Using the ICE-GB Corpus to model the English dative alternation

Daphne Theijssen
Department of Linguistics
Radboud University Nijmegen
d.theijssen@let.ru.nl

Abstract

In this paper, we applied Bresnan et al.'s (2007) Generalized Linear Model approach to model the English dative alternation to a corpus that shows more variation in text genre and discourse type: the ICE-GB corpus.

In a direct comparison, using only variables currently available for both data sets, we are able to explain 90.8% of the variability in the ICE-GB data as compared to 94.5% in the Switchboard corpus, possibly showing that the variation in genre in ICE-GB decreases the predictive power of the model. As expected, both models showed that the theme is especially pronominal and the recipient is often indefinite and non-local (third person) in the NP-PP construction, while in the NP-NP construction, the theme is longer than the recipient, the recipient is pronominal and the theme is indefinite.

Next, we extended the model for ICE-GB by including a number of syntactic variables. Only word order had a significant effect. This observation and investigation of errors has led us to wonder whether the dative alternation should be modelled in the traditional fashion or perhaps (also) according to the order of the recipient and the theme.

1. Introduction

There are many situations where speakers can choose between several syntactic 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.*). Bresnan et al. (2007) show that such choices can be modelled on the basis of a number of linguistic and paralinguistic properties of the construction and the clause it is embedded in. The larger part of Bresnan et al.'s

(2007) article concerns transcribed spoken data from the Switchboard Corpus. The mixed-effect model explains 95.0% of the dative alternation in 2360 spoken instances. They extended the data with 905 instances from the Wall Street Journal texts in the Penn Treebank and reached a prediction accuracy of 93.4% for the combined data. De Marneffe et al. (2007) have added 530 instances in child speech (CHILDES database) to the 611 instances with *do* and *give* in the SWB data, and were able to predict 95.7%. In both approaches, however, the data sets contain only two different genres.

In the present research we also aim at modelling the dative alternation, building on Bresnan et al.'s work. Our research goal is two-fold: (1) to investigate whether their models are also suitable for predicting the dative alternation in a corpus that shows more variation in text genre (the ICE-GB Corpus, Greenbaum 1996), and (2) to add syntactic variables that we expect to be relevant on the basis of the results found in (1). Since the research is still in progress, the results are preliminary.

The structure of this paper is as follows: The experimental setup for achieving both goals is introduced in Section 2. In Section 3, we present our method and results for comparing Bresnan et al.'s (2007) model for the Switchboard corpus to the our model found for the ICE-GB data. Section 4 introduces new variables, shows the results and gives a short error analysis. Some concluding remarks can be found in Section 5.

2. Experimental setup

2.1 Data

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) from the syntactically annotated ICE-GB Corpus. Following Bresnan et al. (2007), we ignored constructions with a preposition other than *to*, with a passivized object as subject, with a clausal object and with heavy NP shift (*give to the student the book*). We deleted constructions with coordinated verbs or verb phrases and with phrasal verbs from our data set as well. Next, we removed all cases with verbs that were present in instances with only one of the two dative constructions. Characteristics of the resulting data set ICE (919 instances) can be found in Table 1.

Table 1: Characteristics of ICE-GB and extracted datives (*ICE*)

	Spoken		*Written*		*Total*
	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	308	142	102	155	707
NP-PP	83	49	29	51	212
NP-NP / text	1.7	1.2	2.0	1.0	1.4
NP-PP / text	0.5	0.4	0.6	0.3	0.4

Bresnan et al. (2007) extracted 2360 occurrences of the dative alternation in the three-million-word Switchboard corpus (SWB): 1859 NP-NP and 501 NP-PP cases.

2.2 Method

Following Bresnan et al. (2007), we apply Generalized Linear Mixed Modelling (linear regression)[1] with *verb sense* as a random effect. The verb sense is the lemma of the verb together with its semantic class: 'abstract' (e.g. 'give.a' in *give it some thought*), 'transfer of possession' (e.g. 'give.t' in *give him the book*), 'future transfer of possession' (e.g. 'promise.f' in *promise her the money*), 'prevention of possession' (e.g. 'deny.p' in *deny them their rights*) and 'communication' (e.g. 'tell.c' in *tell him a story*). Verb sense is included as a random effect because it is expected to be biased with respect to certain properties of the two NP objects and to the construction realized (e.g. shown in Gries and Stefanowitsch's (2004) collostructional analysis).

