisation of relations described in Chapter 5. For each relation, I provide its description based on <https://universaldependencies.org/u/dep/> and describe my assessment of whether a model’s performance on that relation is likely to be affected by treebank alterations that change noun lexemes.

## F.1 Categories of relations (based on LAS)

1. **Not at all (concerned mostly with verbs or trivial):**  
punct, root, aux, aux:clitic, aux:end, aux:pass, aux:imp, advmod, advmod:neg, advmod:arg, compound:prt, list, csubj, ccomp, xcomp, expl:pv, nmod:pred
2. **Should not be affected (but concerned with nouns):**  
case, det, det:poss, det:numgov, det:nummod, cop, mark, advmod:emph, amod, discourse, discourse:emo, discourse:intj, conj, cc, cc:preconj, parataxis, parataxis:obj, parataxis:insert, obl:arg, fixed, nummod, nsubj:cop, iobj, obj, nsubj
3. **Likely to be affected:**  
orphan, amod:flat, flat, flat:name, flat:foreign, compound, compound:nn, obl, nmod, nmod:arg, nmod:flat, nmod:poss, nmod:gobj, nmod:gsubj, appos

## F.2 Rationale

1. **punct**

*Used for any piece of punctuation in a clause, tokens with the relation punct always attach to content words.*

**LAS category: 1, UAS category: 1**

This relation is determined by POS and the structure of the sentence (word order in particular). In the training split of the corpus it is attached to a lot of different content words, with various meanings, so changing lexical information should not affect its accuracy. The label is also trivial.

**2. root**

*The root grammatical relation points to the root of the sentence. If the main predicate is not present (due to ellipsis) and there are multiple orphaned dependents, one of these is promoted to the head (root) position and the other orphans are attached to it.*

**LAS category: 1, UAS category: 1**

Most of the time the root is a verb, not changed during noun lexeme mixing. In those cases a model should still identify the root correctly. In situations where the main predicate is missing, the root can become a noun, but in those cases word order is likely to be the main signal, rather than the lexical information.

**3. case**

*It is used for any case-marking element which is treated as a separate syntactic word (including prepositions, postpositions, and clitic case markers). Case-marking elements are treated as dependents of the noun they attach to or introduce.*

**LAS category: 2, UAS category: 2**

Each preposition will typically modify only a specific subset of nouns of particular lexical category/class. As an example, Polish preposition ‘w’ can attach to: plural nouns in accusative (to describe a pattern), days of the week in accusative, locations in locative, years in locative, and various other objects in locative (meaning inside or ‘wearing’).<sup>1</sup>

The model could rely on such lexical cues to determine the head for an adposition. However, as a native speaker I have no problem identifying the correct heads of prepositions in the Polish mixed data, given the case marking and the fact that the head is usually located very close to the preposition. This is even in cases where the

---

<sup>1</sup>There are also many fixed phrases with the preposition ‘w’, but these are most likely to be linked with UD relation *fixed*.

meaning of the resulting prepositional phrase (after mixing) is not clear; e.g., the combination of the ‘w’ preposition and the case + lexeme does not cleanly match any of the interpretations outlined above.

Consequently, I believe the performance on this relation should not be affected *if the model learns what case is governed by the specific preposition*. Further, there should be no difference in UAS and LAS performance, given that there is a closed set of words that function as prepositions within a language.

#### 4. **det**

*The relation determiner holds between a nominal head and its determiner.*

**LAS category: 2, UAS category: 2**

The agreement with the head and its proximity are clear signals to the correct attachment. The label assignment should also be trivial (unless there are some inconsistencies in the training data). Generally, there is no lexical relationship between the set of determiners and the set of words they modify (apart from POS).

#### 5. **det:poss**

*Used for possessive determiners.*

**LAS category: 2, UAS category: 2**

See *det*.

#### 6. **det:numgov**

*Pronominal quantifiers in Slavic languages are labeled det:numgov instead of det because they normally do not agree with the quantified noun in case (unlike non-quantifying determiners). The quantifier requires the counted noun to be in its genitive form.*

**LAS category: 2, UAS category: 2**

The quantifier governs a specific case and its head usually follows it immediately. This makes it straightforward to identify.

#### 7. **det:nummod**

*Used for pronominal quantifiers in situations where the whole phrase (quantifier + noun) fills a role where genitive, dative, locative or instrumental noun phrases are expected.*

**LAS category: 2, UAS category: 2** (see *det*)

## 8. aux

*An aux (auxiliary) of a clause is a function word associated with a verbal predicate that expresses categories such as tense, mood, aspect, voice or evidentiality. It is often a verb (which may have non-auxiliary uses as well) but many languages have nonverbal TAME markers and these are also treated as instances of aux.*

**LAS category: 1, UAS category: 1**

The verbs do not change during mixing and the relation is independent of the arguments that a verb takes.

## 9. aux:clitic

*The aux:clitic relation is used in the Polish PDB-UD and PUD treebanks for mobile inflections (e.g., *piliśmy* is split into *pili* + *śmy* and *śmy* is attached as aux:clitic).*

**LAS category: 1, UAS category: 1**

This relation is trivial – a unique set of inflections attach to a word directly preceding them.

## 10. aux:cnd

*The aux:cnd relation is used in the Polish PDB-UD and PUD treebanks for the conditional particle ‘by’ (e.g. ‘mógłby’, ‘moglibyśmy’).*

**LAS category: 1, UAS category: 1** (see *aux*)

## 11. aux:pass

*A passive auxiliary of a clause is a non-main verb of the clause which contains the passive information.*

**LAS category: 1, UAS category: 1** (see *aux*)

## 12. aux:imp

*The aux:imp relation in the Polish PDB-UD treebank for the imperative particle ‘niech’ and its variants ‘niechaj’, ‘niechże’, ‘niechby’ (let).*

**LAS category: 1, UAS category: 1** (see *aux*)

## 13. cop

*A cop (copula) is the relation of a function word used to link a subject to a nonverbal predicate.*

**LAS category: 2, UAS category: 2**

The position of the head with respect to the copula and the case + agreement marking should be sufficient signals to the attachment. Plus, copulas attach to a variety of different words (a big lexical spread). Assignment of a label is trivial, given the attachment.

**14. mark**

*A marker is the word marking a clause as subordinate to another clause (in English e.g., ‘that’, ‘whether’, ‘although’, in Polish ‘że’, ‘chociaż’, ‘żeby’, ‘bo’, ‘zanim’). The marker is a dependent of the subordinate clause head.*

**LAS category: 2, UAS category: 2**

Most of the markers attach to verbal heads of the clause. For those that attach to nouns, those nouns can be identified via morphosyntactic cues.

**15. advmod**

*An adverbial modifier of a word is a (non-clausal) adverb or adverbial phrase that serves to modify a predicate or a modifier word.*

**LAS category: 1, UAS category: 1**

This relation is concerned with modifiers of predicates/adjectives/adverbs and not nouns.

**16. advmod:neg**

*The advmod:neg relation is used in the Polish PDB-UD and PUD treebanks for adverbials realised by the negative particle nie (“not”).*

**LAS category: 1, UAS category: 1 (see advmod)**

**17. advmod:emph**

*Used for emphasizers. While other adverbial modifiers usually modify verbs, adjectives or adverbs, these emphasizers often modify noun phrases, including prepositional phrases.*

**LAS category: 2, UAS category: 2**

Although some of the emphasizers can attach to nouns, there is no lexical regularity regarding the types of nouns to which they attach. The attachment should be clear based on morphosyntactic clues.

**18. advcl**

*An adverbial clause modifier is a clause which modifies a verb or other predicate (adjective, etc.), as a modifier not as a core complement. This includes things such as a temporal clause, consequence, conditional clause, purpose clause, etc. The dependent must be clausal (or else it is an advmod) and the dependent is the main predicate of the clause.*

**LAS category: 1, UAS category: 1**

The relation concerns mostly verbs.

**19. acl**

*acl stands for finite and non-finite clauses that modify a nominal. The acl relation contrasts with the advcl relation, which is used for adverbial clauses that modify a predicate. The head of the acl relation is the noun that is modified, and the dependent is the head of the clause that modifies the noun. Some languages allow finite clausal complements for nouns with a subset of nouns like fact or report. These look roughly like relative clauses, but do not have any omitted role in the dependent clause. This is the class of ‘content clauses’ in Huddleston and Pullum 2002). These are also analyzed as acl.*

**LAS category:2 , UAS category: 2**

While the relation is concerned with nouns, it is signaled primarily through morphosyntactic means, so in most cases it should not be affected by noun lexeme mixing. On the other hand, mixing the nouns ‘fact’, ‘report’ etc., can be problematic. Assuming that this problem will be infrequent, I assigned this relation to class 2.

**20. acl:relcl**

*A relative clause modifier of a nominal is a clause that modifies the nominal, whereas the nominal is coreferential with a constituent inside the relative clause (here the constituent may be realised as a relative pronoun, another relative word, or it may not be overtly realised at all).*

**LAS category: 2, UAS category: 2**

**21. advmod:arg**

*The advmod:arg relation is used in the Polish PDB-UD and PUD treebanks for adverbial complements that are required to complete the meaning of a verb. The adverbial complements are realised as adverbs or adverbial phrases.*

**LAS category: 1, UAS category: 1**

This relation only concerns adverbs and verbs.

**22. amod**

*An adjectival modifier of a noun (or pronoun) is any adjectival phrase that serves to modify the meaning of the noun (or pronoun).*

**LAS category: 2, UAS category: 2**

The agreement with the noun and the position of the adjective with respect to the noun clearly identifies the head. The performance would only drop if a model was relying on ‘typical features’ that nouns are associated with (which would be incorrect).

**23. amod:flat**

*The amod:flat relation is used in the Polish PDB-UD and PUD treebanks for adjectival parts (ADJ) of named entities, e.g., organisation names with clear syntactic structure, dates. These adjectival parts of named entities are not proper names (PROPN). Their combinations with other words (PROPN and non-PROPN) build named entities.*

**LAS category: 3, UAS category: 2**

The UAS should not be affected, but the recognition of the relation as *amod:flat*, instead of *amod* could become more challenging. Although, in many cases where this relation is used, both the head and the dependent are capitalised, which is a cue for the *amod:flat* label.

