The dative alternation revisited: fresh insights from contemporary British spoken data

Gard B Jenset, Barbara McGillivray, Michael Rundell

Motivation

A well-known feature of English grammar is the dative alternation, whereby a verb may be used in a V-NP-NP construction (*Give me the money*) or with a prepositional phrase in the pattern V-NP-PP, typically with the preposition *to* (*Give the money to me*). In this study, we use data from the Early-Access Subset (EAS) of the Spoken British National Corpus 2014 to investigate the behaviour of six high-frequency verbs whose argument structure preferences include the dative alternation. Given that speakers have both patterns available to them, our goal is to discover whether the choice of pattern is motivated rather than random — and if so, what factors influenced that choice.

Although the dative alternation is a well-researched topic, most published work draws either on introspection or on data from written sources. Using contemporary unscripted spoken text from face-to-face conversations takes us into new territory, especially as the linguistic data in the EAS corpus are complemented by a wide range of sociolinguistic information on participating speakers. By “sociolinguistic information” we mean the social phenomena that co-occur with linguistic variables (Bayley 2002, 118). This represents a powerful new research resource, and in this chapter we show how it yields new insights into the use of the dative alternation.

Previous research

The dative alternation is one of several English constructions that offer the speaker a choice in how to order the information in an utterance. Although such variation can be investigated from several perspectives (Arnold et al. 2000, 28), our starting point is the identification of factors that correctly predict the choice of information ordering. In the case of the dative alternation, it is well established that the structural complexity, or heaviness, of constituents play a role (Arnold et al. 2000), in accordance with Behagel's Law of ordering short constituents before long ones (Köhler 1999; Arnold et al. 2000, 29). Additionally, Arnold et al. demonstrated that information status, or "givenness", plays a role, and that givenness and heaviness are partially independent constraints modulated by the strength of their effects. In their ground-breaking quantitative study of the English dative alternation, Bresnan et al. (2007) showed that intuitions are insufficient for investigating the intricacies of the dative alternation, and that the range of possible variation is greater than had previously been assumed. Using multivariate statistical techniques (generalized linear regression modelling), they confirmed that discourse factors have an independent role to play, and that these effects are significant even when conditioned on specific verbs and verb senses. Jenset and Johansson (2013) used a large set of data from the Web to study the effects of the semantics of the theme, e.g. the thing being given in "gave the X to her". They found semantic effects that suggest the choice of construction is moderated by the semantics of the theme. However, to retrieve the data from unannotated text, Jenset and Johansson extracted only data with pronominal recipients, which might have affected the results.

In contrast with the discourse and grammatical factors influencing the dative alternation, the effects of sociolinguistic variables at the level of the individual is less clear. So far, sociolinguistic studies have mainly identified differences between major varieties of English, such as Australian English (Bresnan and Ford 2010), New Zealand English (Bresnan and Hay 2008), and African-American English (Kendall, Bresnan, and Van Herk 2011). The results in these studies all point to a largely shared English grammar with respect to the dative alternation, but with minor probabilistic variations, such as the New Zealand tendency to produce more inanimate recipients in the V-NP-NP construction (Bresnan and Hay 2008). However, this still leaves open the status of individual sociolinguistic variables such as age or education. In their study of the dative alternation in African-American English, Kendall, Bresnan, and Van Herk (2011) found that gender and age were not statistically significant predictors of the dative alternation using a binary logistic regression. This result might be due to the use of written data, but other studies using spoken data such as Bresnan and Hay (2008) and Bresnan and Ford (2010) reach the same conclusion. However, these studies rely in part on the Switchboard corpus (Godfrey et al. 1992), which consists of recordings of telephone conversations between strangers. It is possible that a corpus of unscripted, face to face dialogue between speakers who know one another might lead to different results.

Research goals

The objective of this study is to identify those factors which might influence a speaker — in any given situation — to prefer one dative pattern over the other. The range of potentially significant features is broad, so there are plenty of candidates to choose from. In the EAS dataset, the raw language data are semantically tagged (using UCREL's semantic analysis system, USAS, whose tags are available at http://ucrel.lancs.ac.uk/usas/semtags.txt), and further supplemented by rich sociolinguistic metadata, which provide a snapshot of speakers' age, gender, level of education, occupation, dialect, and socio-economic status. It also tells us about the closeness (or otherwise) of the relationships between the participants in any interaction. Any or all of these factors could have a bearing on the choice of construction, as could contextual features such as the type of direct and indirect objects (which brings in questions of the verb’s selectional preferences) and their length. Our goal, therefore, is to determine the effects, if any, of these variables on speakers’ choices.

High-frequency verbs like the ones examined in this study are typically complex in terms of their semantics, syntax, and phraseology. So the initial stage of data extraction, aimed at identifying relevant instances of the dative constructions, generated a good deal of noise. A manual process of cleaning and annotation followed, and the resulting set of around 2000 concordance lines formed the basis for the investigation. However, with such a large number and wide range of potentially significant variables, the task of discovering whether the choice of construction is non-random (and if so, which factors play a part in motivating it) is one of considerable complexity. An important feature of this study, therefore, is our use of state-of-the-art multivariate statistical techniques, in order to account for the interplay of the potentially significant variables. Moreover, in seeking to model speakers’ choices as a function of several variables, our research exploits many of the unique features of this rich dataset.

Our general goal is to discover why speakers select one dative pattern over another, and at a more specific level this will involve determining whether speakers’ choices are affected by linguistic or semantic aspects of the co-text, and/or by sociolinguistic factors such as age, gender, and social status. Finally, we aimed to learn whether the evidence in the EAS data can help to confirm the results of previous studies.

Description of corpus data and metadata

The source data for this study are the EAS of the Spoken BNC 2014. The data consists of transcripts of spontaneous, informal conversation, recorded between 2012 and 2015, and runs to

4,789,185 words. Spoken BNC 2014 has (by design) many similarities with the “demographic” spoken component of the original BNC (BNC XML Edition, http://www.natcorp.ox.ac.uk/). As in the BNC of the early 1990s, the sampling frame includes a range of sociolinguistic categories, such as age, gender, and socio-economic status. Speakers represent various combinations of these, and in the EAS this information is available as a rich set of metadata which complements the transcribed recordings.

It is important to note, however, that for many of the key sociolinguistic indicators, speakers are not optimally distributed across the sampling frame. Looking at participants’ ages, for example, over 40% of speakers in the data we studied were in the age range 19-29. Representation in the “middle” age ranges (30-49) shows a sharp decline, and rises again for the cohort of people aged 50 and over. The sample is similarly skewed in the case of social status. Speakers from social grades A and B (roughly speaking, higher-status middle-class individuals) are hugely over-represented, with close to half of all sentences in our sample coming from members of these groups, while A’s and B’s make up only 27% of the UK population at large. Social grades C1, C2, and D are thinly populated in the EAS data, but grade E (non-working, pensioners etc.) is again over-represented, comprising almost a third of speakers in the data we studied, as against 8% in the national figures. The distribution of dialects is skewed towards southern English varieties; for example, the number of data points corresponding to speakers whose dialect is labelled as “south” is 970 (50% of the total in the dataset), and 387 (20%) are unspecified. It is useful to be aware of these imbalances, and a degree of caution is required when extrapolating any findings to the wider population.

