# Modelling the English dative construction in varied written and spoken text

*Daphne Theijssen, Nelleke Oostdijk, Hans van Halteren, Lou Boves*
*Radboud University Nijmegen*

## Introduction

Over the past decades, linguistic theorists have attempted to design sets of deterministic rules that account for all-and-only the sentences of a language that are deemed 'grammatical'. However, intuitions about what constructions are (im)possible virtually always appear to be at odds with usage data (e.g. Chater and Manning 2006). The problems related to graded grammaticality and coverage have resulted in probabilistic approaches to linguistics, which consider grammaticality as a function that can take values between 0 (categorically ungrammatical) and 1 (no other option available), with most values being between the two extremes.

There are situations where speakers can choose between several options that are equally grammatical, but that may differ in their acceptability in the given context. An example is the dative construction in English, for which speakers and writers can choose between structures with a double object (NP-NP, e.g. *She handed the student the book.*) or prepositional dative structure (NP-PP, e.g. *She handed the book to the student.*). What we need is yet another kind of descriptive model, which can explain such choices on the basis of a (potentially large) number of linguistic, paralinguistic and extralinguistic properties of a sentence or a paragraph in written texts and their discourse equivalents in spoken language.

Until recently, the development of such models has been hampered by the lack of advanced statistical techniques that can deal with phenomena such as syntactic structures and their elements. Fortunately, linguistics can profit from recent advances in what used to be called nonparametric statistics, where powerful models have been and are being developed for handling this type of variables. The models with which Bresnan et al. (2007) explain the selection between the two dative constructions in English represent arguably the most advanced attempt today to show that the choice between the two options can be explained by way of a combination of (para-)linguistic factors.

In the present research we also aim at modelling the dative alternation, building on Bresnan et al.'s (2007) work. Since their source data is not available due to restrictions on the additions and corrections to the Switchboard Corpus they applied (Bresnan, personal communication), we are forced to create a new data set. This enables us to apply the linguistic features and the statistical modelling techniques they used to data that shows more variation in text genre. Also, we attempt to improve the model by adding new features that we believe are relevant for explaining the variation. Since the research is still in progress, this abstract will only describe our methods, while the results will be presented at the workshop.

## Varied written and spoken text

The larger part of Bresnan et al.'s (2007) article concerns transcribed spoken data from the Switchboard Corpus. The model explains 94% of the dative alternation in previously unseen data. They extended the data with instances from the Wall Street Journal texts in the Penn Treebank and concluded that the found model for the spoken data generalizes to written data.

The variety in Bresnan et al.'s data, however, is very narrow. The spoken data contains spontaneous conversations on fixed topics solely, and the written data consists only of financial newspaper articles. Therefore we investigate whether, and if so how, an increase in the range of text and discourse types affects the quality of the model. For this purpose, we employ the syntactically annotated ICE-GB Corpus (Greenbaum 1996). The corpus consists of one million words in British English and contains spoken dialogues (private and public) and monologues (unscripted and scripted), and written texts that are non-printed (student writing and letters) and printed (academic, popular, reportage, instructional, persuasive and creative).

With the help of a Perl script, we automatically extracted sentences with an indirect and a direct object (NP-NP) and sentences with a direct object and a prepositional phrase with the preposition *to* (NP-PP).

The found instances have subsequently been manually checked to filter irrelevant structures such as (1a), which contains a locative *to*-PP instead of a prepositional dative construction. For the present research, we ignore constructions with prepositions other than *to*, with coordinated verbs or verb phrases, with phrasal verbs, and with passive voice. Also, we remove all instances with verbs that are present in instances with only one of the two dative constructions. Characteristics of the resulting data set can be found in Table 1.

(1) a.  *Fold the short edges to the centre.*                              (ICE-GB W2D-019_144:1)
     b.  *\*Fold the centre the short edges.*

Table 1. Characteristics of our data set

|  | *Spoken* | | *Written* | | *Total* |
| number of | Dialogues | Monologues | Non-printed | Printed | |
| --- | --- | --- | --- | --- | --- |
| texts | 180 | 120 | 50 | 150 | 500 |
| words | 360,000 | 240,000 | 100,000 | 300,000 | 1,000,000 |
| NP-NP | 433 | 222 | 133 | 214 | 1002 |
| NP-PP | 84 | 53 | 31 | 52 | 220 |
| NP-NP / texts | 2.4 | 1.9 | 2.7 | 1.4 | 2.0 |
| NP-PP / texts | 0.5 | 0.4 | 0.6 | 0.3 | 0.4 |