**24. orphan**

*The ‘orphan’ relation is used in cases of head ellipsis where simple promotion would result in an unnatural and misleading dependency relation. The typical case is predicate ellipsis where one of the core arguments has to be promoted to clausal head.*

**LAS category: 3, UAS category: 3**

The lexical relationship/similarity between the core argument promoted to the head, the dependent and the other predicates in a sentence is likely helpful for correct parsing of ellipsis instances.

**25. discourse**

*This is used for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way). We generally follow the guidelines of what the Penn Treebanks count as an INTJ. They define this to include: interjections ('oh', 'uh-huh', 'Welcome'), fillers ('um', 'ah'), and non-adverbial discourse markers ('well', 'like', but not 'you know' or 'actually'). These discourse elements are attached to the head of the most relevant nearby clause, which is why they are grouped with non-core clausal dependents even though they are normally not dependents of the predicates as such.*

**LAS category: 2, UAS category: 2**

The relation concerns only a specific subset of words, so the only potentially challenging task is to identify the most suitable head. This should not be any harder after mixing noun lexemes.

**26. discourse:emo**

*Used for emoticons/emojis.*

**LAS category: 2, UAS category: 2** (see *discourse*)

**27. discourse:intj**

*The discourse:intj relation in the Polish PDB-UD and PUD treebanks for exclamations (e.g. 'ach!'), response particles (e.g., 'no'), greetings (e.g., 'serwus'), curses (e.g., 'kurwa'), etc.*

**LAS category: 2, UAS category: 2** (see *discourse*)

**28. conj**

*A conjunct is the relation between two elements connected by a coordinating conjunction, such as 'and', 'or', etc. We treat conjunctions asymmetrically: The head of the relation is the first conjunct and all the other conjuncts depend on it via the conj relation.*

**LAS category: 2, UAS category: 2**

It is a very general relation and can link all sorts of lexemes. It is also signaled through specific words/punctuation. It is possible that in some situations incorrect handling of a multi-word expression could result in conj being attached incorrectly, e.g., “The (futures) and (exchange markets)” vs “The (stock and exchange) markets”, but such instances are likely to be uncommon.

29. **cc**

*A cc is the relation between a conjunct and a preceding coordinating conjunction. A coordinating conjunction may also appear at the beginning of a sentence. This is also called a cc, even though there is no preceding conjunct (except implicitly or in a preceding sentence).*

**LAS category: 2, UAS category: 2**

The coordinating conjunctions attach to all sorts of lexemes so there is no reason for the lexical information (beyond POS) to be involved in this attachment.

30. **cc:preconj**

*A preconjunct is the relation between the head of an NP and a word that appears at the beginning bracketing a conjunction (and puts emphasis on it), such as ‘either’, ‘both’, ‘neither’.*

**LAS category: 2, UAS category: 2 (see cc)**

31. **fixed**

*The fixed relation is one of the three relations for multi-word expressions (MWEs) (the other two being flat and compound). It is used for certain fixed grammaticised expressions that behave like function words or short adverbials.*

**LAS category: 2, UAS category: 2**

I do not mix nouns that are linked through a *fixed* relation.

32. **flat**

*The flat relation is used for exocentric (headless) semi-fixed MWEs like names (Hillary Rodham Clinton) and dates (24 December). It contrasts with fixed, which applies to completely fixed grammaticised (function word-like) MWEs (like ‘in spite of’), and with compound, which applies to endocentric (headed) MWEs (like ‘apple pie’).*

**LAS category: 3, UAS category: 3**

When parts of an MWE get replaced it might be difficult to recognise the phrase as an MWE.

**33. flat:name**

*The flat:name relation is a specialization of flat used for names.*

**LAS category: 3, UAS category: 3** (see *flat*).

**34. flat:foreign**

*We use flat:foreign to label sequences of foreign words.*

**LAS category: 3, UAS category: 3** (see *flat*).

**35. compound**

*The compound relation is one of three relations for multi-word expressions (MWEs) (the other two being fixed and flat). It is used for any kind of  $X^0$  compounding (noun compounds (e.g., phone book), but also verb and adjective compounds that are more common in other languages), for particle verbs and for serial verbs.*

**LAS category: 3, UAS category: 3** (see *flat*).

**36. compound:nn**

*The dependency type compound:nn, which stands for noun compound modifier has two basic uses in the Finnish UD scheme. First, In Finnish, compounds are generally written as a single word, but for instance some compounds involving foreign words or proper names are written separately using a dash. These are annotated using compound:nn. The second use of the type compound:nn is to mark appellation modifiers, which are modifying, non-inflecting noun phrases that generally express profession, rank, position, assignment or other such classifiable property.*

**LAS category: 3, UAS category: 3** (see *compound*).

**37. compound:prt**

*The phrasal verb particle relation identifies an idiomatic phrasal verb, and holds between the verb and its particle (tagged as ADP). It is a subtype of the compound relation.*

**LAS category: 1, UAS category: 1**

The relation concerns exclusively verbs and phrasal verb particles.

**38. list**

*The list relation is used for chains of comparable items. In lists with more than two items, all items of the list should modify the first one. Informal and web text often contains passages which are meant to be interpreted as lists but are parsed as single sentences.*

**LAS category: 1, UAS category: 1**

This relation can be identified based on the sentence structure.

**39. parataxis**

*The parataxis relation (from Greek for “place side by side”) is a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a “:” or a “;”, placed side by side without any explicit coordination, subordination, or argument relation with the head word. Parataxis is a discourse-like equivalent of coordination, and so usually obeys an iconic ordering. Hence it is normal for the first part of a sentence to be the head and the second part to be the parataxis dependent, regardless of the headedness properties of the language. But things do get more complicated, such as cases of parentheticals, which appear medially.*

**LAS category: 2, UAS category: 2**

The relation usually links two verbs and in these cases it should not be affected by lexeme mixing. In some languages, including Russian it is also used to connect parts of the sentence written as explanation in brackets. Here, LAS could be affected by the mixing, but the assumption is that such cases are relatively rare.

**40. parataxis:obj**

*The specialization is used in for paratactic sentences with an implicit object role with respect to the governing predicate. The relation is often used with direct speech when governed by a verb.*

**LAS category: 2, UAS category: 2 (see parataxis)**

**41. parataxis:insert**

*The parataxis:insert relation is used in the Polish PDB-UD and PUD treebanks for explanatory or commenting word, clause or sentence inserted into a sentence. A paranthesis is usually marked with punctuation marks (e.g. commas, brackets, or dashes), but sometimes it is not marked.*

**LAS category: 2, UAS category: 2** (see *parataxis*)

#### 42. **csubj**

*A clausal subject is a clausal syntactic subject of a clause, i.e., the subject is itself a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb.*

**LAS category: 1, UAS category: 1**

In most cases the relation does not involve a noun and should not be affected by noun lexeme mixing.

#### 43. **csubj:cop**

*A clausal copular subject (csubj:cop) is a clause that acts as the subject of another, copular clause. As in all copular clauses, the predicative acts as the head of the clause and hence it is also the governor of the copular subject.*

**LAS category: 1, UAS category: 1**

The relation is primary concerned with non-nouns.

#### 44. **ccomp**

*A clausal complement of a verb or adjective is a dependent clause which is a core argument. That is, it functions like an object of the verb, or adjective.*

*It is also used in reported speech. With a speech verb like ‘say’, the content of reported speech is considered to be part of the verb’s valency. It therefore attaches as ccomp—not only when integrated within the clause as an indirect quotation (‘said that...’), but also when set off as a direct quotation, even with inverted order. Quoted content is considered to be ccomp even if it is a sentence fragment.*

**LAS category: 1, UAS category: 1**

While the relation can at times involve a noun, e.g., in quoted content, I expect these cases to be rare.

#### 45. xcomp

*An open clausal complement (xcomp) of a verb or an adjective is a predicative or clausal complement without its own subject. The reference of the subject is necessarily determined by an argument external to the xcomp (normally by the object of the next higher clause, if there is one, or else by the subject of the next higher clause). The clausal complement can be headed by various parts of speech, including a VERB, ADJ, or NOUN. The xcomp-taking predicate of the higher clause can be a VERB or ADJ.*

**LAS category: 1, UAS category: 1** (see *ccomp*)

#### 46. xcomp:ds

*The dependency type xcomp:ds, which stands for clausal complement with different subject, is a subtype of xcomp (open clausal complement). It is used for clausal complements where the subject of the complement clause is inherited from the higher clause, but it's not a subject of the governing clause (if it is a subject, xcomp is used instead).*

**LAS category: 1, UAS category: 1** (see *ccomp*)

#### 47. xcomp:pred

*The xcomp:pred relation is used in the Polish PDB-UD and PUD treebanks for predicative adjectives and nominals depending on verbs, e.g. *czuć się* ('to feel'), *zostać* ('to become'), *wydawać się* ('to seem to be'), which are not considered copular verbs.*

**LAS category: 2 , UAS category: 2**

This relation involves nouns, but it is signalled via morphosyntactic means and should not be affected. The label can be identified based on the verb.

#### 48. nsubj

*A nominal subject (nsubj) is a nominal which is the syntactic subject and the proto-agent of a clause. That is, it is in the position that passes typical grammatical test for subjecthood, and this argument is the more agentive, the do-er, or the proto-agent of the clause. This nominal may be headed by a noun, or it may be a pronoun or relative pronoun or, in ellipsis contexts, other things such as an adjective.*

**LAS category: 2, UAS category: 2**

Case and agreement marking provide a strong signal for subjecthood.

**49. nsubj:cop**

*The dependency type nsubj:cop is used for the nominal subject of a copular clause. The predicative is the head of the copular clause, and also the governor of the nsubj:cop dependency.*

**LAS category: 2, UAS category: 2** (see *nsubj*)

**50. nsubj:pass**

*A passive nominal subject is a noun phrase which is the syntactic subject of a passive clause.*

**LAS category:2 , UAS category:2** (see *nsubj*)

**51. obj**

*The object of a verb is the second most core argument of a verb after the subject. Typically, it is the noun phrase that denotes the entity acted upon or which undergoes a change of state or motion (the proto-patient). In general, if there is just one object, it should be labeled obj, regardless of the morphological case or semantic role that it bears. If there are two or more objects, one of them should be obj and the others should be iobj.*

**LAS category: 2, UAS category: 2**

In most cases the object is clearly indicated by morphosyntax.