We created a matrix with the values of a number of variables (see Sections 3.1 and 4.1) and the construction used (NP-NP or NP-PP) for each data case. The matrix can be seen as a multi-dimensional space (each variable being one dimension) in which all data points are represented. With the help of (multi-level) linear regression, a function is found that best splits the data into the NP-NP and the NP-PP cases. For each variable, a coefficient is established that 'bends' the function in the appropriate direction. We will examine the coefficients to analyze the importance and direction of the variables. Also, we use the function to evaluate how

[1] We used Linear Mixed Effects Regression with the help of the function 'lmer()' in the R package *lme4* (Bates 2005).

many data instances are on the correct 'side' of the function (the model fit). For a detailed description of logistic regression (and other) modelling, see Baayen (in press).

3. Varied written and spoken text

In this section we investigate whether, and if so how, an increase in the range of text and discourse types affects the quality of Bresnan et al.'s model (2007). As described in Section 2.1, we employ the syntactically annotated ICE-GB Corpus. The corpus 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).

3.1 Variables

In order to evaluate how well Bresnan et al.'s model (2007) generalizes to the ICE-GB data, the same variables need to be applied to both data sets.[2] For both theme (e.g. what is sent) and recipient (to whom it is sent), we established the pronominality, the definiteness, the animacy, the person, the number and the concreteness. Also, the length difference between the theme and the recipient (log scale) and the semantic verb class were added to the model. Since the concreteness of the recipient (not annotated) and the person and animacy of the theme (too sparse) were not present in SWB, we had to remove them from ICE as well. Similarly, we deleted discourse accessibility (previous mentioning or common ground) of theme and recipient, and structure parallelism (same variant used previously) from SWB because they have not been annotated in ICE yet. All feature values have been manually determined to reduce the risk of erroneous data. The annotation manual was based on Bresnan et al. (2007).

3.2 Results

The model fit can be found in Table 2. Surprising is the fact that the proportion of correctly predicted constructions in SWB (94.5%) almost equals that reported in Bresnan et al.'s article (2007): 95.0%. This means that the variables 'discourse

[2] I thank Dr. Joan Bresnan for providing me with the full data set, also including variables not present in the version available through the R package *languageR*

accessibility' and 'structure parallelism', which both have significant effects in their model, hardly have additional value for the simpler model applied here. The correctly predicted proportion for our data set (90.8%) is much lower, which could mean that the genre differences affect the predictability of the dative alternation.

Table 2: Classification table for SWB and ICE

		SWB: predicted			*ICE: predicted*		
		NP-NP	NP-PP	% correct	NP-NP	NP-PP	% correct
observed	NP-NP	1808	51	97.3%	673	34	95.2%
	NP-PP	79	422	84.2%	51	161	75.9%
			overall:	**94.5%**		*overall:*	**90.8%**

% correct from always guessing NP-NP: 80.0% (SWB) and 78.8% (ICE)

The coefficients of significant effect in the model for Bresnan et al.'s and our data can be found in Table 3 and 4 respectively. The directions are as expected: in the NP-PP construction, the theme is especially pronominal and the recipient is often indefinite and non-local (third person), while in the NP-NP construction, the theme is longer than the recipient, the recipient is pronominal and the theme is indefinite. In SWB, we find in addition that the recipient is often inanimate in the NP-PP variant and the theme singular in the NP-NP variant.

Table 3: Significant effects in SWB

variable + value	*direction*	*coefficient*	*z-value*	*significance*	*level*
Pronominality of theme	NP-PP	2,34	7,89	2,96E-15	***
Inanimacy of recipient	NP-PP	1,67	3,55	3,79E-04	***
Indefiniteness of recipient	NP-PP	1,32	4,13	3,63E-05	***
Non-local person of recipient	NP-PP	0,54	1,99	4,66E-02	*
Singular number of theme	NP-NP	-0,81	-3,19	1,42E-03	**
Length difference (log)	NP-NP	-1,58	-9,12	2,00E-16	***
Pronominality of recipient	NP-NP	-2,28	-7,78	7,18E-15	***
Indefiniteness of theme	NP-NP	-2,37	-9,43	2,00E-16	***

Table 4: Significant effects in ICE

variable + value	*Direction*	*coefficient*	*z-value*	*significance*	*level*
Non-local person of recipient	NP-PP	1,52	4,02	5,86E-05	***
Indefiniteness of recipient	NP-PP	1,33	3,36	7,83E-04	***
Indefiniteness of theme	NP-NP	-0,99	-3,19	1,43E-03	**
Pronominality of recipient	NP-NP	-1,05	-3,03	2,43E-03	**
Length difference (log)	NP-NP	-1,73	-8,51	2,00E-16	***

4. 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 some syntactic characteristics that are potentially relevant.