Scope of the research

To investigate the dative alternation, we needed to find verbs for which both patterns were equally acceptable, even if not equally frequent. A number of mid-frequency verbs were looked at, including *award*, *hand*, *grant*, and *mail*, all of which include the two dative patterns among their syntactic preferences. These yielded too few usable data points, and it became clear that, if we were to collect adequate material for the study, in a corpus of fairly modest size, we would need to focus on high-frequency words.

The six verbs on which this study is based are: *give*, *lend*, *offer*, *sell*, *send*, and *show*, and basic frequency data is shown in Table 9.1:

Verb	**Frequency in EAS**	**Datives**
give	3980	1000*
lend	65	38
offer	238	72
show	1066	276
sell	1063	103
send	1527	570

Table 1: the six verbs with their raw frequency in the corpus and the number of dative instances in these data

*Note that *give* is a special case: its frequency in the corpus (it occurs far more often than any of our other verbs) raised the possibility that the data for this one verb might skew our results. To minimize this risk we took a sample, and selected only *the first 1000 dative instances* from the concordances for *give*.

In this study, we look only at dative uses from the "core" meaning of each verb. The verbs are all polysemous, and dative uses can be found in other senses — and in some cases, the dative alternation is as normal as it is in the core meaning. For example:

…you know someone loves you they ***show*** *you love* (V-NP-NP)
… everybody had to ***show*** *allegiance to Hitler* (V-NP-PP)

…they are ***selling*** *you the idea that you can publish….*(V-NP-NP)
…didn't try to ***sell*** *their faith to you…* (V-NP-PP)

But we felt that a focus on core usage would minimize interference from other factors and yield a cleaner set of data to work with. This raises the question of what a core meaning is, and how a given occurrence of a verb can be reliably assigned to a particular sense. Identifying syntactic patterns (V-NP-NP, V-NP-PP) is straightforward, but disambiguating the meanings of a polysemous word is notoriously difficult — not least because the category "word sense" is inherently unstable (e.g. Hanks 2013, 65-83). Having said that, the interplay of patterns and co-text (notably selectional restrictions) provides a robust working basis for word sense disambiguation. In the unsorted data for one of our verbs, for example, we find this dative (V-NP-PP) pattern:

> *…erm who actually is a lecturer and* ***lends credibility to*** *the whole thing*

The V-NP-NP pattern would be equally normal here (*this* ***lends it*** *some* ***credibility***). This does not seem like a core instance of *lend*, and that intuition is supported by the nature of the direct and indirect objects — the "implicatures" as they are termed in Hanks's *Pattern Dictionary of English Verbs* (PDEV; http://pdev.org.uk). In PDEV's analysis of *lend*, the first two constructions listed are identical, and what distinguishes them are the things being lent (in the first case, physical objects or assets, in the second, things such as "weight, credibility, credence, support"). The core sense of *sell* can be similarly identified on the basis of what is being sold: in the FrameNet database, for example, the basic meaning of *sell* belongs to a Frame called "Commerce_sell" (https://framenet.icsi.berkeley.edu/fndrupal/framenet_data), which describes "commercial transactions involving a buyer and a seller exchanging money and goods". This rules out a sentence about someone "selling you *the idea* that you can publish…".

We feel confident that the dative sentences forming the raw material for our study all instantiate core meanings of the selected verbs. The data for *give* represent a possible exception. Not only is *give* highly polysemous, but its frequent use as a light verb (*I'll give you a call*) as well as its appearance in numerous idioms with dative-like structures (*give it a go, give her the benefit of the doubt,* etc.) further complicates matters. Most of the cases where we had doubts were resolved by the non-availability of one or other of the patterns under investigation: *give me a call* is common, but **give a call to me* would be aberrant. So although a few borderline cases may have survived the cut, there are too few to compromise the overall analysis.

Data collection and manual annotation

First of all, we queried the CQPWeb interface of the EAS data to collect all corpus concordances of each of the six verbs we selected. This step (which is described in more detail in the next section) resulted in six text files, each containing all concordance lines of one of the six verbs. Next, we identified only those concordance lines instantiating the core meaning of each verb.

In order to identify instances of the two patterns involved in the dative alternation, it was necessary to add syntactic information to the corpus data. Initially, we tried automatically parsing the concordance lines using the PCFG Stanford Parser (Klein & Manning 2003). However, the quality of the parsed results was poor when using the default parser, particularly on constructions typical of the spoken register, such as *it sort of gives you a list.* For this reason, we opted for manual annotation of the patterns in each concordance line, and filtered the subset of data for each verb to retain only the dative instances of the core meanings. Each line was then manually annotated in a spreadsheet, to indicate which of the two dative patterns it represented, and to show the *recipient* and the *theme*, thus:

> *the easiest way is to sell the print to a big company*

pattern: V-NP-PP; recipient: *a big company*; theme: *the print*

We further enriched the annotation to identify the *head* of the noun phrases representing the recipient and theme. So in the sentence

you can just send Christmas cards … to people you don't see from year to year

the recipient (in full) is “people you don't see from year to year”, but the head is simply “people”. Even without further analysis, this procedure (and the spreadsheets it populated) revealed a number of unmistakeable trends. The two most striking observations (which we discuss more fully later) were: first, the V-NP-NP pattern is the dominant dative form for all the verbs except *sell*, accounting for 69.6% of all dative instances in our data; and secondly, that personal pronouns dominate the recipient column when the pattern is V-NP-NP (*he showed me the letter; no-one's going to give you a job*, etc.). An interesting revelation in the dataset for *send* (though not specifically relevant to the present study) is that the vast majority of cases refer to sending things by electronic media. Although there are a few instances of cards or packages being sent via the postal system, the “theme” column here is mainly populated by words like *email, money, CV, link, photos,* and above all *message*. This marks a striking change from the way the verb is used in the original BNC. The annotation process also showed up instances of a third dative pattern, in sentences like these:

my uncle sold it me for a fiver

[she] showed it me on DVD

remember not to give it him as breakfast cereal

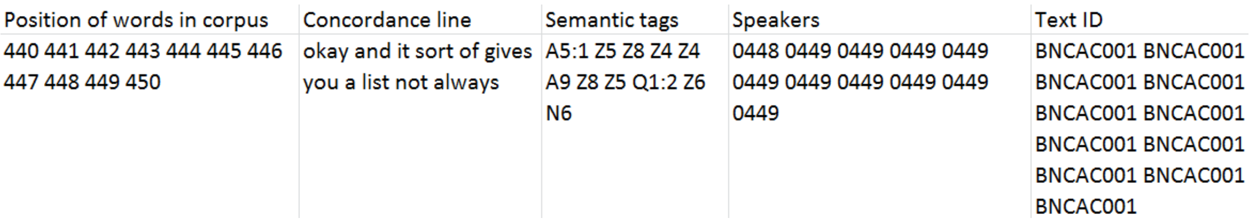
This is a known dialectal variant (in some varieties of British English) on the more usual dative constructions, but as there were fewer than ten instances in the 2000-odd concordance lines in our sample, its presence in the data was unlikely to affect the analysis and such instances were excluded from the final multivariate analysis.