**52. iobj**

*The indirect object of a verb is any nominal phrase that is a core argument of the verb but is not its subject or (direct) object. The prototypical example is the recipient of ditransitive verbs of exchange.*

**LAS category: 2, UAS category: 2** (see *obj*)

**53. obl**

*The obl relation is used for a nominal (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct. This means that it functionally corresponds to an adverbial attaching to a verb, adjective or other adverb. The obl relation can be further specified by the case.*

**LAS category: 3, UAS category: 3**

Without lexical information, it could be unclear what the adjunct/oblique argument should attach to (e.g., it could be recognised as *nmod* and attached to a noun). Further, in the mixed lexeme case *obl* could be more likely confused with *obl:arg*.

#### 54. **obl:arg**

*(Polish specific) The relation obl:arg is used for oblique arguments and distinguishes them from adjuncts, which use the plain obl relation. The relation obl:arg is a language-specific subtype (as opposed to universal type) because the argument-adjunct distinction is not made at the universal level (some discussion is here). Arguments are selected by the predicate. Their coding (preposition and morphological case) is determined by the predicate; within the set of arguments of this predicate, the coding maps the argument to a particular semantic role. In contrast, the semantics of an adjunct is relatively independent of the predicate, and typical adjuncts (such as specifications of time, location, manner or instrument) can combine with a large number of different predicates.*

**LAS category: 2, UAS category: 2**

Given that oblique arguments are selected by predicates and signaled through inflectional morphology, lexical noun mixing should not affect performance on this relation.

#### 55. **obl:cmpr**

*The obl:cmpr relation is used in the Polish PDB-UD and PUD treebanks for comparative expressions realised as phrases. Comparative expressions are preceded by a comparative conjunction either *niż* ('than') in comparatives of inequality or e.g. *jak* ('like') in comparatives of equality.*

**LAS category: 1 , UAS category: 1**

The relation is signaled through a comparative conjunction and typically attaches to the directly preceding verb. In the training data it appears with a wide lexical variety of dependents.

## 56. **obl:agent**

*The relation **obl:agent** is used for agents of passive verbs. In Polish PDB it mostly follows the preposition ‘przez’, in contexts where it means ‘by’ – e.g., ‘wysłany przez nich’ (‘sent by them’).*

### **LAS category: 3, UAS category: 2**

The preposition ‘przez’ has many meanings. Changing the lexeme of its head when that head is labeled as *obl:agent* can affect which of the meanings best fits the context, making the *obl:agent* label less likely. The attachment, however, should be relatively straightforward.

## 57. **expl:pv**

*(Polish specific) Reflexive clitic of an inherently reflexive verb. Reflexive pronouns usually replace objects of verbs. However, some verbs are inherently reflexive, i.e., the verb always occurs with a reflexive pronoun, and the pronoun cannot be replaced by a non-reflexive pronoun. With these verbs, the reflexive pronoun is attached as **expl:pv** instead of **obj**.*

### **LAS category: 1, UAS category: 1**

In Polish there is only one reflexive clitic ‘się’, and it will trivially attach to the first preceding verb.

## 58. **nmod**

*The **nmod** relation is used for nominal dependents of another noun or noun phrase and functionally corresponds to an attribute, or genitive complement.*

### **LAS category: 3, UAS category: 3**

The attachment of the prepositional phrase could be ambiguous without the lexical information; e.g., *nmod* could be confused with the *obl* relation and attach to a verb, rather than a noun.

## 59. **nmod:arg**

*The **nmod:arg** relation is used in the Polish PDB-UD and PUD treebanks for nominal dependents required by nouns. The governing nouns are often derived from verbs, but they are not gerunds, e.g., ‘zmiana’ (a change) is derived from ‘zmieniać’ (to change).*

**LAS category: 3, UAS category: 2**

In Polish this relation is mostly used for noun + noun (GEN) instances, with the nouns placed right next to one another. The attachment of the noun in genitive is unlikely to be affected, but the label is likely to be confused, e.g., with *nmod* or *appos*. I expect the drop on LAS to be substantial given that only some nouns can take nominal arguments.

**60. nmod:flat**

*The nmod:flat relation is used in the Polish PDB-UD and PUD treebanks for nominal parts (NOUN) of named entities, e.g., organisation names with clear syntactic structure. These nominal parts of named entities are not proper names (PROPN). Their combinations with other words (PROPN and non-PROPN) build named entities.*

**LAS category: 3, UAS category: 2** (see *amod:flat*)**61. nmod:poss**

*The dependency type poss stands for possessive in the original SD scheme, but in UD Finnish, the corresponding type nmod:poss is used for genitive modifiers in general, which in Finnish often but not nearly always imply possession. There are two kinds of genitive modifiers that are not annotated using the general genitive modifier type: the genitive object, nmod:gobj and the genitive subject, nmod:gsubj*

*In Polish this relation seems to be reserved for ‘zdaniem X’ (according to X).*

**LAS category: 3, UAS category: 2** (see *amod:flat*)

In Polish UD, LAS performance will probably drop considerably, because lexeme ‘zdaniem’ will be replaced by another lexeme. UAS will likely be less affected. In Finnish, the LAS performance is less likely to drop, but the relation could be confused with e.g., *nmod:gobj* and *nmod:gsubj*.

**62. nmod:gobj**

*(Finnish specific) Certain nouns, those which have been directly derived from a verb or otherwise have a verb counterpart, can take an object in Finnish. These objects closely resemble more general genitive modifiers nmod:poss.*

**LAS category: 3, UAS category: 2** (see *nmod:poss*)

63. **nmod:gsubj**

*(Finnish specific) Genitive subjects are subject-like arguments taken by a noun. This is in parallel to genitive objects nmod:gobj.*

**LAS category: 3, UAS category: 2** (see *nmod:poss*)

64. **nmod:pred**

*The nmod:pred relation is used in the Polish PDB and PUD treebanks for predicative expressions depending on the gerund form of the copula być (to be).*

**LAS category: 1, UAS category: 1**

In Polish PDB this relation always applies to phrases ‘być X (INS)’ (to be X). The gerunds are not mixed and the position with respect to the copula, combined with the instrumental case are very strong signals for this relation.

65. **appos**

*An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun. It includes parenthesised examples, as well as defining abbreviations in one of these structures.*

**LAS category: 3, UAS category: 2**

It should be affected to the same extent as *nmod(s)*. The relation might be hard to distinguish from *nmod(s)* after noun lexeme mixing.

66. **nummod**

*A numeric modifier of a noun is any number phrase that serves to modify the meaning of the noun with a quantity.*

**LAS category: 2, UAS category: 2**

The relation is used with a variety of nouns and the lexical content should not have an effect on its recognition.

---

SUPPLEMENTARY RESULTS

---

| #                                    | Polish |       |     |      | Russian |       |     |      | Finnish |       |     | Estonian |       |     |
|--------------------------------------|--------|-------|-----|------|---------|-------|-----|------|---------|-------|-----|----------|-------|-----|
|                                      | all    | nsubj | obj | iobj | all     | nsubj | obj | iobj | all     | nsubj | obj | all      | nsubj | obj |
|                                      | 28.9K  | 365   | 293 | 73   | 95.7K   | 818   | 762 | 56   | 15.6K   | 148   | 148 | 37.0K    | 324   | 324 |
| LAS, Dozat and Manning (2017) + BERT |        |       |     |      |         |       |     |      |         |       |     |          |       |     |
| M                                    | 0.1    | 0.5   | 0.8 | 5.0  | 0.1     | 0.2   | 0.3 | 0.9  | 0.2     | 0.4   | 1.0 | 2.9      | 1.2   | 2.3 |
| L                                    | 0.1    | 1.1   | 1.8 | 4.0  | 0.1     | 0.5   | 1.4 | 3.2  | 0.1     | 1.5   | 0.5 | 0.1      | 1.8   | 1.3 |
| D                                    | 0.2    | 0.7   | 1.0 | 2.1  |         |       |     |      |         |       |     |          |       |     |
| UAS, Dozat and Manning (2017) + BERT |        |       |     |      |         |       |     |      |         |       |     |          |       |     |
| M                                    | 0.1    | 0.4   | 0.4 | 0.0  | 0.1     | 0.2   | 0.0 | 0.0  | 0.1     | 0.4   | 0.7 | 2.3      | 1.2   | 0.8 |
| L                                    | 0.1    | 1.0   | 0.7 | 1.1  | 0.1     | 0.4   | 0.2 | 0.8  | 0.1     | 1.5   | 0.9 | 0.1      | 0.7   | 0.6 |
| D                                    | 0.1    | 0.2   | 0.4 | 1.4  |         |       |     |      |         |       |     |          |       |     |

**Table G.1:** Complement to Chapter 3. Standard deviation scores for LAS/UAS results for morphologically aware BERT-based model (M) and BERT-based baselines trained and tested on lemmatised language (L) and decased language (D).

|                                     | Polish |       |     |      | Russian |       |     |      | Finnish |       |     | Estonian |       |     |
|-------------------------------------|--------|-------|-----|------|---------|-------|-----|------|---------|-------|-----|----------|-------|-----|
|                                     | all    | nsubj | obj | iobj | all     | nsubj | obj | iobj | all     | nsubj | obj | all      | nsubj | obj |
|                                     |        |       |     |      |         |       |     |      |         |       |     |          |       |     |
| LAS, Dozat and Manning (2017) + CNN |        |       |     |      |         |       |     |      |         |       |     |          |       |     |
| M                                   | 0.4    | 1.1   | 1.7 | 3.6  | 0.1     | 1.3   | 1.8 | 7.2  | 0.5     | 1.6   | 1.8 | 0.3      | 1.8   | 3.5 |
| L                                   | 0.6    | 2.2   | 3.1 | 4.5  | 10.5    | 5.0   | 7.9 | 30.7 | 0.6     | 5.0   | 4.7 | 0.6      | 3.5   | 2.6 |
| D                                   | 0.5    | 0.7   | 2.0 | 5.2  |         |       |     |      |         |       |     |          |       |     |
| UAS, Dozat and Manning (2017) + CNN |        |       |     |      |         |       |     |      |         |       |     |          |       |     |
| M                                   | 0.2    | 0.1   | 1.5 | 1.6  | 0.1     | 0.5   | 0.3 | 0.8  | 0.4     | 0.8   | 2.3 | 0.1      | 1.0   | 1.0 |
| L                                   | 0.3    | 1.5   | 0.8 | 1.7  | 7.8     | 6.7   | 4.6 | 4.7  | 0.4     | 0.4   | 3.9 | 0.4      | 1.4   | 1.8 |
| D                                   | 0.4    | 1.7   | 0.4 | 2.9  |         |       |     |      |         |       |     |          |       |     |