4.1 Additional variables

Linguists (including Bresnan et al. 2007) generally agree on the existence of the *principle of end weight*: the tendency to place long constituents at the end of the clause. We believe the effect of the principle may increase when the dative construction is embedded deeper in the sentence. An example in the ICE-GB Corpus can be found in (3a). Although the NP-PP variant we constructed in (3b) is equally grammatical, it is less easy to read and therefore less acceptable. We thus add a variable denoting the type of the clause (main, subordinate or relative) to the model.

(3) a. *I don't know if a million words would be enough to give [you]* $_{\text{RECIPIENT}}$ *[that statistical <,> uhm information to start off with]* $_{\text{THEME}}$.

(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]* $_{\text{THEME}}$ *[to you]* $_{\text{RECIPIENT}}$.

Similarly, we add a number of other characteristics that describe the clause in which the construction is embedded: its mode (declarative, interrogative or imperative) and word order (unmarked or fronting). For cases found in dependent clauses, we include their type (clausal or phrasal) and for cases found in clausal dependent clauses, also their importance (adjunct or complement).

Another feature that is possibly relevant is the presence or absence of an adverb between the theme and the recipient, as exemplified in (4). We therefore add two variables denoting the length of such intervening phrases in words and in characters.

(4) *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)

4.2 Results

The predicted proportions of the extended model for ICE are shown in Table 5. The model accuracy (91.9%) is higher than what we found without including the syntactic variables (90.9%), though not significantly (chi-square = 0.833, p = 0.361). Most improvement is achieved because more NP-PP cases are now also predicted to be NP-PP.

Table 5: Classification table for ICE (with syntax)

		ICE: predicted with syntax added		
		NP-NP	NP-PP	% correct
observed	NP-NP	674	33	95.3%
	NP-PP	41	171	80.7%
			overall:	**91.9%**

% correct from always guessing NP-NP: 78.8%

In Table 6, the coefficients of the significant effects in model B can be found. The effects of the first model (Table 4) are also significant in this model, and show the same direction. Of the newly added syntactic variables, only word order (unmarked) has a significant effect. The importance of word order shows that we might be modelling the wrong variants. We have assumed that in NP-NP constructions, the recipient precedes the theme, and in NP-PP constructions, the other way around. However, in cases of fronting, the order is altered and therefore difficult to model.[3] Since we plan to include passivized objects in the future, and perhaps even

[3] It is not clear whether Bresnan et al. (2007) included or excluded these instances.

instances with heavy NP shift (*She gave to the student the book*), word order is a crucial aspect to consider.

Table 6: Significant effects in ICE (with syntax)

variable + value	*direction*	*coefficient*	*z-value*	*significance*	*level*
Unmarked word order	NP-PP	2,22	3,23	1,25E-03	**
Non-local person of recipient	NP-PP	1,60	3,91	9,40E-05	***
Indefiniteness of recipient	NP-PP	1,56	3,72	2,02E-04	***
Pronominality of theme	NP-PP	1,18	2,57	1,02E-02	*
Pronominality of recipient	NP-NP	-1,11	-3,03	2,44E-03	**
Indefiniteness of theme	NP-NP	-1,18	-3,62	3,00E-04	***
Length difference (log)	NP-NP	-1,90	-8,71	2,00E-16	***

4.3 Error analysis

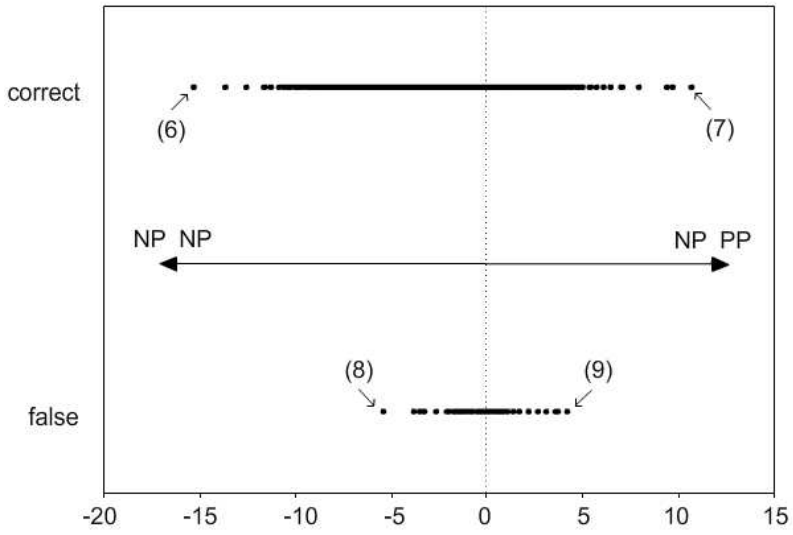
The distance of a data point from the regression function can be used to measure the certainty of the predicted construction (here represented as the log odds). The more this distance differs from zero, the more certain the prediction is. Negative values direct at the NP-NP variant, positive ones at the NP-PP. In Figure 1, the found values have been plotted for both the correctly and the falsely predicted data cases (approach taken from Gries 2003). The lower graph (representing false predictions) is smaller than the upper graph, indicating that for incorrectly predicted cases, the model is less certain than for those correctly predicted (which is good). Also, we can consider the extremes (6) and (7) to be prototypical examples of the variants NP-NP and NP-PP respectively, whereas extremes (8) and (9) are unexpected constructions according to the model (Gries 2003):