Data processing

In order to conduct the sociolinguistic analysis for this study, it was necessary to gather the relevant corpus data and metadata in a format that allowed further statistical processing. For this reason, we defined a pipeline to process the data automatically; this ensures future reproducibility and replicability of our results. In this section, we describe the data processing pipeline.

1. Export of concordance lines

First of all, we queried the CQPWeb interface of the EAS data to collect all corpus concordances of each of the six verbs we selected. The corpus query specified the lemma and the part of speech "verb". For example, to retrieve all concordances for the target verb *give*, we used the query "{give/V}". We selected the widest context size allowed, which consists of 50 words before and 50 words after the target word (option "50 words each way"). Finally, we downloaded the concordances together with all the corpus metadata fields available (option "Method: Download all text metadata"). This step resulted in six text files, each containing all concordance lines of one of the six verbs. Figure 9.1 shows the first line of the file for *give*.

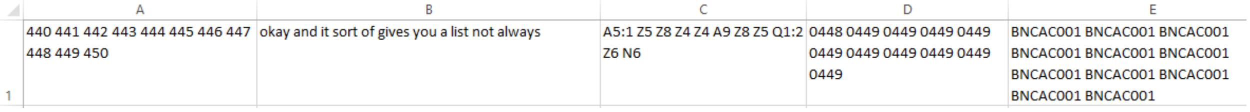


Position of words in corpus	Concordance line	Semantic tags	Speakers	Text ID
440 441 442 443 444 445 446 447 448 449 450	okay and it sort of gives you a list not always	A5:1 Z5 Z8 Z4 Z4 A9 Z8 Z5 Q1:2 Z6 N6	0448 0449 0449 0449 0449 0449 0449 0449 0449 0449 0449	BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001

Figure 9.1: Screenshot of the file containing the first corpus concordance line for *give* extracted in the first step of the pipeline.

2. Export of corpus metadata

The concordance files contain information about the corpus text ID, the target verb, its right and left context (both as raw text and with part-of-speech tagging), the date on which the conversation was recorded, its length and its location, the number and identifiers of the speakers, their relationship, the subject and topics covered in the conversation, the url corresponding to the concordance line, and its position in the corpus. In order to conduct our analysis, we needed to retrieve the corpus semantic annotation, which follows the definitions of the English semantic tagger of the UCREL semantic analysis system. Therefore, the second step consisted in exporting the metadata from the corpus via the CQPWeb interface. Again, we queried the corpus specifying the verb lemma and its part-of-speech; then, we selected the option "Download query as plain-text tabulation", and chose an offset of -5 and 5 for each of the following fields from the drop-down menu: "Corpus position number", "word", "semtag", "u_who", and "text_id". This allowed us to record various types of information about each of the words occurring in the concordance line for the verbs of interest and within a window of size five, which is the widest context allowed. For each such context word, the information recorded concerned its position in the corpus, its form, its semantic tag, the identifier of the speaker uttering it, and the identifier of the text in which it occurred. Figure 9.2 shows an example of such metadata file for the verb *give.*

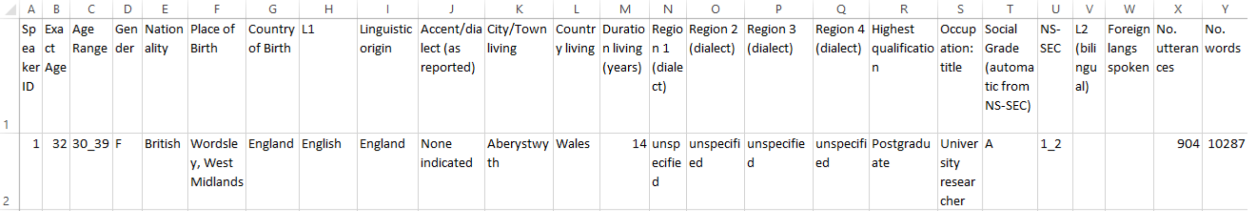


	A	B	C	D	E
1	440 441 442 443 444 445 446 447 448 449 450	okay and it sort of gives you a list not always	A5:1 Z5 Z8 Z4 Z4 A9 Z8 Z5 Q1:2 Z6 N6	0448 0449 0449 0449 0449 0449 0449 0449 0449 0449 0449	BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001 BNCAC001

Figure 9.2: Screenshot of the file containing the corpus metadata for the first concordance line of *give* extracted in the second step of the pipeline.

3. Collect and clean speaker data

Further, our analysis required more details about the speakers, so the next step in the data collection phase consisted in gathering and cleaning the speaker data. These were delivered to us as an Excel spreadsheet, whose rows correspond to each speaker and whose columns record the following information: identifier, exact age, age range, gender, nationality, place of birth, country of birth, native language, linguistic origin, accent/dialect, city/town of residence, country of residence, the number of years the speaker has spent living there, four different categorizations of the region of residence, highest qualification, occupation, social grade, second native language for bilingual speakers, foreign languages spoken, number of utterances, and number of words uttered in the corpus. Figure 9.3 shows an example of the first speaker's data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Speaker ID	Exact Age	Age Range	Gender	Nationality	Place of Birth	Country of Birth	L1	Linguistic origin	Accent/dialect (as reported)	City/Town living	Country living	Duration living (years)	Region 1 (dialect)	Region 2 (dialect)	Region 3 (dialect)	Region 4 (dialect)	Highest qualification	Occupation: title	Social Grade (automatic from NS-SEC)	NS-SEC	L2 (bilingual)	Foreign langs spoken	No. utterances	No. words
2	1	32	30_39	F	British	Wordsley, West Midlands	England	English	England	None indicated	Aberystwyth	Wales	14	unspecified	unspecified	unspecified	unspecified	Postgraduate	University researcher	A	1_2			904	10287

Figure 9.3: Screenshot of the file containing the data for the first speaker processed in the third step of the pipeline.

The speaker data contained several inconsistencies, for example the column "Duration living (years)" indicates the number of years that the speaker has spent living in the location indicated, and in the overall dataset it contains values such as "30 years", "30", and also "all my life" and "all her life". This is probably due to the fact that the data were entered manually. In order to enable the subsequent statistical analysis, we decided to clean these data by replacing inconsistent values such as the ones listed above, with more consistent ones.