**Table G.2:** Complement to Chapter 3. Standard deviation scores for LAS/UAS results for morphologically aware CNN-based model (M) and CNN-based baselines trained and tested on lemmatised language (L) and decased language (D).

|              | Polish<br>383 clauses |       |      |      | Russian<br>749 clauses |       |      |      | Finnish<br>227 clauses |       |      | Estonian<br>283 clauses |       |      |
|--------------|-----------------------|-------|------|------|------------------------|-------|------|------|------------------------|-------|------|-------------------------|-------|------|
|              | all                   | nsubj | obj  | iobj | all                    | nsubj | obj  | iobj | all                    | nsubj | obj  | all                     | nsubj | obj  |
| DM +BERT     |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO          | 88.8                  | 99.7  | 98.1 | 35.9 | 95.6                   | 98.3  | 94.6 | 91.9 | 94.2                   | 99.6  | 99.6 | 94.3                    | 96.5  | 95.4 |
| OVS          | 86.6                  | 96.6  | 93.4 | 20.3 | 94.5                   | 92.8  | 90.1 | 85.5 | 93.6                   | 98.2  | 98.7 | 92.8                    | 95.1  | 92.6 |
| SOV          | 86.3                  | 97.4  | 89.7 | 18.8 | 94.1                   | 95.6  | 90.4 | 80.6 | 89.0                   | 95.2  | 93.4 | 89.0                    | 87.6  | 86.2 |
| OSV          | 85.9                  | 94.3  | 91.9 | 18.8 | 94.1                   | 93.6  | 92.3 | 87.1 | 90.4                   | 99.1  | 93.8 | 87.1                    | 91.2  | 76.0 |
| VSO          | 79.9                  | 77.5  | 76.6 | 17.2 | 90.8                   | 82.2  | 80.8 | 79.0 | 88.4                   | 86.8  | 96.5 | 91.7                    | 89.4  | 91.2 |
| VOS          | 84.4                  | 84.9  | 94.7 | 25.0 | 93.8                   | 89.9  | 93.5 | 88.7 | 84.7                   | 72.2  | 86.8 | 88.4                    | 84.8  | 83.0 |
| DM +CNN      |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO          | 82.1                  | 96.1  | 90.3 | 26.6 | 90.7                   | 89.6  | 91.0 | 67.7 | 81.2                   | 94.3  | 89.4 | 84.8                    | 88.0  | 79.2 |
| OVS          | 67.1                  | 56.1  | 44.4 | 20.3 | 80.5                   | 57.9  | 62.6 | 61.3 | 65.9                   | 49.8  | 55.5 | 78.3                    | 72.1  | 60.1 |
| SOV          | 69.4                  | 86.7  | 42.5 | 6.2  | 80.6                   | 75.6  | 56.0 | 46.8 | 62.0                   | 57.3  | 36.1 | 67.7                    | 49.8  | 42.8 |
| OSV          | 63.4                  | 54.6  | 37.2 | 20.3 | 78.2                   | 46.9  | 63.4 | 61.3 | 67.0                   | 85.9  | 33.0 | 68.1                    | 68.2  | 21.2 |
| VSO          | 63.9                  | 50.1  | 42.8 | 14.1 | 74.8                   | 45.0  | 47.6 | 41.9 | 61.7                   | 41.9  | 66.5 | 72.7                    | 50.2  | 60.8 |
| VOS          | 64.2                  | 25.8  | 78.8 | 23.4 | 77.9                   | 31.9  | 82.0 | 66.1 | 55.9                   | 33.0  | 43.2 | 72.9                    | 56.5  | 47.7 |
| DM +FASTTEXT |                       |       |      |      |                        |       |      |      |                        |       |      |                         |       |      |
| SVO          | 82.1                  | 91.1  | 93.4 | 39.1 | 91.4                   | 93.1  | 90.6 | 80.6 | 80.1                   | 88.5  | 87.2 | 81.8                    | 85.9  | 87.3 |
| OVS          | 68.3                  | 58.5  | 56.6 | 10.9 | 81.1                   | 69.2  | 55.3 | 46.8 | 59.4                   | 50.2  | 32.6 | 69.8                    | 62.2  | 51.6 |
| SOV          | 72.8                  | 72.8  | 68.1 | 14.1 | 83.4                   | 73.4  | 68.2 | 54.8 | 64.8                   | 55.5  | 59.9 | 70.9                    | 55.5  | 71.0 |
| OSV          | 64.9                  | 68.4  | 44.7 | 7.8  | 80.0                   | 70.6  | 56.7 | 30.6 | 64.2                   | 85.0  | 23.8 | 66.6                    | 72.1  | 32.5 |
| VSO          | 61.0                  | 49.6  | 62.8 | 15.6 | 62.6                   | 37.1  | 41.5 | 40.3 | 45.2                   | 35.7  | 47.6 | 49.8                    | 26.5  | 64.0 |
| VOS          | 69.3                  | 46.7  | 88.4 | 21.9 | 72.1                   | 41.1  | 62.1 | 51.6 | 42.1                   | 24.2  | 29.1 | 53.0                    | 29.0  | 47.7 |

**Table G.3:** Complement to Chapter 4. Mixed-treebank experiment (see Section 4.3). Reported numbers are LAS scores on the dev splits of the *unambiguous* counterfactual treebanks (concatenation of all available cf. treebanks for a language). All results are for models trained with a random seed set to 1.

|                                              |     | Polish |       |      |       | Russian |       |      |       | Finnish |       |       | Estonian |       |       |
|----------------------------------------------|-----|--------|-------|------|-------|---------|-------|------|-------|---------|-------|-------|----------|-------|-------|
|                                              |     | all    | nsubj | obj  | iobj  | all     | nsubj | obj  | iobj  | all     | nsubj | obj   | all      | nsubj | obj   |
| svo attachment bias metric (LAS)             |     |        |       |      |       |         |       |      |       |         |       |       |          |       |       |
| BERT                                         | UA. | 4.2    | 9.6   | 8.8  | 15.9  | 2.1     | 7.5   | 5.2  | 7.7   | 5.0     | 9.3   | 5.8   | 4.5      | 6.9   | 9.6   |
|                                              | A.  | 5.0    | 16.4  | 18.4 | 20.0  | 4.0     | 27.9  | 22.6 | 20.0  | 3.0     | 22.5  | 15.0  | 4.4      | 19.2  | 21.1  |
| CNN                                          | UA. | 16.5   | 41.4  | 41.2 | 9.7   | 12.3    | 38.1  | 28.7 | 12.2  | 18.7    | 40.7  | 42.5  | 12.9     | 28.6  | 32.7  |
|                                              | A.  | 10.2   | 46.4  | 45.5 | 12.0  | 9.0     | 51.3  | 48.1 | 8.0   | 8.1     | 50.0  | 33.4  | 8.0      | 42.6  | 28.6  |
| FT.                                          | UA. | 14.8   | 31.9  | 29.3 | 25.0  | 15.6    | 34.8  | 33.8 | 35.8  | 25.0    | 38.4  | 48.6  | 19.8     | 36.8  | 33.9  |
|                                              | A.  | 11.6   | 34.2  | 34.2 | 12.0  | 9.6     | 46.4  | 50.0 | 12.0  | 9.8     | 49.1  | 40.0  | 10.5     | 44.1  | 29.3  |
| svo LAS difference (unambiguous - ambiguous) |     |        |       |      |       |         |       |      |       |         |       |       |          |       |       |
| BERT                                         |     | ↑1.5   | ↓3.3  | ↑0.6 | ↑4.1  | ↓3.1    | ↓2.4  | ↑3.5 | ↓11.9 | ↓17.8   | ↓7.9  | ↓3.8  | ↓10.7    | ↓13.2 | ↓4.7  |
| CNN                                          |     | ↑0.4   | ↓0.9  | ↑0.8 | ↓6.6  | ↓2.7    | ↑2.6  | ↑3.4 | ↓47.7 | ↓12.9   | ↓6.8  | ↓22.7 | ↓13      | ↓8.4  | ↓19.9 |
| FT.                                          |     | ↓0.3   | ↓9.0  | ↑0.3 | ↓19.1 | ↓6.0    | ↓4.1  | ↓1.8 | ↓40.6 | ↓12     | ↓5.2  | ↓16.4 | ↓12.6    | ↓11.8 | ↓16.9 |

**Table G.4:** Complement to Chapter 4. Mixed-treebank experiment on ambiguous counterfactual treebanks (see Section 4.3). Top: svo attachment bias metric (LAS) for the unambiguous counterfactual treebanks (UA.) (see Section 4.4) and the ambiguous counterfactual treebanks (A.) (see Section 4.5). Bottom: The difference between the svo LAS on the *unambiguous* counterfactual treebanks and the svo LAS on *ambiguous* counterfactual treebanks.