(6) *You have <u>given</u> {me}$_{REC}$ [you]$_{TH}$ and you have restored to me myself.*
(ICE-GB W1B-006_16:1)

(7) *And secondly I obviously can't <u>do</u> [justice]$_{TH}$ in sus in such a short time <,> {to the exposition of the ways in which this theory differed from other views at the time}$_{REC}$ <,,>* (ICE-GB S2B-049_5:1:A)

(8) *But why on earth should <,> why on earth should Mr Neil make that comment unless Mr <,> uh Slipper had <u>given</u> [the appearance]$_{TH-1}$ {to him}$_{REC}$ [uh of uh ignorance of the extradition treaty]$_{TH-2}$* (ICE-GB S2A-064_82:2:A)

Figure 1: Coefficients and prediction accuracy for ICE with syntax



(9) *So I think uh Perez de Cuellar has probably been prevailed on to uh to to come out with some kind of platitude that will uh give {all these reporters who were sitting around here all day waiting for something to happen}$_{REC}$ [something to report]$_{TH}$* (ICE-GB S2B-010_86:1:B)

Sentences (8) and (9) are both spoken data showing hesitation. This is especially clear in (8) where the theme has been interrupted by the recipient, splitting the theme in two. Again, we see that a distorted order of the objects influences the predictive power of the model. In (9), the verb *give* is preceded by a hesitation (*uh*), which could mean that the speaker is determining what he wants to say and actually saying it simultaneously. This would explain the fact that the principle of end weight is heavily violated.

5. Concluding remarks

In this paper, we have attempted to establish whether Bresnan et al.'s (2007) linear mixed-effect regression model for the dative alternation in SWB is also suitable for modelling the alternation in a corpus with more genre variation (ICE). The proportion of correctly predicted constructions for ICE was lower (90.8%) than that for SWB (94.5%), which could indicate that the text type affects the performance (or fit) of the model. In the future, we should thus include the text type as an additional variable (provided that the data is not too sparse).

However, there may be other causes for the lower prediction accuracies. We have annotated the data following Bresnan et al. as well as we could, but there may still be annotation differences. Also, the ICE-GB corpus consists of language used by British English speakers, while Switchboard is American English. Lastly, both SWB and ICE may have been hampered by the fact that certain variables had to be ignored to arrive at a model that contained variables annotated and sufficiently frequent in both data sets. In near future, when we have completed our variable set, we need to establish the benefit of the syntactic variables again. It would also be interesting to discover how well the models for SWB and ICE generalize to unseen data, for example by applying the SWB model (including its coefficients) to ICE and vice versa.

Furthermore, we have seen that word order has a significant effect in ICE, and that split objects are difficult to model. These are very useful findings for the continuation of our research. We will have to ask ourselves whether we want to model according to the traditional variants (NP-NP and NP-PP), with the possible consequence of removing all instances with marked word order, or perhaps (also) attempt to model the alternation in the ordering of theme and recipient.

References

Baayen, R. H. (in press). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Bates, D. 2005. Fitting linear mixed models in R. *R News*, 5 (1): 27-30.

Bresnan, J., A. Cueni, T. Nikitina and R.H. Baayen 2007. Predicting the Dative Alternation. In Bouma, G, I. Kraemer and J. Zwarts (eds.), *Cognitive Foundations of Interpretation:* 69-94. Amsterdam: Royal Netherlands Academy of Science.

De Marneffe, M-C, S. Grimm, U.C. Priva, S. Lestrade, G. Ozbek, T. Schnoebelen, S. Kirby, M. Becker, V. Fong and J. Bresnan 2007. A Statistical Model of Grammatical Choices in Childrens' Productions of Dative Sentences. Presented at FAVS 2007, York, UK.

Godfrey, J., E. Holliman and J. McDaniel 1992. Switchboard: Telephone speech corpus for research and development. *Proceedings of ICASSP-92*, San Francisco: 517-20.

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

Gries, S. Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1: 1-27.

Gries, S. Th. and A. Stefanowitsch 2004. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. *International Journal of Corpus Linguistics* 9: 97-129.