4. Combine corpus data with metadata

Finally, we wrote a Python script that combined the details from the concordance files with the metadata files and the speaker information in one file per verb. This was followed by another script (written in R) which merged the individual datasets for each verb into a single dataset and cleaned some of the variables and defined new ones. The values of the speaker's education level were mapped to a set of consistent categories and semantic tags were mapped to their least granular class; the additional variables we defined are: length of theme/recipient, pronominality of theme/recipient, definitiveness of theme, animacy of theme/recipient.

Analysis of corpus data

As mentioned earlier, the V-NP-NP pattern is dominant for all the verbs except *sell* (where V-NP-PP outnumbers V-NP-NP by almost three to one). Table 9.2 shows the distribution of the two patterns for our six verbs:

Verb	Datives	V-NP-NP	V-NP-PP
give	1000	882 (88%)	118 (12%)
lend	38	27 (71%)	11 (29%)
offer	72	62 (86%)	10 (14%)
sell	103	26 (25%)	77 (75%)
send	570	436 (76%)	134 (24%)
show	276	242 (88%)	34 (12%)

Table 9.2: distribution of the two dative patterns in the data for the six verbs

Our central research question is: what factors influence a speaker to select one pattern rather than the other? There are plenty of possibilities, and the EAS data allow us to look at a wide range of linguistic and sociolinguistic variables. To assess the effects of these features, we conducted a multivariate analysis, and we discuss the output of this model in the next section.

Here, we will give an overview of the different factors considered. The dataset comprises 1,938 observations, corresponding to occurrences of the six verbs in the corpus. Each observation is characterized by a range of attributes, which we can group in the following categories:

- Syntactic, semantic, lexical features: verb lemma; syntactic realization of the dative construction (either as V-NP-NP or V-NP-PP), lexical realization of the recipient, lexical realization of the theme, syntactic head of the recipient, syntactic head of the theme, semantic tag of the syntactic head of the recipient phrase, semantic tag of the syntactic head of the theme phrase, semantic tag of the verb occurrence in context, length of recipient, length of theme (both measured as number of characters), length of theme, pronominality of recipient, pronominality of theme;

- Corpus metadata: number of speakers, location, relation, topics;
- Speaker metadata: exact age, age range, gender, nationality, place of birth, country of birth, L1, bilingualism; linguistic origin, accent, city of residence, country of residence, number of years they have lived in the city of residence, level 1 dialect, level 2 dialect, level 3 dialect, level 4 dialect (where a speaker's dialect is identified with varying degrees of granularity), highest qualification, occupation, foreign languages spoken, number of utterances, number of words.

We analyzed the different variables in order to assess their quality and relevance to the phenomenon under study, as detailed below. As a consequence, we decided to exclude the following variables from further analysis: bilingualism (because the large majority of speakers are monolingual)[1], level 1, 2, and 3 dialect, accent, and city/town (because they were too granular and the data were too sparse), nationality (due to the overwhelming proportion of speakers categorized as British, 99% of the data points), L1 (due to the overwhelming proportion of the value "English", 97% of the data points), country (due to the overwhelming proportion of speakers from England, corresponding to 91% of the data points), number of years they have lived in the city of residence (all values are not-applicable), place of birth, country of birth, and linguistic origin (due to the inconsistency of the annotation), occupation (too many values and

[1] We also excluded the 98 observations whose speaker was bilingual because they represented only 0.05% of the total, which is a number not sufficiently high to show any effect.

inconsistent coding), foreign languages, relation, topics, and location (too much variation and inconsistent coding).

Exploratory analysis

In order to answer our research questions, we conducted a preliminary exploratory analysis of the rich dataset we had collected, and this is the focus of the current section, which aims to show if there is any correlation between an individual feature and either of the two patterns. This analysis gave us insights into the nature of the data we worked with, and informed our later investigations on the statistical models reported on in the next section.

1. Speaker's age

Focussing on sociolinguistic variables first, the speaker's *age* appears to have no effect on their preference for either pattern, as confirmed by a Welch 2 sample t test (t = -1.79, degrees of freedom = 460.675, p-value > 0.05). If we consider another age-related metadata field, namely *age range*, with values 11-18, 19-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99, we see no significant difference in the choice of syntactic pattern by age range category (chi-square = 10.9223, degrees of freedom = 8, p-value > 0.05).

2. Speaker's level of education

With regard to people's level of education, there is some evidence that that speakers with postgraduate degrees are a little more likely than others to use a V-NP-PP pattern: 24% of the instances with a speaker with postgraduate degree display this pattern, as opposed to 15% of those with a college/6th form education, 18% for graduates, and 21% for those without post-GCSE qualifications. A Pearson's chi-square test showed a significant association between pattern and education level (chi-square = 14.9012, degrees of freedom = 3, p-value < 0.05), but a small effect (Cramér V = 0.088).

3. Speaker’s gender

There appears, also, to be a small gender effect, with male speakers showing a slight preference for V-NP-PP when compared with female speakers: V-NP-PP occurs in 21% of the instances with male speakers and in 19% of the instances with female speakers. However, this difference is not statistically significant, as confirmed by a Pearson’s chi-square test (chi-square = 0.8198, degrees of freedom = 1, p-value > 0.05).

4. Region

We analyzed the distribution of the two patterns by region, according to the level 3 of granularity, which includes the following categories: Midlands, non UK, North, Scottish, South, Unspecified, and Welsh. There does not seem to be any major difference between these different regions in the way the two patterns are distributed. Table 9.3 shows the proportion of the two patterns by region, and the absolute frequencies. A Pearson’s chi-square test showed no significant association between region and pattern (chi-square, degrees of freedom = 6, p-value > 0.05).