|              |                      | Polish |       |      |      | Russian |       |      |      | Finnish |       |      | Estonian |       |      |
|--------------|----------------------|--------|-------|------|------|---------|-------|------|------|---------|-------|------|----------|-------|------|
|              |                      | all    | nsubj | obj  | iobj | all     | nsubj | obj  | iobj | all     | nsubj | obj  | all      | nsubj | obj  |
| DM-LSTM+BERT | AB. (UAS)            | 2.2    | 4.7   | 6.5  | 8.7  | 1.1     | 1.9   | 2.6  | 7.3  | 2.2     | 3.7   | 5.1  | 2.8      | 6.6   | 7.0  |
|              | AB. (LAS)            | 4.8    | 15.0  | 11.0 | 11.1 | 2.3     | 6.4   | 5.6  | 5.2  | 3.6     | 8.9   | 7.2  | 3.8      | 8.1   | 8.8  |
|              | PB. (H)              | 0.03   | 0.05  | 0.07 | 0.09 | 0.02    | 0.02  | 0.03 | 0.07 | 0.03    | 0.04  | 0.05 | 0.05     | 0.08  | 0.08 |
|              | PB. (L)              | 0.05   | 0.14  | 0.10 | 0.12 | 0.02    | 0.06  | 0.06 | 0.04 | 0.02    | 0.08  | 0.04 | 0.03     | 0.05  | 0.06 |
|              | <i>Ctrl. PB. (H)</i> | 0.01   | 0.00  | 0.00 | 0.00 | 0.01    | 0.00  | 0.00 | 0.00 | 0.01    | 0.00  | 0.00 | 0.01     | 0.01  | 0.00 |
|              | <i>Ctrl. PB. (L)</i> | 0.02   | 0.01  | 0.01 | 0.02 | 0.00    | 0.01  | 0.01 | 0.02 | 0.00    | 0.00  | 0.00 | 0.01     | 0.01  | 0.01 |
| DM+BERT      | AB. (UAS)            | 2.5    | 4.8   | 6.0  | 9.8  | 1.1     | 2.2   | 2.5  | 10.0 | 1.9     | 3.8   | 3.7  | 3.6      | 6.8   | 8.0  |
|              | AB. (LAS)            | 5.1    | 13.6  | 11.3 | 12.1 | 2.3     | 6.9   | 5.9  | 5.3  | 2.9     | 7.8   | 4.4  | 4.6      | 8.6   | 10.8 |
|              | PB. (H)              | 0.03   | 0.05  | 0.06 | 0.12 | 0.02    | 0.03  | 0.03 | 0.10 | 0.03    | 0.05  | 0.04 | 0.06     | 0.12  | 0.10 |
|              | PB. (L)              | 0.05   | 0.09  | 0.12 | 0.13 | 0.02    | 0.05  | 0.08 | 0.03 | 0.02    | 0.06  | 0.04 | 0.03     | 0.05  | 0.07 |
|              | <i>Ctrl. PB. (H)</i> | 0.01   | 0.00  | 0.00 | 0.00 | 0.01    | 0.00  | 0.00 | 0.01 | 0.01    | 0.00  | 0.00 | 0.03     | 0.03  | 0.01 |
|              | <i>Ctrl. PB. (L)</i> | 0.02   | 0.00  | 0.02 | 0.05 | 0.01    | 0.00  | 0.01 | 0.00 | 0.01    | 0.00  | 0.00 | 0.02     | 0.01  | 0.02 |
| DM+CNN       | AB. (UAS)            | 10.1   | 23.1  | 20.3 | 23.8 | 6.0     | 17.9  | 12.8 | 21.2 | 10.2    | 16.7  | 25.4 | 8.0      | 18.5  | 17.2 |
|              | AB. (LAS)            | 17.4   | 47.6  | 41.0 | 28.9 | 12.4    | 43.2  | 28.1 | 11.8 | 18.4    | 35.7  | 45.7 | 12.7     | 29.4  | 33.5 |
|              | PB. (H)              | 0.12   | 0.25  | 0.24 | 0.25 | 0.08    | 0.21  | 0.15 | 0.25 | 0.15    | 0.21  | 0.31 | 0.14     | 0.25  | 0.21 |
|              | PB. (L)              | 0.13   | 0.41  | 0.34 | 0.27 | 0.09    | 0.36  | 0.25 | 0.13 | 0.12    | 0.30  | 0.36 | 0.10     | 0.23  | 0.31 |
|              | <i>Ctrl. PB. (H)</i> | 0.03   | 0.02  | 0.04 | 0.02 | 0.04    | 0.05  | 0.02 | 0.02 | 0.07    | 0.04  | 0.05 | 0.06     | 0.06  | 0.04 |
|              | <i>Ctrl. PB. (L)</i> | 0.07   | 0.07  | 0.09 | 0.16 | 0.04    | 0.12  | 0.10 | 0.17 | 0.06    | 0.07  | 0.13 | 0.06     | 0.11  | 0.14 |
| DM+FAST.     | AB. (UAS)            | 10.4   | 18.0  | 23.9 | 28.8 | 12.3    | 26.5  | 24.5 | 32.1 | 18.3    | 33.4  | 34.8 | 14.8     | 26.7  | 31.9 |
|              | AB. (LAS)            | 15.2   | 28.6  | 30.3 | 37.6 | 16.6    | 34.7  | 35.0 | 34.1 | 23.1    | 44.6  | 39.9 | 17.2     | 28.3  | 33.3 |
|              | PB. (H)              | 0.14   | 0.24  | 0.29 | 0.32 | 0.16    | 0.29  | 0.26 | 0.35 | 0.25    | 0.37  | 0.40 | 0.23     | 0.36  | 0.35 |
|              | PB. (L)              | 0.11   | 0.28  | 0.25 | 0.34 | 0.09    | 0.31  | 0.24 | 0.26 | 0.14    | 0.37  | 0.29 | 0.12     | 0.28  | 0.25 |
|              | <i>Ctrl. PB. (H)</i> | 0.05   | 0.07  | 0.04 | 0.02 | 0.03    | 0.03  | 0.01 | 0.02 | 0.08    | 0.07  | 0.07 | 0.08     | 0.09  | 0.05 |
|              | <i>Ctrl. PB. (L)</i> | 0.06   | 0.12  | 0.07 | 0.11 | 0.03    | 0.08  | 0.03 | 0.10 | 0.06    | 0.09  | 0.10 | 0.09     | 0.16  | 0.11 |

**Table G.5:** Complement to Chapter 4. Fine-grained bias metrics’ results. *AB.* and *PB.* stand for SVO attachment bias and SVO probability bias, respectively. *(H)* stands for head, *(L)* stands for label. *Ctrl.* stands for *control*.

|                | Polish<br>295 clauses |       |      |      |     | Russian<br>632 clauses |       |      |      |      | Finnish<br>127 clauses |       |     |     | Estonian<br>258 clauses |       |     |
|----------------|-----------------------|-------|------|------|-----|------------------------|-------|------|------|------|------------------------|-------|-----|-----|-------------------------|-------|-----|
|                | all                   | nsubj | obj  | iobj |     | all                    | nsubj | obj  | iobj |      | all                    | nsubj | obj |     | all                     | nsubj | obj |
| DM -LSTM +BERT |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO            | 0.5                   | 0.7   | 0.7  | 1.7  | 0.1 | 0.4                    | 0.3   | 0.0  | 0.1  | 0.0  | 0.0                    | 0.2   | 0.6 | 0.6 |                         |       |     |
| OVS            | 0.8                   | 2.2   | 1.3  | 2.5  | 0.3 | 1.0                    | 1.0   | 1.2  | 0.1  | 1.2  | 1.2                    | 0.2   | 0.5 | 1.2 |                         |       |     |
| SOV            | 0.6                   | 2.1   | 1.7  | 2.8  | 0.2 | 1.3                    | 1.1   | 1.3  | 0.2  | 0.0  | 0.8                    | 0.7   | 0.8 | 1.8 |                         |       |     |
| OSV            | 0.1                   | 0.7   | 2.3  | 1.0  | 0.2 | 1.4                    | 0.8   | 0.0  | 0.5  | 0.0  | 1.2                    | 0.7   | 1.2 | 2.3 |                         |       |     |
| VSO            | 0.5                   | 6.3   | 3.6  | 4.1  | 0.1 | 2.2                    | 1.2   | 3.3  | 0.5  | 0.8  | 0.0                    | 0.2   | 1.1 | 1.4 |                         |       |     |
| VOS            | 0.4                   | 0.6   | 0.9  | 1.9  | 0.2 | 0.1                    | 0.8   | 0.0  | 1.5  | 8.2  | 1.4                    | 0.7   | 1.4 | 1.2 |                         |       |     |
| DM +BERT       |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO            | 0.1                   | 0.3   | 0.0  | 3.0  | 0.2 | 0.4                    | 0.2   | 0.0  | 0.5  | 0.4  | 0.8                    | 1.6   | 1.4 | 1.3 |                         |       |     |
| OVS            | 0.6                   | 0.8   | 2.8  | 3.6  | 0.6 | 1.1                    | 1.7   | 1.1  | 0.2  | 0.7  | 0.4                    | 1.3   | 1.1 | 3.4 |                         |       |     |
| SOV            | 0.5                   | 1.0   | 3.1  | 4.3  | 0.2 | 0.4                    | 1.5   | 2.2  | 0.7  | 1.5  | 1.4                    | 2.9   | 3.5 | 2.8 |                         |       |     |
| OSV            | 0.8                   | 0.4   | 1.5  | 8.5  | 0.4 | 0.5                    | 1.4   | 2.0  | 0.2  | 0.4  | 1.1                    | 2.5   | 2.7 | 7.6 |                         |       |     |
| VSO            | 0.7                   | 6.3   | 3.5  | 3.6  | 0.8 | 2.4                    | 3.4   | 5.6  | 0.8  | 1.0  | 1.3                    | 1.3   | 1.8 | 0.4 |                         |       |     |
| VOS            | 0.4                   | 3.2   | 1.6  | 3.4  | 0.2 | 0.6                    | 0.7   | 1.1  | 1.1  | 4.7  | 2.4                    | 1.9   | 1.3 | 4.2 |                         |       |     |
| DM +CNN        |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO            | 0.9                   | 0.3   | 3.9  | 5.4  | 1.3 | 3.9                    | 2.1   | 6.3  | 1.1  | 1.7  | 3.8                    | 1.1   | 2.6 | 3.1 |                         |       |     |
| OVS            | 2.0                   | 10.5  | 2.8  | 5.4  | 1.5 | 4.7                    | 6.4   | 3.2  | 1.8  | 6.4  | 3.2                    | 1.2   | 3.4 | 1.1 |                         |       |     |
| SOV            | 0.9                   | 2.4   | 4.0  | 9.9  | 1.0 | 5.8                    | 3.8   | 3.7  | 1.7  | 5.3  | 7.7                    | 1.2   | 3.5 | 0.9 |                         |       |     |
| OSV            | 0.8                   | 6.5   | 4.7  | 7.6  | 1.6 | 2.7                    | 9.9   | 4.1  | 2.0  | 3.4  | 2.2                    | 1.1   | 3.0 | 1.9 |                         |       |     |
| VSO            | 0.6                   | 7.0   | 2.5  | 6.4  | 0.4 | 4.3                    | 2.3   | 4.2  | 1.6  | 7.9  | 5.4                    | 1.7   | 4.7 | 5.9 |                         |       |     |
| VOS            | 0.6                   | 3.2   | 4.6  | 5.2  | 0.4 | 3.3                    | 3.9   | 10.0 | 1.2  | 3.4  | 2.7                    | 0.7   | 2.9 | 3.9 |                         |       |     |
| DM +FASTTEXT   |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO            | 1.1                   | 4.0   | 1.8  | 4.3  | 0.2 | 2.9                    | 0.3   | 2.1  | 0.9  | 3.2  | 2.9                    | 0.7   | 3.2 | 2.7 |                         |       |     |
| OVS            | 3.2                   | 9.4   | 12.4 | 4.9  | 0.7 | 1.4                    | 5.9   | 14.5 | 1.8  | 5.4  | 8.8                    | 1.5   | 1.9 | 8.3 |                         |       |     |
| SOV            | 0.4                   | 4.7   | 2.1  | 2.4  | 1.2 | 7.0                    | 2.3   | 1.8  | 1.3  | 4.2  | 4.8                    | 0.5   | 2.8 | 4.4 |                         |       |     |
| OSV            | 1.8                   | 6.5   | 8.9  | 4.3  | 0.8 | 1.6                    | 5.5   | 6.3  | 1.7  | 4.2  | 3.5                    | 0.7   | 0.8 | 3.0 |                         |       |     |
| VSO            | 1.3                   | 8.8   | 3.4  | 3.5  | 5.5 | 6.0                    | 9.2   | 7.4  | 5.4  | 10.3 | 12.1                   | 1.9   | 6.5 | 4.1 |                         |       |     |
| VOS            | 1.6                   | 3.7   | 7.2  | 5.8  | 6.4 | 4.4                    | 13.0  | 14.0 | 6.7  | 6.6  | 11.2                   | 1.0   | 7.6 | 4.2 |                         |       |     |