One of the linguistic features applied by Bresnan et al. (2007) is the semantic class of the verb: 'abstract' (e.g. *give it some thought*), 'transfer of possession' (e.g. *send*), 'future transfer of possession' (e.g. *promise*), 'prevention of possession' (e.g. *deny*) and 'communication' (e.g. *tell*). In the example in the introduction, two noun phrases are important: *the book* (what has been given) and *the student* (who it has been given to). Bresnan et al. call these the 'theme' and the 'recipient', respectively. For both theme and recipient, the discourse accessibility is established as are the pronominality, the definiteness, the animacy, the person, the number and the concreteness (the latter only for the theme). Discourse accessibility is defined as 'given' or 'not given' in the preceding context, or 'accessible' to the addressee. Also, they checked which construction (NP-NP or NP-PP) has been used previously in the dialogue, resulting in the feature 'structure parallelism in dialogue'. Lastly, the length difference between the theme and the recipient is added to the model (log scale). The features 'person of theme' and 'animacy of theme' were removed from Bresnan et al.'s research since they were too sparse. We will follow a similar approach in which we include all features unless they appear to be too infrequent in our data to base conclusions on them. All feature values will be manually determined to reduce the risk of erroneous data.

The statistical modelling techniques Bresnan et al. (2007) apply are Linear Regression Modelling and Generalized Linear Mixed Modelling. The latter is a generalization of the former, in which random effects can be included in the predictor. This results in a model that reveals correlations between the feature effects. Bresnan et al. employ this technique in order to establish the correlation between the verb sense and the other features. We will build similar models for our data set and evaluate the results in comparison with those of Bresnan et al.

**Extending the model**
Although Bresnan et al. (2007) have based their list of potentially relevant features on a large number of existing theories of and approaches to the dative alternation, we believe there are further linguistic characteristics that are potentially relevant.
Gries and Stefanowitsch (2004), for example, have tried to predict the dative alternation on the basis of the verb form solely. They extracted dative constructions from the ICE-GB Corpus and applied the Fisher exact test to the distribution of each verb form found in both constructions. The results seem promising: for the verb forms with a significant bias towards one of the two constructions (19 of 40), 82.2% of the alternation is correctly predicted, compared to 65.0% when simply selecting the most frequent construction. Therefore, we will include their 'collostructional analysis' in our research as well.
Example (2a) is taken from ICE-GB Corpus, and shows an NP-NP construction in an embedded clause.

Although the NP-PP variant we constructed in (2b) is equally grammatical, it is less easy to read and therefore seems less acceptable. This effect can be explained by the principle of end weight, which has also been mentioned in Bresnan et al. (2007). We believe the effect of the principle may increase when the dative construction is embedded deeper in the sentence.

(2) a. *I don't know if a million words would be enough to give [you]* RECIPIENT *[that statistical <,> uhm information to start off with]* THEME.                    (ICE-GB S1B-076_123:1:B)

    b. *I don't know if a million words would be enough to give [that statistical <,> uhm information to start off with]* THEME *[to you]* RECIPIENT.

Having seen instances such as (2), it seems useful to investigate a number of characteristics that relate to the syntactic environment in which the construction is found. Thus, we include information on the level (main or embedded) and type of clause (subordinate or relative), the mode (declarative, interrogative or imperative) and word order (unmarked, clefting or extraposition) of the clause in which the construction occurs, and also information on the polarity (positive or negative) of the clause.

Another feature that does not appear in the feature set of Bresnan et al. (2007) is the presence or absence of an adverb between the theme and the recipient, as exemplified in (3). We will include information on the form and the length of such intervening phrases.

(3)     *Ukraine lacks oil, but much Soviet oil comes from the Transcaucasian republics, now also aspiring to independence, which could try to bypass Moscow by selling [oil]*THEME **directly** *[to Ukrainian nationalists]*RECIPIENT.                    (ICE-GB W2C-008_20:1)

At the workshop, we will present our results and relate them to the findings of Bresnan et al. (2007) and Gries and Stefanowitsch (2004).

### References

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana, and Baayen, Harald 2007. Predicting the Dative Alternation. In *Cognitive Foundations of Interpretation*, Bouma, Gerlof, Kraemer, Ineke, and Zwarts, Joost (eds.), pp. 69-94. Amsterdam: Royal Netherlands Academy of Science.

Chater, Nick and Manning, Christopher D. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10 (7), pp. 335-344.

Greenbaum, Sidney (ed.) 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, Stefan Th. and Stefanowitsch, Anatol 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1), pp. 97-129.