Region	**V-NP-NP**	**V-NP-PP**
Midlands	64 (74%)	22 (26%)
Non UK	9 (82%)	2 (18%)
North	371 (84%)	72 (16%)
Scottish	2 (100%)	0 (0%)
South	755 (78%)	210 (22%)
Unspecified	249 (82%)	54 (18%)
Welsh	18 (78%)	5 (22%)

Table 9.3: distribution of dative patterns according to speaker’s region

5. Length of arguments

Coming to the linguistic variables, there is strong evidence that the *length* of the arguments (recipient and themes) predicts the selected pattern in many cases. Broadly speaking, where the recipient is instantiated by a short NP (such as *me, my mum, the boys*, or a personal name), the V-NP-NP pattern is preferred:

we gave him a drink of water

take you and show you our wonderful country

I sent my nan a postcard from Barcelona

The converse is also generally true: a longer recipient tends to imply the V-NP-PP pattern.

probably never show it to yeah any males in your life

you can just send Christmas cards … to people you don't see from year to year

By the same logic, a longer theme predicts V-NP-NP:

[name] lent me a ski jacket, the helmet, the skis and the poles

But they offer me a cup of coffee and a biscuit

and sell you a picture of your house from the air

But where both recipient and theme are realized by very short words, V-NP-PP tends to be preferred. Taking account of the far higher number of V-NP-NP instances, there is a marked bias towards V-NP-PP sentences like these:

give it to him

I won't lend it to you

I'll show that to [name]

Figure 9.4 shows the relationship between the selected pattern and the length of the theme and recipient:

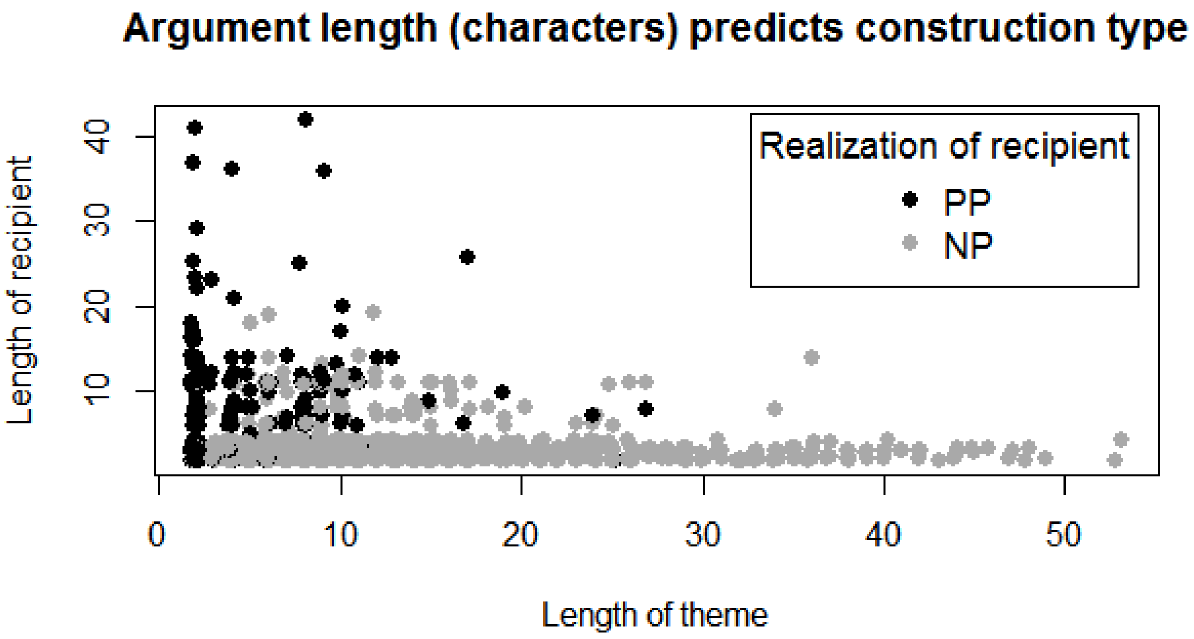


Figure 9.4: chart showing the length of theme plotted against length of recipient for the two patterns

6. Pronominality of arguments

It will be obvious from the previous examples that a high percentage of the short themes and recipients are pronouns. Where the pattern is V-NP-NP, the recipient is overwhelmingly likely to be a personal pronoun, with *you* and *me* being especially common. In the case of *show*, for example, a mere 18 of the 242 V-NP-NP instances are *not* personal pronouns. In V-NP-PP sentences, the theme (which is typically inanimate), is frequently realised by the pronouns *it* or *them*, as in the following examples:

> *you sent it to a hundred and thirty people*
>
> *sort out all our books and give them to that [name] book shop*

But the distribution of pronouns (personal for recipients, impersonal for themes) is not symmetric. The V-NP-NP pattern overwhelmingly prefers a personal pronoun in the recipient slot (93% of cases) - or to put it another way, the choice of a pronoun as recipient overwhelmingly predicts the V-NP-NP pattern; on the other hand, this pattern only has 9% of pronominal themes. For V-NP-PP sentences, the situation is more complicated: here, pronominal recipients account for 56% of the cases, and pronominal themes are 65% of the cases. Chi-square tests on the distribution of pronominal recipients and themes by patterns show that these differences are significant and have a medium-sized effect (pronominality of recipient: chi-square= 313.1513, degrees of freedom = 1, p-value < 0.05; Cramér V = 0.42; pronominality of theme: chi-square = 546.6041, degrees of freedom = 1, p-value < 0.05; Cramér V = 0.55).

7. Semantic tag of recipient and theme

The most frequent semantic tag for recipients by a large margin is Z8, which corresponds to pronouns; the same holds for themes, although in this case the distribution of semantic tags is less skewed.

If we ask whether there is a significant difference in the way the two patterns are distributed by semantic tag of the recipient, we find that this is not the case (chi-square = 14.4643, degrees of freedom = 12, p-value > 0.05).

However, if we consider the semantic tag of the theme, we find a significant difference. A chi-square test shows a medium-sized statistically significant difference in the distributions of the two patterns by semantic tag (chi-square = 616.0445, degrees of freedom = 92, p-value < 0.05, Cramér V = 0.577); generally, the V-NP-NP pattern is more common than V-NP-PP, with the exception of the theme semantic tags “Crime, law and order” and “Pronouns, etc.”, for which prepositional recipients are more common than non-prepositional ones.

Statistical multivariate models

For the multivariate analysis, we used a binary logistic mixed-effects model fitted in R with the lme4 package (Bates et al. 2008). Such models are now regularly used in both Corpus

Linguistics and Sociolinguistics, see e.g. Baayen (2008) and Tagliamonte and Baayen (2012). The response was a binary variable indicating the V-NP-NP pattern (0/False) or V-NP-PP (1/True). A binary logistic regression model attempts to model the probability of the response value (essentially switching from 0 to 1, or False to True) as a function of a series of linear combinations of predictor variables. For technical reasons, the probabilities are estimated on a logarithmic scale as odds ratios. Unlike fixed-effects regression models, mixed-effects models are generally better at dealing with specific sampling biases. The advantage of mixed-effects models over simpler fixed-effects generalized linear models is that adjustments for known biases in the data can be made explicit in the model itself. In this case, we used a random effect for verbs, thus adjusting for different overall frequencies and different rates of construction use in the verbs. Another reason for choosing a random effect for verb is that, unlike variables such as gender or social class where we have covered all relevant values, our sample only covers a subset of all verbs (Baayen 2008, 241). For the multivariate regression analysis, we employed a smaller subset of the data encompassing only complete cases, totalling 1602 observations.