**Table G.6:** Complement to Chapter 4. Standard deviation scores for LAS results in Table 4.10 (unambiguous counterfactual treebanks).

|              | Polish<br>295 clauses |       |      |      |     | Russian<br>632 clauses |       |      |      |      | Finnish<br>127 clauses |       |     |     | Estonian<br>258 clauses |       |     |
|--------------|-----------------------|-------|------|------|-----|------------------------|-------|------|------|------|------------------------|-------|-----|-----|-------------------------|-------|-----|
|              | all                   | nsubj | obj  | iobj |     | all                    | nsubj | obj  | iobj |      | all                    | nsubj | obj |     | all                     | nsubj | obj |
| DM +BERT     |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO          | 0.5                   | 1.1   | 1.2  | 23.1 | 0.2 | 0.3                    | 0.8   | 0.0  | 0.5  | 0.0  | 3.3                    | 3.2   | 3.0 | 2.9 |                         |       |     |
| OVS          | 0.9                   | 2.8   | 8.5  | 0.0  | 0.6 | 3.7                    | 4.7   | 0.0  | 1.6  | 0.0  | 3.3                    | 3.3   | 1.0 | 6.9 |                         |       |     |
| SOV          | 0.7                   | 1.7   | 6.3  | 11.5 | 0.3 | 2.0                    | 2.5   | 16.7 | 1.2  | 0.0  | 10.0                   | 3.0   | 5.5 | 4.6 |                         |       |     |
| OSV          | 0.9                   | 3.7   | 4.9  | 0.0  | 0.3 | 2.0                    | 5.5   | 16.6 | 1.5  | 3.4  | 10.0                   | 3.4   | 3.6 | 4.5 |                         |       |     |
| VSO          | 0.6                   | 7.2   | 4.2  | 19.1 | 0.5 | 4.2                    | 7.3   | 16.7 | 1.2  | 6.4  | 8.6                    | 3.5   | 5.5 | 2.1 |                         |       |     |
| VOS          | 1.1                   | 6.1   | 0.0  | 19.1 | 0.3 | 2.3                    | 4.3   | 0.0  | 0.8  | 5.5  | 5.4                    | 3.4   | 4.0 | 1.7 |                         |       |     |
| DM +CNN      |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO          | 1.5                   | 2.0   | 5.7  | 11.5 | 0.4 | 2.7                    | 3.5   | 19.2 | 2.0  | 6.4  | 6.7                    | 1.7   | 2.0 | 6.0 |                         |       |     |
| OVS          | 0.3                   | 7.6   | 4.6  | 10.0 | 0.9 | 6.3                    | 6.9   | 0.0  | 1.0  | 0.0  | 6.6                    | 1.8   | 2.6 | 6.6 |                         |       |     |
| SOV          | 1.4                   | 4.4   | 7.0  | 10.0 | 0.3 | 2.3                    | 4.2   | 16.7 | 1.3  | 10.0 | 10.0                   | 2.4   | 6.7 | 6.0 |                         |       |     |
| OSV          | 0.9                   | 6.6   | 8.8  | 16.3 | 0.5 | 4.5                    | 5.7   | 16.7 | 2.5  | 8.4  | 8.6                    | 1.3   | 2.6 | 3.6 |                         |       |     |
| VSO          | 1.3                   | 5.3   | 6.1  | 19.1 | 0.4 | 2.7                    | 2.6   | 33.4 | 0.6  | 5.5  | 8.4                    | 1.5   | 5.7 | 6.9 |                         |       |     |
| VOS          | 0.3                   | 4.4   | 3.0  | 0.0  | 0.3 | 5.3                    | 4.1   | 19.3 | 1.1  | 7.7  | 6.7                    | 1.0   | 7.0 | 9.5 |                         |       |     |
| DM +FASTTEXT |                       |       |      |      |     |                        |       |      |      |      |                        |       |     |     |                         |       |     |
| SVO          | 1.2                   | 2.1   | 2.0  | 11.5 | 0.5 | 3.0                    | 2.0   | 0.0  | 2.5  | 11.6 | 14.8                   | 2.0   | 5.5 | 6.0 |                         |       |     |
| OVS          | 1.2                   | 9.8   | 7.4  | 10.0 | 0.4 | 4.0                    | 6.5   | 0.0  | 2.4  | 3.3  | 8.6                    | 1.6   | 3.4 | 5.5 |                         |       |     |
| SOV          | 1.3                   | 2.4   | 3.0  | 0.0  | 0.5 | 6.5                    | 2.6   | 0.0  | 1.9  | 12.8 | 12.8                   | 0.6   | 1.2 | 7.9 |                         |       |     |
| OSV          | 1.3                   | 5.5   | 3.4  | 0.0  | 0.4 | 4.0                    | 8.5   | 16.6 | 1.9  | 5.4  | 8.4                    | 0.8   | 6.5 | 2.7 |                         |       |     |
| VSO          | 2.0                   | 10.9  | 2.8  | 10.0 | 1.3 | 5.6                    | 6.7   | 16.6 | 2.3  | 11.4 | 6.4                    | 1.6   | 3.8 | 4.9 |                         |       |     |
| VOS          | 1.9                   | 5.9   | 14.1 | 11.5 | 1.2 | 3.7                    | 7.6   | 0.0  | 2.5  | 3.8  | 13.8                   | 1.0   | 4.3 | 3.0 |                         |       |     |

**Table G.7:** Complement to Chapter 4. Standard deviation scores for LAS results in Table 4.13 (ambiguous counterfactual treebanks).

|      |   | Polish<br>302 clauses |       |     |      | Russian<br>654 clauses |       |     |      | Finnish<br>129 clauses |       |      | Estonian<br>268 clauses |       |     |
|------|---|-----------------------|-------|-----|------|------------------------|-------|-----|------|------------------------|-------|------|-------------------------|-------|-----|
|      |   | core                  | nsubj | obj | iobj | core                   | nsubj | obj | iobj | core                   | nsubj | obj  | core                    | nsubj | obj |
| BERT | O | 0.5                   | 0.3   | 0.9 | 3.9  | 0.1                    | 0.3   | 0.2 | 1.1  | 0.5                    | 0.4   | 0.7  | 1.5                     | 0.7   | 2.3 |
|      | R | 0.9                   | 1.5   | 1.0 | 4.6  | 0.7                    | 1.2   | 1.6 | 2.0  | 0.9                    | 1.7   | 1.6  | 1.3                     | 0.9   | 1.8 |
| CNN  | O | 1.0                   | 0.8   | 1.2 | 5.5  | 1.3                    | 1.5   | 1.6 | 8.4  | 1.8                    | 1.5   | 2.1  | 1.6                     | 1.7   | 2.8 |
|      | R | 1.5                   | 1.5   | 4.0 | 3.5  | 1.4                    | 2.0   | 1.8 | 6.8  | 1.7                    | 1.9   | 2.9  | 1.9                     | 1.7   | 4.4 |
| FT.  | O | 3.9                   | 5.5   | 4.8 | 2.8  | 0.9                    | 4.3   | 3.0 | 2.1  | 4.5                    | 6.0   | 13.2 | 2.6                     | 2.1   | 5.8 |
|      | R | 3.9                   | 5.5   | 4.8 | 2.8  | 0.9                    | 4.3   | 3.0 | 2.1  | 4.5                    | 6.0   | 13.2 | 2.6                     | 2.1   | 5.8 |

**Table G.8:** Complement to Chapter 5. Standard deviation scores for LAS results in Table 5.3 (unambiguous rotated results).