The first step involved identifying a model that converged. In practice, this means correctly specifying the list of predictors (also variously known as fixed effects or independent variables) and the random effect(s) in a manner that correctly predicts the response, in our case the realization of the recipient as either a PP or NP. In this step we removed a number of variables, notably semantic variables such as semantic tag or semantic field of the theme, and the sociolinguistic variables of dialect and speaker’s social grade. The problem with these categorical variables was that the model did not converge due to estimated probabilities close to one or zero, a problem that can occur when a factor value is rare, but when it appears it perfectly predicts the response (Venables and Ripley 2002, 198-199). Put differently, a number of rare semantic tags and sociolinguistic variable values always predicted one response value. With no variation, the model could not correctly estimate the required parameter values. Another way of viewing this is that in our dataset these variables, the semantic field of the theme or the dialect, might contain values that are good predictors of certain response values (NP or PP), but they are not overall good predictors of the variation between the two response values. In the case of the speaker's social grade, we found a large correlation with the speaker's highest qualification (= 844.5, degrees of freedom = 15, $p < 0.05$, Cramér V = 0.39). Hence, we decided to omit the social grade variable since exploratory testing indicated that the highest qualification was a better predictor for the response.

Having omitted these variables, we arrived at a full model that we could use as a starting point for our analysis. The maximal converging model has the following structure:

Response: Probability of V-NP-PP

Random effect: Verb

Fixed effects: Gender + RecPrn + ThemePrn + log(RecLen) + log(ThemeLen) + DefTheme + AnimateRec + AnimateTheme + SpeakerHighestQual + AgeImputed + SpeakerHighestQual X AgeImputed

The numeric variables length of theme (ThemeLen) and length of recipient (RecLen) were logarithmically transformed to better adhere to the model assumption of normally distributed numerical predictors. Furthermore, we had to add an interaction term between the age and highest qualification of the speaker, since these two variables are sufficiently closely related to create problems for the model specification if no interaction term is specified. The predictor "AgeImputed" corresponds to the speaker's age, when that is available, and to the midpoint of the speaker's age range, when the exact range was not available. We also defined the variables

“DefTheme” (theme expressed as a definite phrase) and ‘’AnimateRec" (animacy of recipient). “RecPrn” and “ThemePrn” refer to the pronominality of the recipient and theme, respectively. The model’s predictors are shown in Table 9.4.

Predictors	**Coef β**	**SE(β)**	**z**	**p**
(Intercept)	-0.63	1.08	-0.6	>0.6
GenderM	0.48	0.24	2.0	**<.05**
RecPrnTRUE	-2.05	0.42	-4.9	**<.0001**
ThemePrnTRUE	1.90	0.30	6.4	**<.0001**
log(RecLen)	1.39	0.30	4.7	**<.0001**
log(ThemeLen)	-2.11	0.22	-9.4	**<.0001**
DefThemeTRUE	-0.12	0.29	-0.4	>0.7
AnimateRecTRUE	0.76	0.32	2.4	**<.05**
AnimateThemeTRUE	1.54	1.01	1.5	>0.1
SpeakerHighestQualGraduate	1.61	0.69	2.3	**<.05**
SpeakerHighestQualPostgraduate	1.94	0.82	2.4	**<.05**
SpeakerHighestQualSecondary School	1.24	0.90	1.4	>0.2
AgeImputed	0.03	0.01	2.2	**<.05**
SpeakerHighestQualGraduate:AgeImputed	-0.03	0.02	-2.1	**<.05**
SpeakerHighestQualPostgraduate:AgeImputed	-0.03	0.02	-1.9	>0.1
SpeakerHighestQualSecondary School:AgeImputed	-0.02	0.02	-1.0	>0.3

Table 4: coefficients, standard errors, z-values, and p-values for the selected variables.

The table shows the name of the predictor, the size of the effect (coefficient), the uncertainty of the effect (coefficient standard error), the z-value used to determine statistical significance, and the p-value. Statistically significant predictors (below the conventional 0.05 threshold) are highlighted in bold. The coefficient represents the chance of switching from an NP recipient to a PP recipient, expressed as an odds ratio on a logarithmic scale. A positive number indicates an increase in PP recipients with the predictor, whereas a negative number indicates an increase in NP recipients; a value of zero equates to even odds. Odds ratios are not intuitively interpretable, so we proceed with discussing them as probabilities. Log odds ratios can be transformed into probabilities by taking the inverse logit function of the coefficient together with the intercept. A more convenient approach is what Gelman and Hill (2007, 82) call the “divide by 4 rule”. To interpret the coefficients on a probability scale where it represents the midpoint of the logistic curve, we can simply divide the coefficient by four. Based on this heuristic, we see that out of the coefficients that are statistically significant, the linguistic variables pronominality of theme and recipient, and length of theme and recipient stand out. Of the sociolinguistic variables, we see that a graduate or postgraduate qualification lean towards a PP recipient, as does male gender, albeit with a smaller effect. Some predictors, viz. definiteness of theme and animacy of theme, are not statistically significant.

However, this full model is not necessarily the best one. The aim is to find a model that explains the data in the simplest manner, without adding any unnecessary variables (Faraway 2005, 121). A simple model for our purposes is one that, in accordance with Occam’s razor, will both fit the data and adequately describe the response variable (V-NP-NP vs. V-NP-PP) with as few predictors as possible. The criteria used for finding an optimal model involved using log-likelihood ratio tests to identify any significant difference between the smaller and the larger model, while at the same time maintaining a high C-index value. The C-index is a measure of how well the model predicts the data, and a C-value of 0.95 indicates an excellent fit (Baayen

2008, 204). Finally, we inspected plots of the model residuals to ensure that they did not show signs of serious problems with the model structure (Faraway 2005, 53-56; Gelman and Hill 2007, 97-98).

Based on these criteria, we found that the variables definiteness of theme, animacy of theme, speaker's highest qualification, and speaker age, could all be eliminated. The final model has the following structure:

Response: Probability of V-NP-PP

Random effect: Verb

Fixed effects: Gender + RecPrn + ThemePrn + log(RecLen) + log(ThemeLen) + AnimateRec

Table 5 summarizes the fixed effects of the model:

	Coef β	**SE(β)**	**z**	**p**
(Intercept)	0.67	0.94	0.7	>0.5
GenderM	0.56	0.22	2.5	**<.05**
RecPrnTRUE	-1.97	0.41	-4.8	**<.0001**
ThemePrnTRUE	1.99	0.29	6.9	**<.0001**
log(RecLen)	1.40	0.30	4.7	**<.0001**
log(ThemeLen)	-2.04	0.21	-9.8	**<.0001**
AnimateRecTRUE	0.69	0.31	2.2	**<.05**

Table 9.5: coefficients, standard errors, z-values, p-values for the selected predictors.