|      |   | Polish<br>49 clauses |       |     |      | Russian<br>181 clauses |       |     |      | Finnish<br>15 clauses |       |      | Estonian<br>49 clauses |       |     |
|------|---|----------------------|-------|-----|------|------------------------|-------|-----|------|-----------------------|-------|------|------------------------|-------|-----|
|      |   | core                 | nsubj | obj | iobj | core                   | nsubj | obj | iobj | core                  | nsubj | obj  | core                   | nsubj | obj |
| BERT | O | 1.3                  | 1.2   | 2.2 | 23.1 | 0.3                    | 0.3   | 0.7 | 0.0  | 2.0                   | 3.3   | 3.3  | 2.8                    | 4.1   | 1.7 |
|      | R | 6.6                  | 9.8   | 4.7 | 10.0 | 2.1                    | 2.7   | 2.0 | 16.7 | 2.7                   | 5.5   | 5.4  | 1.0                    | 2.0   | 2.0 |
| CNN  | O | 1.7                  | 3.7   | 2.3 | 11.5 | 1.7                    | 1.7   | 2.5 | 0.0  | 4.3                   | 9.4   | 6.7  | 2.1                    | 3.5   | 7.0 |
|      | R | 3.0                  | 2.0   | 4.9 | 11.5 | 0.7                    | 2.4   | 1.4 | 0.0  | 12.3                  | 9.4   | 15.9 | 4.5                    | 4.1   | 7.0 |
| FT.  | O | 2.2                  | 3.5   | 1.9 | 10.0 | 0.7                    | 2.5   | 2.1 | 0.0  | 10.0                  | 14.4  | 15.9 | 4.0                    | 6.7   | 1.7 |
|      | R | 7.2                  | 9.2   | 6.3 | 11.5 | 2.5                    | 7.0   | 2.4 | 16.6 | 6.4                   | 14.8  | 6.4  | 3.3                    | 5.4   | 6.1 |

**Table G.9:** Complement to Chapter 5. Standard deviation scores for LAS results in Table 5.4 (ambiguous rotated results).

| Polish<br>clause count: 295          |     |       |      | Russian<br>clause count: 632 |  |     |       | Finnish<br>clause count: 127 |      |  | Estonian<br>clause count: 258 |       |      |  |
|--------------------------------------|-----|-------|------|------------------------------|--|-----|-------|------------------------------|------|--|-------------------------------|-------|------|--|
|                                      | all | nsubj | obj  | iobj                         |  | all | nsubj | obj                          | iobj |  | all                           | nsubj | obj  |  |
| MORPHOLOGICALLY BLIND BASELINES      |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| DM + BERT                            |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| Sim.                                 | 0.1 | 1.4   | 0.8  | 2.8                          |  | 0.1 | 0.5   | 1.4                          | 1.7  |  | 0.8                           | 3.7   | 1.1  |  |
| svo                                  | 0.5 | 4.6   | 2.4  | 6.9                          |  | 1.4 | 3.5   | 4.3                          | 3.5  |  | 0.5                           | 1.0   | 3.3  |  |
| ovs                                  | 0.9 | 1.1   | 1.2  | 3.0                          |  | 0.6 | 3.2   | 2.2                          | 3.2  |  | 0.5                           | 0.4   | 1.0  |  |
| sov                                  | 0.6 | 2.9   | 0.5  | 3.4                          |  | 0.6 | 2.5   | 1.0                          | 1.1  |  | 0.5                           | 1.9   | 1.0  |  |
| osv                                  | 0.7 | 1.3   | 1.3  | 3.9                          |  | 0.3 | 1.5   | 4.9                          | 8.4  |  | 0.2                           | 0.8   | 2.4  |  |
| vso                                  | 0.4 | 1.7   | 0.9  | 4.1                          |  | 0.4 | 1.3   | 1.4                          | 2.1  |  | 0.4                           | 1.0   | 3.0  |  |
| vos                                  | 0.9 | 2.1   | 3.3  | 6.3                          |  | 0.8 | 3.0   | 4.1                          | 2.1  |  | 0.4                           | 0.5   | 3.2  |  |
| DM + CNN                             |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| Sim.                                 | 0.7 | 1.3   | 4.6  | 7.9                          |  | 7.3 | 6.7   | 7.8                          | 31.5 |  | 1.7                           | 6.0   | 4.2  |  |
| svo                                  | 1.1 | 2.0   | 5.9  | 3.4                          |  | 2.4 | 6.3   | 5.9                          | 29.5 |  | 1.9                           | 9.6   | 4.5  |  |
| ovs                                  | 0.2 | 2.7   | 1.2  | 3.4                          |  | 4.2 | 1.5   | 2.4                          | 5.6  |  | 1.8                           | 0.4   | 5.3  |  |
| sov                                  | 0.7 | 4.9   | 0.8  | 2.4                          |  | 2.6 | 17.3  | 6.4                          | 8.2  |  | 1.5                           | 2.1   | 10.5 |  |
| osv                                  | 0.8 | 2.9   | 2.8  | 3.1                          |  | 3.6 | 10.5  | 8.2                          | 5.2  |  | 1.6                           | 9.0   | 1.7  |  |
| vso                                  | 0.7 | 3.2   | 2.4  | 1.3                          |  | 2.2 | 3.0   | 2.8                          | 5.4  |  | 0.3                           | 1.3   | 4.2  |  |
| vos                                  | 1.3 | 1.1   | 5.0  | 3.9                          |  | 3.0 | 2.9   | 2.0                          | 21.3 |  | 1.2                           | 0.5   | 6.3  |  |
| MODELS PROCESSING UNALTERED LANGUAGE |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| DM +BERT                             |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| Sim.                                 | 0.2 | 0.6   | 0.4  | 1.7                          |  | 0.2 | 0.4   | 0.2                          | 1.0  |  | 0.5                           | 0.4   | 0.8  |  |
| svo                                  | 0.3 | 1.6   | 1.2  | 1.8                          |  | 0.2 | 1.2   | 1.9                          | 2.4  |  | 0.8                           | 1.9   | 1.5  |  |
| ovs                                  | 0.3 | 2.2   | 2.0  | 1.7                          |  | 0.8 | 2.8   | 2.8                          | 2.8  |  | 0.9                           | 2.0   | 1.2  |  |
| sov                                  | 0.5 | 8.0   | 6.4  | 5.3                          |  | 0.9 | 5.0   | 1.0                          | 2.8  |  | 0.9                           | 2.3   | 5.8  |  |
| osv                                  | 1.3 | 7.5   | 2.4  | 3.4                          |  | 0.5 | 2.6   | 2.9                          | 2.1  |  | 0.6                           | 1.4   | 1.4  |  |
| vso                                  | 0.7 | 6.6   | 1.6  | 6.2                          |  | 0.4 | 3.8   | 3.4                          | 7.4  |  | 0.3                           | 7.5   | 6.3  |  |
| vos                                  | 0.3 | 2.6   | 1.8  | 5.3                          |  | 0.5 | 2.3   | 4.2                          | 2.7  |  | 1.1                           | 5.6   | 3.2  |  |
| DM +CNN                              |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| Sim.                                 | 0.3 | 2.5   | 3.1  | 5.1                          |  | 1.0 | 2.8   | 1.8                          | 3.2  |  | 1.3                           | 1.7   | 3.0  |  |
| svo                                  | 1.1 | 1.9   | 7.4  | 2.4                          |  | 1.5 | 4.3   | 4.7                          | 8.9  |  | 0.8                           | 3.6   | 2.3  |  |
| ovs                                  | 0.9 | 7.2   | 2.1  | 1.6                          |  | 1.5 | 6.6   | 6.3                          | 8.2  |  | 1.3                           | 4.5   | 3.0  |  |
| sov                                  | 1.3 | 4.4   | 4.9  | 6.7                          |  | 0.7 | 6.5   | 3.7                          | 4.7  |  | 1.8                           | 3.9   | 3.9  |  |
| osv                                  | 0.9 | 6.6   | 8.2  | 2.8                          |  | 0.7 | 4.7   | 7.0                          | 8.9  |  | 1.3                           | 1.6   | 2.1  |  |
| vso                                  | 0.8 | 4.1   | 1.8  | 4.4                          |  | 0.2 | 2.8   | 1.6                          | 3.7  |  | 2.0                           | 7.7   | 6.3  |  |
| vos                                  | 1.3 | 0.5   | 6.3  | 4.1                          |  | 0.2 | 3.2   | 3.0                          | 9.6  |  | 0.8                           | 3.2   | 3.2  |  |
| DM +FASTTEXT                         |     |       |      |                              |  |     |       |                              |      |  |                               |       |      |  |
| Sim.                                 | 1.0 | 3.1   | 0.6  | 1.6                          |  | 0.4 | 2.7   | 0.3                          | 3.2  |  | 1.2                           | 2.7   | 3.0  |  |
| svo                                  | 2.3 | 9.3   | 5.6  | 4.3                          |  | 0.8 | 4.2   | 2.5                          | 8.6  |  | 2.6                           | 5.1   | 14.3 |  |
| ovs                                  | 1.7 | 6.9   | 11.7 | 5.9                          |  | 0.4 | 5.3   | 2.9                          | 5.3  |  | 0.9                           | 8.4   | 11.1 |  |
| sov                                  | 1.5 | 7.6   | 2.5  | 1.9                          |  | 0.9 | 5.5   | 2.4                          | 1.3  |  | 2.4                           | 5.2   | 10.0 |  |
| osv                                  | 1.3 | 5.4   | 9.6  | 7.0                          |  | 0.6 | 3.7   | 2.2                          | 2.4  |  | 1.4                           | 9.5   | 7.2  |  |
| vso                                  | 0.7 | 10.0  | 2.8  | 5.1                          |  | 1.9 | 4.6   | 5.1                          | 9.8  |  | 5.0                           | 10.8  | 8.7  |  |
| vos                                  | 2.4 | 3.1   | 6.5  | 6.1                          |  | 2.3 | 6.0   | 7.1                          | 11.7 |  | 7.2                           | 11.6  | 9.0  |  |

**Table G.10:** Complement to Chapter 5. Standard deviation scores for LAS results in Table 5.6 (results for unambiguous rotated counterfactual treebanks).

|                                                                       | BERT |       |      |      | CNN                       |      |      |      | FASTTEXT |       |      |      |
|-----------------------------------------------------------------------|------|-------|------|------|---------------------------|------|------|------|----------|-------|------|------|
|                                                                       | all  | nsubj | obj  | iobj | Polish, clause count: 295 |      |      |      | all      | nsubj | obj  | iobj |
| Unaltered Polish (4 models)                                           |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 87.6 | 89.7  | 87.4 | 83.1 | 81.8                      | 86.4 | 81.9 | 59.0 | 72.0     | 66.1  | 73.6 | 49.6 |
| ovs                                                                   | 82.0 | 68.4  | 70.4 | 72.1 | 64.9                      | 29.2 | 30.0 | 28.3 | 58.0     | 30.6  | 32.5 | 16.8 |
| sov                                                                   | 75.3 | 71.8  | 38.0 | 41.5 | 68.3                      | 80.1 | 25.1 | 19.3 | 60.4     | 55.7  | 41.3 | 18.0 |
| osv                                                                   | 80.9 | 70.0  | 70.4 | 66.7 | 61.5                      | 32.6 | 30.6 | 23.4 | 54.7     | 37.8  | 25.9 | 13.5 |
| vso                                                                   | 67.8 | 35.4  | 23.8 | 47.0 | 61.0                      | 27.2 | 30.5 | 27.9 | 56.1     | 32.7  | 33.8 | 20.5 |
| vos                                                                   | 78.7 | 47.2  | 85.0 | 78.1 | 65.1                      | 11.2 | 70.5 | 51.2 | 59.4     | 21.6  | 56.7 | 33.6 |
| Case tag concatenations (4 models)                                    |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 90.3 | 99.2  | 95.3 | 90.2 | 87.5                      | 98.4 | 93.6 | 84.8 | 67.4     | 70.7  | 69.8 | 58.2 |
| ovs                                                                   | 89.3 | 97.5  | 93.9 | 84.8 | 85.4                      | 97.9 | 90.9 | 66.8 | 55.4     | 35.9  | 35.8 | 28.7 |
| sov                                                                   | 83.6 | 98.4  | 68.8 | 52.0 | 79.6                      | 95.9 | 69.3 | 41.0 | 56.6     | 62.1  | 36.1 | 22.1 |
| osv                                                                   | 87.4 | 97.1  | 86.9 | 75.4 | 84.2                      | 95.2 | 88.3 | 62.3 | 53.9     | 39.0  | 36.2 | 27.9 |
| vso                                                                   | 78.2 | 81.2  | 56.9 | 54.1 | 77.3                      | 94.2 | 62.5 | 36.9 | 48.2     | 32.0  | 27.8 | 12.3 |
| vos                                                                   | 86.2 | 87.0  | 92.7 | 81.6 | 83.8                      | 91.6 | 90.0 | 74.6 | 52.8     | 23.8  | 51.3 | 36.9 |
| Case tag concatenations; first character only (4 models)              |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 89.6 | 98.9  | 95.6 | 84.4 | 83.0                      | 89.4 | 87.3 | 70.9 | 68.8     | 75.8  | 74.1 | 54.5 |
| ovs                                                                   | 88.6 | 95.8  | 95.5 | 76.7 | 79.2                      | 82.6 | 76.7 | 56.5 | 56.0     | 37.1  | 39.4 | 13.9 |
| sov                                                                   | 81.9 | 97.0  | 64.6 | 40.6 | 73.3                      | 85.0 | 56.8 | 37.3 | 56.7     | 67.5  | 36.2 | 7.4  |
| osv                                                                   | 87.5 | 96.9  | 91.0 | 68.0 | 77.7                      | 83.6 | 72.8 | 54.5 | 54.4     | 41.0  | 40.5 | 11.1 |
| vso                                                                   | 74.2 | 68.1  | 46.0 | 43.0 | 71.3                      | 77.3 | 51.2 | 41.0 | 46.0     | 28.8  | 35.9 | 8.6  |
| vos                                                                   | 84.1 | 77.5  | 92.8 | 76.2 | 76.2                      | 70.2 | 77.0 | 68.4 | 50.7     | 18.2  | 54.5 | 25.0 |
| Case tag concatenations; first character only + syncretism (4 models) |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 87.5 | 89.7  | 88.2 | 85.7 | 81.0                      | 74.7 | 82.8 | 67.6 | 68.2     | 67.7  | 76.9 | 47.1 |
| ovs                                                                   | 79.7 | 62.9  | 68.8 | 66.0 | 73.0                      | 53.1 | 60.3 | 52.0 | 52.9     | 19.8  | 31.8 | 15.6 |
| sov                                                                   | 74.7 | 71.8  | 40.0 | 38.5 | 69.6                      | 69.5 | 46.8 | 35.2 | 56.0     | 57.2  | 40.2 | 7.8  |
| osv                                                                   | 79.0 | 73.3  | 64.4 | 61.1 | 69.8                      | 56.5 | 50.1 | 47.5 | 52.5     | 30.5  | 36.9 | 15.2 |
| vso                                                                   | 65.8 | 27.9  | 30.3 | 48.0 | 63.3                      | 24.6 | 52.1 | 35.6 | 49.8     | 23.8  | 36.2 | 9.8  |
| vos                                                                   | 77.5 | 46.7  | 84.5 | 77.4 | 69.7                      | 30.1 | 73.6 | 65.2 | 53.0     | 14.9  | 54.9 | 27.9 |
| Case tag concatenations; first character only + allomorphy (4 models) |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 90.5 | 99.3  | 94.8 | 86.9 | 83.7                      | 87.7 | 86.5 | 72.6 | 69.3     | 77.5  | 77.0 | 52.9 |
| ovs                                                                   | 89.0 | 95.6  | 92.5 | 81.5 | 79.8                      | 80.2 | 79.2 | 64.8 | 55.8     | 33.4  | 43.4 | 14.8 |
| sov                                                                   | 81.5 | 97.2  | 57.9 | 40.2 | 73.9                      | 83.8 | 53.5 | 35.7 | 58.1     | 71.4  | 38.1 | 7.4  |
| osv                                                                   | 87.3 | 96.8  | 86.7 | 70.9 | 77.3                      | 78.6 | 71.5 | 57.0 | 54.3     | 41.4  | 40.1 | 13.6 |
| vso                                                                   | 75.8 | 77.9  | 40.2 | 43.0 | 72.0                      | 76.6 | 52.2 | 36.5 | 45.8     | 27.6  | 31.7 | 6.1  |
| vos                                                                   | 84.2 | 79.3  | 92.2 | 78.3 | 76.6                      | 69.0 | 76.4 | 70.1 | 51.2     | 15.8  | 52.9 | 27.9 |
| Case encoded in a preposition (4 models)                              |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 93.8 | 99.3  | 96.6 | 89.8 | 92.4                      | 98.7 | 95.7 | 91.8 | 91.8     | 98.5  | 94.5 | 86.9 |
| ovs                                                                   | 93.2 | 95.8  | 96.8 | 81.2 | 91.3                      | 98.7 | 95.1 | 77.0 | 90.2     | 96.4  | 91.5 | 76.7 |
| sov                                                                   | 91.1 | 98.9  | 80.0 | 62.7 | 89.6                      | 97.7 | 84.6 | 66.0 | 84.9     | 88.4  | 79.5 | 57.0 |
| osv                                                                   | 92.7 | 95.7  | 93.9 | 79.5 | 90.7                      | 96.1 | 93.6 | 67.2 | 88.0     | 90.1  | 87.2 | 75.0 |
| vso                                                                   | 90.4 | 97.7  | 75.3 | 57.8 | 88.4                      | 97.2 | 81.8 | 66.8 | 85.3     | 95.5  | 77.1 | 57.0 |
| vos                                                                   | 90.7 | 80.4  | 95.8 | 77.5 | 90.4                      | 90.5 | 95.2 | 85.2 | 89.1     | 88.0  | 93.1 | 77.9 |
| Decased Polish, case tag fed as an additional feature (4 models)      |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 91.1 | 97.9  | 97.2 | 86.9 | 90.0                      | 99.6 | 96.0 | 86.5 | 88.0     | 96.8  | 93.5 | 86.9 |
| ovs                                                                   | 88.6 | 83.0  | 94.5 | 72.2 | 88.5                      | 98.9 | 94.5 | 77.0 | 86.8     | 97.3  | 90.2 | 72.1 |
| sov                                                                   | 85.9 | 96.4  | 74.2 | 49.2 | 84.8                      | 97.5 | 76.2 | 55.8 | 81.1     | 91.0  | 72.1 | 53.7 |
| osv                                                                   | 87.7 | 89.9  | 92.0 | 69.3 | 87.6                      | 96.0 | 91.9 | 70.5 | 84.8     | 92.6  | 84.0 | 66.0 |
| vso                                                                   | 82.8 | 86.5  | 62.5 | 46.3 | 83.2                      | 95.8 | 70.0 | 52.5 | 80.3     | 96.7  | 59.7 | 48.8 |
| vos                                                                   | 85.8 | 77.7  | 94.6 | 78.3 | 86.8                      | 90.0 | 93.1 | 79.9 | 85.1     | 86.2  | 87.5 | 77.0 |
| Unaltered Polish, case tag fed as an additional feature (1 model)     |      |       |      |      |                           |      |      |      |          |       |      |      |
| svo                                                                   | 91.4 | 98.3  | 96.2 | 85.2 | 90.2                      | 98.3 | 96.6 | 86.9 | 88.9     | 97.6  | 92.3 | 91.8 |
| ovs                                                                   | 89.3 | 85.4  | 94.9 | 75.4 | 88.0                      | 98.0 | 94.0 | 78.7 | 85.9     | 94.2  | 88.5 | 72.1 |
| sov                                                                   | 84.3 | 96.6  | 54.9 | 50.8 | 85.1                      | 96.3 | 79.1 | 59.0 | 80.3     | 91.5  | 71.5 | 37.7 |
| osv                                                                   | 88.7 | 87.8  | 91.9 | 80.3 | 87.5                      | 93.9 | 94.5 | 78.7 | 85.0     | 92.9  | 82.6 | 67.2 |
| vso                                                                   | 79.3 | 75.3  | 39.1 | 47.5 | 82.8                      | 92.5 | 65.1 | 55.7 | 79.2     | 94.6  | 51.1 | 42.6 |
| vos                                                                   | 85.5 | 68.8  | 93.2 | 83.6 | 86.4                      | 88.1 | 92.3 | 85.2 | 85.3     | 80.3  | 92.3 | 77.0 |

**Table G.11:** Complement to Chapter 7. Results for the development split of the *unambiguous* rotated counterfactual treebanks for models trained on various synthetic versions of Polish.