From Table 9.5, we see that the effect of gender is 0.56. Dividing it by 4, we obtain a quick estimate of the effect on a probability scale (Gelman and Hill 2008, 82). Men are 14% more likely to use a V-NP-PP construction than women, after controlling for differences between verbs. The coefficient for pronominality of recipients is -1.97, meaning that pronoun recipient reduce the use of V-NP-PP constructions by about 50%. Conversely, the coefficient for pronominality of theme is 1.99, implying an estimate of a 50% increase in the use of V-NP-PP constructions. The coefficient for the logarithmically transformed length of recipients, 1.4, is 0.35, meaning that a one unit increase in recipient length (on a log scale), results in a 35% increase in the use of V-NP-PP. For the logarithmically transformed length of themes we see the opposite tendency. The coefficient of -2.04 suggests a 50% decrease in use of V-NP-PP for every one unit increase. Finally, animate recipients have a 17% higher rate of use of V-NP-PP constructions compared to inanimate recipients.

Figure 9.5 visualizes the fixed effects of the final model. The predictors on the right hand side of the dotted midline are associated with a higher probability of PP recipients. The horizontal bars extending from the points are 95% confidence intervals.

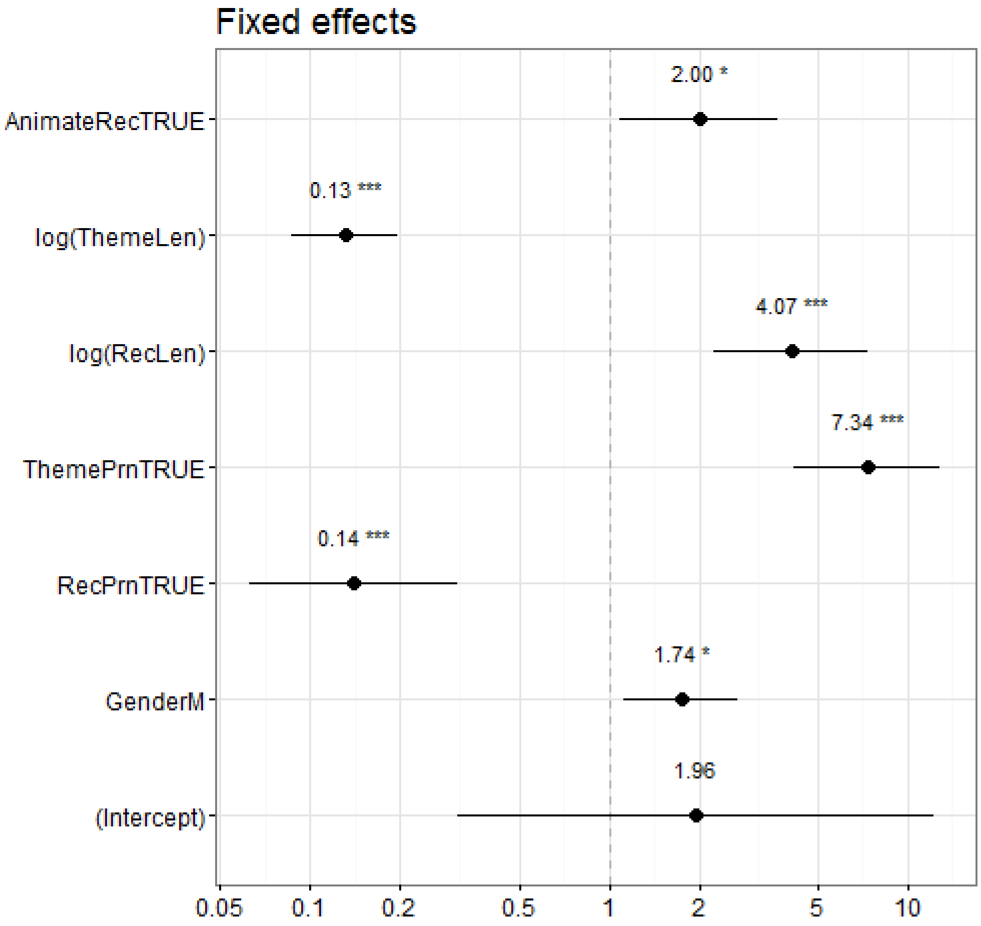


Figure 9.5: Predictors for the minimal model, with 95% confidence intervals for the coefficient estimates. Estimates to the right hand side of the midline are associated with higher probability of PP recipients.

The random effect variable, verbs, has a standard deviation of about 0.84, which translates into an average difference between verbs in their use of the V-NP-PP construction of about plus/minus 21%. In other words, although there are real differences between verbs in how often they occur with the constructions, the model takes this into account, so that the fixed effects in Table 9.5 are the effects of our predictors over and beyond the verb specific effects. Figure 9.6 below shows the random effects, with 95% confidence intervals. *Show* and *give* are notable for their preference for NP recipients, whereas *sell* (as expected from the exploratory analysis) is associated with PP recipients. *Send*, *offer*, and *lend* do not differ greatly from the overall average, as indicated by the overlaps between the intercept and the confidence intervals.

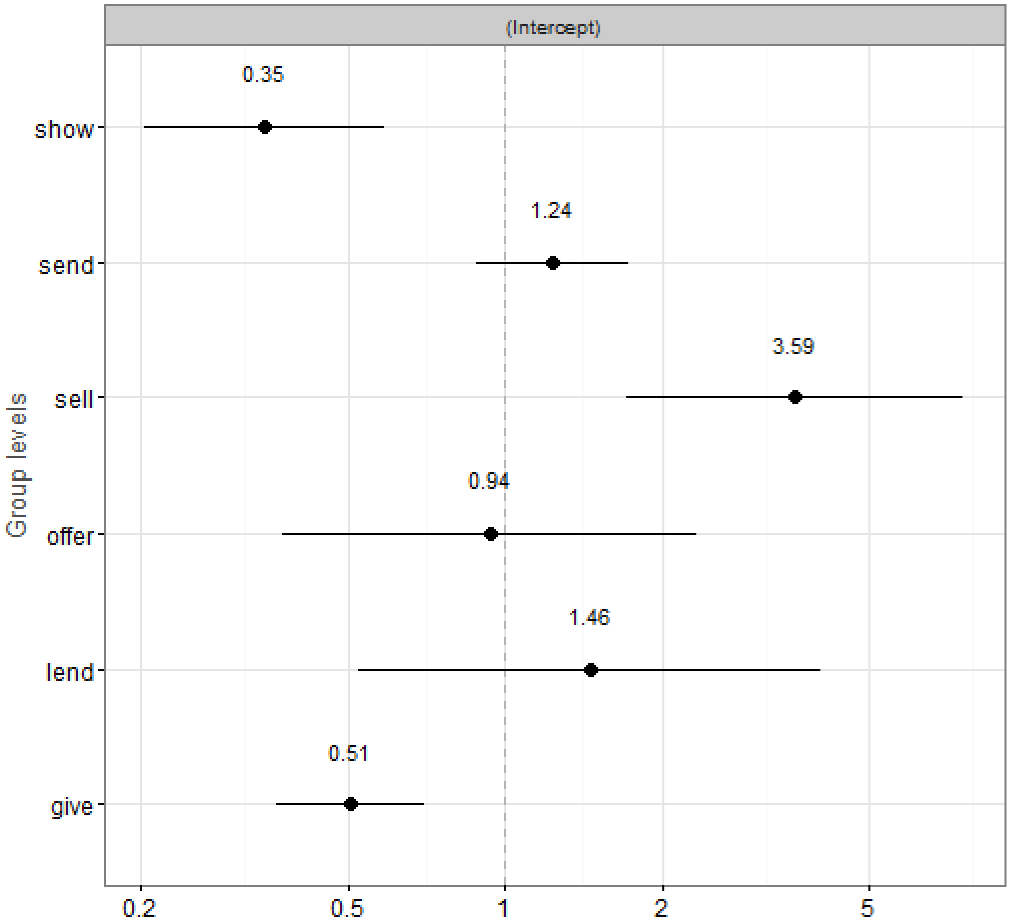


Figure 9.6: Random effects for the minimal model, with 95% confidence intervals. *Sell* is associated with a higher probability of PP recipients, whereas *offer*, *lend*, and *send* are not very different from the overall average.

Sociolinguistic implications and conclusions

Previous studies have investigated the dative alternation from a sociolinguistic perspective, and identified probabilistic differences in the dative alternation at the sociolinguistic macrolevel, e.g. between American and New Zealand English. Moreover, some of the previous research (Bresnan et al. 2007; Bresnan and Hay 2008; Bresnan and Ford 2010) has used spoken data from recordings of phone conversations between strangers. In this study, we used spontaneous spoken data recorded from face-to-face conversations between speakers who know each other. This has given us the opportunity to test whether sociolinguistic features affected the dative alternation in the Spoken BNC Corpus and whether data from this resource could confirm previous results. Our analysis showed the effects of grammatical features on the dative alternation. Firstly, our model downplays the effect of recipient's animacy compared to the model in Bresnan et al. (2007) based on US English. Furthermore, while Bresnan et al. (2007) found a strong effect for indefiniteness of theme, this variable was omitted from our model for reasons of parsimony. What both our model and Bresnan et al.'s share are strong effects for the length of theme and argument, even though we used the log transformed number of characters as opposed to Bresnan et al.'s number of words. The pronominality of recipient also plays a large role in both models. In addition to this, we found a strong effect for the pronominality of the theme. Our results point strongly towards a construction that is modulated in accordance with Behagel's Law, with the concomitant implications for information status (Arnold et al. 2000). This result is in line with previous research showing how discourse features and processing constraints tend to play a lead role in shifting the probabilities for which variant of the construction is used (Arnold et al. 2000; Bresnan et al. 2007; Bresnan and Hay 2008; Bresnan and Ford 2010; Jenset and Johansson 2013).

With regard to the effects of sociolinguistic features, our main finding is two-fold. First, as did Bresnan and Hay (2008), we identified some subtle probabilistic differences between our British data and previous results from other macro-variants of English. As Kendall et al. (2011, 230) point out, this does make the construction a sociolinguistic variable, since aspects of its realization correlate with non-linguistic features. Secondly, at the level of the individual speaker our results point to a somewhat weaker role played by sociolinguistic variables compared to the grammatical ones, also as expected based on previous studies (Bresnan and Hay 2008; Bresnan and Ford 2010; Kendall et al. 2011). The fact that our models found effects for variables such as gender, age, and qualifications could be a feature of the previously mentioned macro-level variation, or it might be a result of the unique data based on spontaneous face to face conversation. Although we identified a model which indicated that speakers with graduate and postgraduate qualifications use the V-NP-PP construction more than those with lower qualifications, with a weak interaction with speaker's age, we eventually opted for a simpler model, where the only strictly sociolinguistic variable that remained is gender. We found that male speakers in the dataset tend to prefer the prepositional realization of recipient, but further investigations would be necessary to assess whether any interaction with other factors is at play, for example conversation topics. The possibility of such confounding effects based on topic is also a possibility with the variables age and qualification. Unfortunately, some of the metadata such as topics or relation between speakers were encoded at such a granular level and displayed such a high degree of inconsistency that we were not able to incorporate them in the model and test their effect. We hope that future research will be able to address these points and add further insights to our findings.

References

Arnold, Jennifer E, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. "Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering." *Language* 76 (1): 28–55.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bates, Douglas, Martin Maechler, and Bin Dai. 2008. *lme4: Linear Mixed-Effects Models Using S4 Classes*. http://lme4.r-forge.r-project.org/.

Bayley, Robert. 2002. "The Quantitative Paradigm." In *The Handbook of Language Variation and Change*, edited by J. K Chambers, Peter Trudgill, and Natalie Schilling-Estes, 117–41. Malden, MA.: Blackwell.

Bresnan, Joan, A. Cueni, T. Nikitina, and R. Harald Baayen. 2007. "Predicting the Dative Alternation." In *Cognitive Foundations of Interpretation*, edited by G. Bouma, I.

Kraemer, and J. Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, Joan, and Marilyn Ford. 2010. “Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English.” *Language* 86 (1): 168–213.

Bresnan, Joan, and Jennifer Hay. 2008. “Gradient Grammar: An Effect of Animacy on the Syntax of Give in New Zealand and American English.” *Lingua* 118 (2): 245–259.

Faraway, Julian J. 2005. *Linear Models with R*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge: Cambridge University Press.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. “SWITCHBOARD: Telephone Speech Corpus for Research and Development.” In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1:517–520. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=225858.

Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Jenset, Gard B., and Christer Johansson. 2013. “Lexical Fillers Influence the Dative Alternation: Estimating Constructional Saliency Using Web Document Frequencies.” *Journal of Quantitative Linguistics* 20 (1): 13–44. doi:10.1080/09296174.2012.754597.

Kendall, Tyler, Joan Bresnan, and Gerard Van Herk. 2011. “The Dative Alternation in African American English: Researching Syntactic Variation and Change across Sociolinguistic Datasets.” *Corpus Linguistics and Linguistic Theory* 7 (2): 229--244.

Klein, Dan, and Christopher D. Manning. 2003. “Accurate Unlexicalized Parsing.” In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 423–30. Sapporo, Japan: Association for Computational Linguistics.

Köhler, Reinhard. 1999. “Syntactic Structures: Properties and Interrelations.” *Journal of Quantitative Linguistics* 6 (1): 46–57. doi:10.1076/jqul.6.1.46.4137.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna. http://www.r-project.org.

Tagliamonte, Sali A., and R. Harald Baayen. 2012. “Models, Forests, and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice.” *Language Variation and Change* 24 (2): 135–78. doi:10.1017/S0954394512000129.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer.