Variable Selection in Logistic Regression: The British English Dative Alternation

Daphne Theijssen

Centre for Language Studies, Radboud University Nijmegen,
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands
d.theijssen@let.ru.nl
http://lands.let.ru.nl/~daphne

Abstract. This paper addresses the problem of selecting the 'optimal' variable subset in a logistic regression model for a medium-sized data set. As a case study, we take the British English dative alternation, where speakers and writers can choose between two – equally grammatical – syntactic constructions to express the same meaning. With 29 explanatory variables taken from the literature, we build two types of models: one with the verb sense included as a random effect, and one without a random effect. For each type, we build three different models by including all variables and keeping the significant ones, by successively adding the most predictive variable (forward selection), and by successively removing the least predictive variable (backward elimination). Seeing that the six approaches lead to six different variable selections (and thus six different models), we conclude that the selection of the 'best' model requires a substantial amount of linguistic expertise.

1 Introduction

There are many linguistic phenomena that researchers have tried to explain on the basis of features on several different levels of description (semantic, syntactic, lexical, etc.), and it can be argued that no single level can account for all observations. Probabilistic modelling techniques can help in combining these partially explanatory features and testing the combination on corpus data. A popular – and rather successful – technique for this purpose is logistic regression modelling. However, *how* exactly the technique is best employed for this type of research remains an open question.

Statistical models built using corpus data do precisely what they are designed to do: find the 'best possible' model for a specific data set given a specific set of explanatory features. The issue that probabilistic techniques model data (while one would actually want to model underlying processes) is only aggravated by the fact that the variables are usually not mutually independent. As a consequence, one set of data and explanatory features can result in different models, depending on the details of the model building process.

Building a regression model consists of three main steps: (1) deciding which of the available explanatory features should actually be included as variables in

T. Icard, R. Muskens (Eds.): ESSLLI 2008/2009 Student Sessions, LNAI 6211, pp. 87–101, 2010.

the model, (2) establishing the coefficients (weights) for the variables, and (3) evaluating the model. The first step is generally referred to as *variable selection* and is the topic of the current paper. Steps (1) and (3) are clearly intimately related.

Researchers have employed at least three different approaches to variable selection: (1) first building a model on all available explanatory features and then keeping/reporting those that have a significant contribution (e.g. [3]), (2) successively adding the most explanatory feature (forward), until no significant gain in model accuracy[1] is obtained anymore (e.g. [9]), and (3) starting with a model containing all available features, and (backward) successively removing those that yield the smallest contribution, as long as the accuracy of the model is not significantly reduced (e.g. [2]). In general, researchers report on only one (optimal) model without giving clear motivations for their choice of the procedure used.

In this paper, we compare the three approaches in a case study: we apply them to a set of 930 instances of the British English dative alternation, taken from the British component of the ICE Corpus. In the dative alternation, speakers choose between the double object (1) and the prepositional dative construction (2).

1. She handed the student the book.
2. She handed the book to the student.

The explanatory features (explanations suggested in the literature) are taken from Bresnan et al.'s work on the dative alternation in American English [3].

Previous research (e.g. [8,3]) has indicated that the verb or verb sense often predicts a preference for one of the two constructions. However, contrary to the fourteen explanatory features suggested by Bresnan et al., which can be treated as fixed variables because of their small number of values (often only two), *verb sense* has so many different values that it cannot be treated as a fixed variable in a regression model. Recently developed logistic regression models can handle variables with too many values by treating these as random effects (cf. [18]). In order to examine the effect of building such *mixed models*, we create models with and without a random effect in each of the three approaches to variable selection described above. This leads to a total of six different models.

Our goal is to investigate whether it is justified to report only one 'optimal' regression model, if models can be built in several different ways. We will also pay attention to the role of a random effect in a model of syntactic variation built with a medium-sized set of observations. The case of the British English dative alternation is used to illustrate the issues and results.

The structure of this paper is as follows: A short overview of the related work can be found in Section 2. The data is described in Section 3. In Section 4, we explain the method applied. The results are shown and discussed in Section 5. In the final Section (6), we present our conclusions.

[1] Obviously, the accuracy measure will also have considerable impact on the result.

2 Related Work

2.1 The Dative Alternation

Bresnan et al. [3] built various logistic regression models for the dative alternation based on 2360 instances they extracted from the three-million word Switchboard Corpus of transcribed American English telephone dialogues [5]. With the help of a mixed-effect logistic regression model, or *mixed model*, with verb sense as a random effect, they were able to explain 95% of the variation. They defined the verb sense as the verb lemma together with its semantic verb class. The semantic verb class is either 'abstract' (e.g. *give it some thought*), 'communication' (e.g. *tell him a story*), 'transfer of possession' (e.g. *give him the book*), 'prevention of possession' (e.g. *deny him the money*) or 'future transfer of possession' (e.g. *promise him help*). To test how well the model generalizes to previously unseen data, they built a model on 2000 instances randomly selected from the total set, and tested on the remaining 360 cases. Repeating this 100 times, 94% of the test cases on average were predicted correctly.

Many of the variables in the model concern the two objects in the construction (*the student* and *the book* in example 1 and 2). In the prepositional dative construction, the object first mentioned is the theme (*the book*), and the second object the recipient (*the student*). In the double object construction, the recipient precedes the theme. Bresnan et al. found that the first object is typically (headed by) a pronoun, mentioned previously in the discourse (*given*), animate, definite and longer (in number of words) than the second object. The characteristics of the second object are generally the opposite: non-pronominal, new, inanimate, indefinite and shorter.

According to Haspelmath [10], there is a slight difference between the dative alternation as it occurs in British English and in American English. When the theme is a pronoun, speakers of American English tend to allow only the prepositional dative construction. In British English, clauses such as *She gave me it* and even *She gave it me* are also acceptable. Haspelmath provides no evidence for these claims (neither from corpora nor from psycholinguistic experiments). He refers to Siewierska and Hollmann [17], who present frequency counts in various corpora of Lancashire (British) English: Of the 415 instances of the dative alternation they found, 8 were of the pattern *She gave me it*, and 15 of *She gave it me*. It must be expected that such differences between language variants result in different behaviour of variables in models for these different language variants. Inappropriate approaches to variable selection may obscure this kind of 'real' difference.

Gries [7] performed analyses with multiple variables that are similar to those in Bresnan et al. [3], but applied a different technique (linear discriminant analysis or LDA) on a notably smaller data set consisting of only 117 instances from the British National Corpus [4]. The LDA model is trained on all instances, and is able to predict 88.9% of these cases correctly (with a majority baseline of 51.3%). There is no information on how the model performs on previously unseen data.

Gries and Stefanowitsch [8] investigated the effect of the verb in 1772 instances from the ICE-GB Corpus [6]. When predicting the preferred dative construction for each verb (not taking into account the separate senses), 82.2% of the constructions could be predicted correctly. Using verb bias as a predictor thus outperforms the majority baseline of 65.0%.

2.2 Variable Selection in Logistic Regression

Variable selection in building logistic regression models is an extremely important issue, for which no hard and fast solution is available. In [11, chapter 5] it is explained that variable selection is often needed to arrive at a model that reaches an acceptable prediction accuracy and is still interpretable in terms of some theory about the role of the independent variables. Keeping too many variables may lead to overfitting, while a simpler model may suffer from underfitting. The risk of applying variable selection is that one optimizes the model for a particular data set. Using a slightly different data set may result in a very different variable subset.

Previous studies aimed at creating logistic regression models to explain linguistic phenomena have used various approaches to variable selection. Grondelaers and Speelman [9], for instance, successively added the most predictive variables to an empty model, while Blackwell [2] successively eliminated the least predictive variables from the full model. The main criticisms of these methods are (1) that the results are difficult to interpret when the variables are highly correlated, (2) that deciding which variable to remove or add is not trivial, (3) that all methods may result in different models that may be sub-optimal in some sense, and (4) that each provides a single model, while there may be more than one 'optimal' subset [11].

A third approach to variable selection used in linguistic research is keeping only the significant variables in a complete model (cf. Bresnan et al. [3]). This is also what Sheather suggests in [16, chapter 8]. Before building a model, however, he studies plots of the variables to select those that he expects to contribute to the model. Where beneficial, he transforms the variables to give them more predictive power (e.g. by taking their log). After these preprocessing steps he builds a model containing all the selected variables, removes the insignificant ones, and then builds a new model. As indicated by Izenman [11], variable selection on the basis of a data set may lead to a model that is specific for that particular set. Since we want to be able to compare our models to those found by Bresnan et al. [3], who did not employ such transformations, we refrain from such preprocessing and we set out using the same set of variables they used in the variable selection process.

Yet another approach mentioned in [11] is to build all models with each possible subset and select those with the best trade-off between accuracy, generalisability and interpretability. An important objection to this approach is that it is computationally expensive to carry out, and that decisions about interpretability may suffer from theoretical prejudice. For these reasons, we do not employ this method.

3 Data

Despite the fact that a number of researchers have studied the dative alternation in English (see Section 2.1), none of the larger data sets used is available in such a form that it enables the research in this paper.[2] We therefore established our own set of instances of the dative alternation in British English. Since we study a syntactic phenomenon, it is convenient to employ a corpus with detailed (manually checked) syntactic annotations. We selected the one-million-word British component of the ICE Corpus, the ICE-GB, containing both written and (transcribed) spoken language [6].

We used a Perl script to automatically extract potentially relevant clauses from the ICE-GB. These were clauses with an indirect and a direct object (double object) and clauses with a direct object and a prepositional phrase with the preposition *to* (prepositional dative). Next, we manually checked the extracted sets of clauses and removed irrelevant clauses such as those where the preposition *to* had a locative function (as, for example, in *Fold the short edges to the centre.*).

Following Bresnan et al. [3], we ignored constructions with a preposition other than *to*, with a clausal object, with passive voice and with reversed constructions (e.g. *She gave it me*). To further limit the influence of the syntactic environment of the construction, we decided to exclude variants in imperative and interrogative clauses, as well as those with phrasal verbs (e.g. *to hand over*). Coordinated verbs or verb phrases were also removed. The characteristics of the resulting data sets can be found in Table 1.

Table 1. Characteristics of the 930 instances taken from the ICE-GB Corpus

Medium	Double object	Prep. dative	Total
Spoken British English	406	152	558
Written British English	266	106	372
Total	672	258	930

4 Method

4.1 Explanatory Features

We adopt the explanatory features and their definitions from Bresnan et al. [3] (Table 2), and manually annotate our data set following an annotation manual based on these definitions.[3]

Our set includes one feature that was not used in [3]: *medium*, which tells us whether the construction was found in written or spoken text. It may well be

[2] Although most of the data set used in [3] is available through the `R` package `LanguageR`, the original sentences and some annotations are not publicly available because they are taken from an unpublished, corrected version of the Switchboard Corpus.

[3] The annotation manual is available online: `http://lands.let.ru.nl/˜daphne/ downloads.html`

Table 2. Explanatory features (th=theme, rec=recipient). All nominal explanatory features are transformed into binary variables with values 0 and 1.

Feature	Values	Description
1. rec = animate	1, 0	human or animal, or not
2. th = concrete	1, 0	with fixed form and/or space, or not
3. rec = definite	1, 0	definite pronoun, proper name or noun preceded by definite determiner, or not
4. th = definite	1, 0	Id.
5. rec = given	1, 0	mentioned/evoked ≤20 clauses before, or not
6. th = given	1, 0	Id.
7. length difference	-3.4-4.2	ln(#words in th) − ln(#words in rec)
8. rec = plural	1, 0	plural in number, or not (singular)
9. th = plural	1, 0	Id.
10. rec = local	1, 0	first or second person (*I*, *you*), or not
11. rec = pronominal	1, 0	headed by a pronoun, or not
12. th = pronominal	1, 0	Id.
13. verb = abstract	1, 0	*give it some thought* is abstract,
verb = communication	1, 0	*tell him a story* is communication,
verb = transfer	1, 0	*give him the book* is transfer
14. structural parallellism	1, 0	preceding instance is prep. dative, or not
15. medium = written	1, 0	type of data is written, or not (spoken)

that certain variables only play a role in one of the two media. In order to test this, we include the 14 (two-way) interactions between the features taken from Bresnan et al. and the medium.[4] Together with the feature *medium* itself, this yields a total number of 29 features.

As mentioned in the Introduction, we will build models with and without including *verb sense* as a random effect. Following [3], we define the verb sense as the lemma of the verb together with its semantic class, e.g. *pay_a* for *pay* with an abstract meaning (*pay attention*) and *pay_t* when *pay* is used to describe a transfer of possession (*pay $10*). In total, our data set contains 94 different verb senses (derived from 65 different verbs). The distribution of the verb senses with 5 or more occurrences can be found in Table 3.

As predicted by Gries and Stefanowitsch [8], many verbs show a bias towards one of the two constructions. The verb *give*, for instance, shows a bias for the double object construction, and *sell* for the prepositional dative construction. Only for *pay* and *send*, the bias differs for the different senses. For example, *pay* shows a clear bias towards the prepositional dative construction when it has an abstract meaning, but no bias when transfer of possession is meant. Nevertheless, we follow the approach in [3] by taking the verb sense, not the verb, as the random effect.

[4] We are aware of the fact that there are other ways to incorporate the medium in the regression models, for instance by building separate models for the written and the spoken data. Since the focus of this paper is on the three approaches in combination with the presence or absence of a random effect, we will limit ourselves to the method described.

Table 3. Distribution of verb senses with 5 or more occurrences in the data set. The verb senses in the right-most list have a clear bias towards the double object (d.obj.) construction, those in the left-most for the prepositional dative (p.dat.) construction, and those in the middle show no clear preference. The *a* represents *abstract*, *c communication* and *t transfer of possession*.

# d.obj. > # p.dat.			# d.obj. ≈ # p.dat.			# d.obj. < # p.dat.		
verb sense	d.obj.	p.dat.	verb sense	d.obj.	p.dat.	verb sense	d.obj.	p.dat.
give_a	255	32	*do_a*	8	10	*pay_a*	2	12
give_t	56	21	*send_c*	9	7	*cause_a*	5	8
give_c	66	10	*lend_t*	8	7	*sell_t*	0	10
tell_c	67	1	*pay_t*	6	5	*owe_a*	2	6
send_t	42	16	*leave_a*	5	4	*explain_c*	0	6
show_c	37	9	*write_c*	4	5	*present_c*	0	6
offer_a	24	9	*bring_t*	3	2	*read_c*	1	4
show_a	6	1	*hand_t*	3	2			
offer_t	6	0						
tell_a	6	0						
wish_c	6	0						
bring_a	4	1						

4.2 Variable Selection

Using the values of the 29 explanatory features (fixed effect factors), we establish a regression function that predicts the natural logarithm (ln) of the odds that the construction C in clause j is a prepositional dative. The prepositional dative is regarded a 'success' (with value 1), while the double object construction is considered a 'failure' (0). The regression function for the models without a random effect is: (1):

$$\ln odds(C_j = 1) = \alpha + \sum_{k=1}^{29}(\beta_k V_{jk}) \ . \tag{1}$$

The α is the intercept of the function. $\beta_k V_{jk}$ are the weights β and values V_j of the 29 variables k. For the model with the random effect (for verb sense i), the regression function is:

$$\ln odds(C_{ij} = 1) = \alpha + \sum_{k=1}^{29}(\beta_k V_{jk}) + e_{ij} + r_i \ . \tag{2}$$

The random effect r_i is normally distributed with mean zero ($r_i \sim N(0, \sigma_r^2)$), independent of the normally distributed error term e_{ij} ($e_{ij} \sim N(0, \sigma_e^2)$). The optimal values for the function parameters α, β_k and (for models with a random effect) r_i and e_{ij} are found with the help of Maximum Likelihood Estimation.[5]

The outcome of the regression enables us to use the model as a classifier: all cases with $\ln odds(C_j = 1) \geq t$ (for the models without a random effect) or

[5] We use the functions `glm()` and `lmer()` [1] in `R` [15].

$\ln odds(C_{ij} = 1) \geq t$ (for models with a random effect) are classified as prepositional dative, all with $\ln odds(C_j = 1) < t$ or $\ln odds(C_{ij} = 1) < t$ as double object, with t the decision threshold, which we set to 0. With this threshold, all instances for which the regression function outputs a negative ln odds are classified as double object constructions, all other instances as prepositional dative.

In the first approach, we include all 29 features in the model formula. We then remove all variables V_k that do not have a significant effect in the model output,[6] and build a model with the remaining (significant) variables.

For the second approach, being forward selection, we start with an empty model and successively add the variable that is most predictive. As Izenman [11] explains, there are several possible criteria for deciding which variable to enter. We decide to enter the variable that yields the highest area under the ROC (Receiver Operating Characteristics) curve of the extended model. The ROC curve shows the proportions of correctly and incorrectly classifies instances as a function of the decision threshold. The area under the ROC curve (AUC) gives the probability that the regression function, when randomly selecting a positive (prepositional dative) and a negative (double object) instance, outputs a higher log odds for the positive instance than for the negative instance. The AUC is thus an evaluation measure for the quality of a model. It is calculated with:

$$\frac{average\_rank(x_{C=1}) - \frac{p+1}{2}}{n-p}, \tag{3}$$

where $average\_rank(x_{C=1})$ is the average rank of the instances x that are prepositional dative (when all instances are ranked numerically according to the log odds), p the number of prepositional dative instances, and n the total number of instances.[7] We add the next most predictive variable to the model as long as it gives an improvement over the AUC of the model without the variable. An interaction of variable V_k with *medium* is only included when the resulting AUC is higher than the value reached after adding the main variable V_k.[8] Two AUC values are considered different when the difference is higher than a threshold. We set the threshold to 0.002.[9]

For the third approach (backward elimination), we use the opposite procedure: we start with the full model, containing all 29 variables, and successively leave out the variable V_k that, after removal, yields the model with the highest AUC value that is not lower than the AUC value for the model with V_k. When the AUC value of a model without variable V_k does not differ from the AUC value of the model without the interaction of V_k with *medium*, we remove the interaction. Again, AUC values are only considered different when the difference exceeds a threshold (again set to 0.002).

[6] We use the P-values as provided by `glm()` and `lmer()`.

[7] We use the function `somers2()` created in `R` [15] by Frank Harrell.

[8] When including an interaction but not the main variables in it, the interaction will also partly explain variation that is caused by the main variables [14].

[9] The threshold value has been established experimentally.

We evaluate the models with and without random effects by establishing the model quality (training and testing on all 930 cases) by calculating the percentage of correctly classified instances (accuracy) and the area under the ROC curve (AUC). Also, we determine the prediction accuracy reached in 10-fold cross-validation (10 sessions of training on 90% of the data and testing on the remaining 10%) in order to establish how well a model generalizes to previously unseen data. In the 10-fold cross-validation setting, we provide the algorithms with the variables selected in the models trained on all 930 cases. The regression coefficients for these subsets of variables are then estimated for each separate training set.

The coefficients in the regression models help us understand which variables play what role in the dative alternation. We will therefore compare the coefficients of the significant effects in the models built on all 930 instances.

5 Results

5.1 Mixed Models

Table 4 gives the model quality and prediction accuracy for the different regression models we built, including verb sense as a random effect. The prediction accuracy (the percentage of correctly classified cases) is significantly higher than the majority baseline (always selecting the double object construction) in all settings, also when testing on new data ($p < 0.001$ for the three models, Wilcoxon paired signed rank test).

Table 4. Number of variables selected, model quality and prediction accuracy of the regression models with *verb sense* as a random effect

			model quality (train=test)		*10-fold cv*
selection	#variables	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.979	0.935	0.819
2. forward	4	0.723	0.979	0.932	0.827
3. backward	4	0.723	0.978	0.928	0.833

When training and testing on all 930 instances, the mixed models reach very high AUC and prediction accuracy (model quality). However, seeing the decrease in accuracy in a 10-fold cross-validation setting, it seems that the mixed models do not generalize very well to previously unseen data.

The significant effects for the variables selected in the three approaches are presented in Table 5. The directions of the main effects are the same as the results presented in Section 2.1 for American English [3].

The forward selection (2) and backward elimination (3) approaches lead to almost the same regression model. The only difference is that in the backward model, the discourse givenness of the recipient is included as a main effect, while it is included as an interaction with medium in the forward model. Both indicate that the choice for the double object construction is more likely when the

Table 5. Coefficients of significant effects in (mixed) regression models with verb sense as random effect, trained on all 930 instances, *** p<0.001 ** p<0.01 * p<0.05. The (negative) effects above the horizontal line draw towards the double object construction, and the (positive) effects below it toward the prepositional dative construction.

Effect	1. significant	2. forward	3. backward
length difference	-2.50 ***	-2.44 ***	-2.39 ***
rec=animate	-1.01 *		
rec=given			-1.44 ***
rec=given, medium=spoken		-0.94 *	
rec=given, medium=written		-1.74 ***	
rec=local	-2.53 ***	-1.82 ***	-1.78 ***
th=pronominal, medium=written	-1.79 *		
(intercept)	2.05 ***	2.32 ***	2.38 ***
th=definite	1.78 ***		
th=given		2.34 ***	2.33 ***
th=pronominal	2.19 ***		

recipient has been mentioned previously in the discourse (and is thus *given*). In the forward model, this effect is a little stronger in writing than in speech.

The animacy of the recipient is only found significant in the model obtained by keeping the significant variables (1). The other differences between the two stepwise models and this model are likely to be caused by the fact that the information contained in the variables shows considerable overlap. Pronominal and definite objects are also often discourse given. A significant effect for the one variable may therefore decrease the possibility of regarding the other as significant. This is exactly what we see: the model obtained through the two stepwise approaches contains a variable denoting the givenness of the theme but none describing its pronominality or definiteness, while it is the other way around for the model with the significant variables from the full model.

The model obtained by keeping the significant variables in the full model also contains one interaction, namely that between medium and a pronominal theme. The main effect (without medium) is also included, but it shows the opposite effect. When the theme is pronominal, speakers tend to use the prepositional dative construction (coefficient 2.15). This effect seems much less strong in writing (remaining coefficient 2.15 - 2.01 = 0.14).

What remains unclear, is which of the three models is more suitable for explaining the British English dative alternation. Seeing the differences between the significant effects in the three models we found, and the relatively low prediction accuracy in 10-fold cross-validation, it seems that the models are modelling the specific data set rather than the phenomenon. A probable cause is that the mixed models are too complex to model a data set consisting of 930 instances. In the next section, we apply the three approaches to build simpler models, namely without the random effect.

5.2 Models without a Random Effect

The model quality and prediction accuracy for the models without a random effect can be found in Table 6.

Table 6. Model fit and prediction accuracy of the regression models without a random effect

			model quality (train=test)		*10-fold cv*
selection	#variables	baseline	AUC	accuracy	aver. accuracy
1. significant	6	0.723	0.938	0.878	0.872
2. forward	7	0.723	0.943	0.878	0.876
3. backward	8	0.723	0.946	0.882	0.876

The estimates of model quality AUC and accuracy are considerably lower than the values obtained with the mixed models (Table 4). On the other hand, the models without a random effect generalize well to new data: the prediction accuracy in 10-fold cross-validation is very similar to the model quality accuracy (when training and testing on all instances). The prediction accuracies reached in 10-fold cross-validation are significantly better than those reached with the best mixed model ($p < 0.001$ for the three regular models compared to the backward mixed model, following the Wilcoxon paired signed rank test). Apparently the simpler models, without a random effect, outperform the mixed models when applying them to previously unseen data.

Table 7 shows the significant effects in the models without random effect. Again, the directions of the coefficients are the same across the three models, but they disagree on the significance of the variables. Three variables are selected in all three approaches: the person of the recipient (local or non-local), the pronominality of the recipient, and the concreteness of the theme. The latter two were not selected at all in the mixed-effect approach of the previous section. Three more variables have significant effects in two of the three models. According to all three models, speakers tend to use the double object construction when the theme is longer than the recipient. The backward elimination model (3), however, shows that the effect of length difference is especially strong in speech. As for the mixed model in the previous section, the forward selection has selected the interaction between the medium and the discourse givenness of the recipient. Writers are thus more likely to choose the double object construction when the recipient has recently been mentioned in the text, than when the recipient is newly (re)introduced.

The semantic verb class is only selected in the backward elimination. In the literature (cf. [13]), it is argued that the prepositional dative construction is especially used to express a change of place (moving the theme), and the double object construction a change of state (possessing the theme). In this perspective, we would expect instances with a *transfer of possession* to be in the prepositional dative construction (*give a book to you*), and instances with *abstract* meanings in the double object construction (*give you moral support*). This is also what

Table 7. Coefficients of significant effects in regression models (without random effect), trained on all 930 instances, *** p<0.001 ** p<0.01 * p<0.05. The (negative) effects above the horizontal line draw towards the double object construction, and the (positive) effects below it toward the prepositional dative construction.

Effect	1. significant	2. forward	3. backward
length difference	-1.73 ***		-2.00 ***
length difference, medium=spoken		-2.35 ***	
length difference, medium=written		-1.71 ***	
rec=definite		-1.01 **	-1.15 ***
rec=given, medium=written		-0.66 *	
rec=local	-1.22 ***	-0.94 **	-1.15 **
rec=pronominal	-1.35 ***	-0.88 **	-1.25 ***
verb=abstract, medium=written			-0.99 *
verb=transfer, medium=spoken			-1.04 *
verb=transfer, medium=written			-1.32 *
(intercept)		0.82 **	1.56 **
th=concrete	1.33 ***	1.48 ***	1.63 ***
th=definite		1.58 ***	1.16 ***
th=given	1.48 ***		0.98 **

Bresnan et al. [3] found for spoken American English. In the backward model, however, the effect is the opposite: a transfer of possession is more strongly drawn towards the double object construction than an abstract meaning. The problem here is that these two semantic verb classes depend largely on the concreteness of the theme (Pearson correlation = 0.739), a feature that has been selected in all three models in Table 7. When the semantic verb class is *transfer of possession*, the theme is very likely to be *concrete*. The backward model thus seems to compensate the positive coefficient of concreteness (1.63) by given a negative coefficient to the semantic verb class (e.g. -1.32 for *transfer of possession* in writing). The resulting effect is still directed at the double object construction (remaining coefficient 1.63 - 1.32 = 0.31), but it is not very strong. In Section 3, we saw that only *pay* and *send* showed different biases towards one of the two constructions in different verb senses. It seems that the biases are mostly due to the verb (see also [8]) and the concreteness of the theme, and not so much to their semantic verb classes *abstract*, *communication* and *transfer of possession.*

6 Discussion and Conclusion

In this paper, we built regular and mixed (i.e. containing a random effect) logistic regression models in order to explain the British English dative alternation. We used a data set of 930 instances taken from the ICE-GB Corpus, and took the explanatory factors suggested by Bresnan et al. [3]. The regular and the mixed models were constructed following three different approaches: (1) providing the algorithms with all 29 variables and keeping the significant ones, (2) starting with an empty model and forwardly successively adding the most predictive variables, and (3) starting with a model with all 29 features and backwardly successively removing the least predictive variables. In total, we thus have built six logistic regression models for the same data set.

The six models show some overlap in the variables that are regarded significant. These variables show the same effects as found for American English [3]: pronominal, relatively short, local (first or second person), discourse given, definite and concrete objects typically precede objects with the opposite characteristics. Contrary to the claims in Haspelmath [10], we found no evidence for the hypothesis that the dative alternation in British English differs from that in American English. With respect to *medium*, there seem to be some differences between the dative alternation in speech and writing. Four variables were selected as interactions with medium. Only one of them, the givenness of the recipient, has been selected in more than one model (i.e. in the two forward selections).

As opposed to the mixed models, the models without a random effect generalize well to previously unseen data. This does not necessarily mean that the British English dative alternation is best modelled with logistic regression models without a random effect. The models fit the data better when verb sense is included as a random effect. The fact that the mixed models do not generalize well to new data could be due to the relatively small size of our data set. In the near future, we therefore aim at extending our data set, employing the British National Corpus [4]. Since manually extending the data set in a way similar to that taken to reach the current data set of 930 instances is too labour-intensive, we aim at automatically extending the data set (in an approach similar to that taken in Lapata [12]), and automatically annotating it for the explanatory features in this paper. With the larger set, we hope to be able to model the underlying processes of the dative alternation, rather than modelling the instances that made it into our data set.

One of the drawbacks of variable selection is that different selection methods can lead to different models [11]. Accordingly, the six methods we applied have led to six different selections of variables and thus to six different models. How can we decide which is the optimal model for our purpose? Of course, the way to approach this issue depends on the goal of a specific research enterprise. For a researcher building a machine translation system, the best approach is probably to choose the highest prediction accuracy on previously unseen data. For linguists, however, the best approach may be less clear. In our project we want to combine the explanatory features suggested in previous research and test the combination on real data. We thus have hypotheses about what are explanatory features and what kind of effect they show in isolation, but it is unclear how specific features behave in combination with others. Also, we want a model that is interpretable in the framework of some linguistic theory and that, ideally, reflects the processes in human brains. It is uncertain how (and if) we can evaluate a model in this sense. Still, despite these difficulties, using techniques such as logistic regression is very useful for gaining insight in the relative contribution that different features have on the choices people make when there is syntactic variability. But contrary to what seems to be common in linguistics, researchers should be careful in choosing a single approach and drawing conclusions from one model only. Firm conclusions about mental processes can only be drawn if similar models are obtained with a number of different data sets. In addition, models derived from corpus data should be tested in psycholinguistic experiments.

Acknowledgments. I am grateful to Lou Boves, Hans van Halteren and Nelleke Oostdijk for their support and their useful comments on this paper.

References

1. Bates, D.: Fitting linear mixed models in R. R News 5(1), 27–30 (2005)
2. Blackwell, A.: Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology. Child Language 32, 535–562 (2005)
3. Bresnan, J., Cueni, A., Nikitina, T., Baayen, H.: Predicting the Dative Alternation. In: Bouma, G., Kraemer, I., Zwarts, J. (eds.) Cognitive Foundations of Interpretation, pp. 69–94. Royal Netherlands Academy of Science, Amsterdam (2007)
4. Burnard, L.: Reference Guide for the British National Corpus (XML Edition). Published for the British National Corpus Consortium. Research Technologies Service at Oxford University Computing Services (2007)
5. Godfrey, J., Holliman, E., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 517–520. IEEE Computer Society, Los Alamitos (1992)
6. Greenbaum, S.: Comparing English Worldwide: The International Corpus of English. Clarendon, Oxford (1996)
7. Gries, S.: Towards a corpus-based identification of prototypical instances of constructions. Annual Review of Cognitive Linguistics 1, 1–27 (2003)
8. Gries, S., Stefanowitsch, A.: Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. International Journal of Corpus Linguistics 9, 97–129 (2004)
9. Grondelaers, S., Speelman, D.: A variationist account of constituent ordering in presentative sentences in Belgian Dutch. Corpus Linguistics and Linguistic Theory 3(2), 161–193 (2007)
10. Haspelmath, M.: Ditransitive alignment splits and inverse alignment. Functions of Language 14(1), 79–102 (2007)
11. Izenman, A.: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer, New York (2008)
12. Lapata, M.: Acquiring lexical generalizations from corpora: a case study for diathesis alternations. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics, pp. 397–404. Morgan Kaufmann, San Francisco (1999)
13. Pinker, S.: Learnability and Cognition: The Acquisition of Argument Structure. MIT Press, Cambridge (1989)
14. Rietveld, T., van Hout, R.: Statistical Techniques for the Study of Language and Language Behavior. Mouton de Gruyter, Berlin (1993)
15. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2008)

16. Sheather, S.: A Modern Approach to Regression with R. Springer, New York (2009)
17. Siewierska, A., Hollmann, W.: Ditransitive clauses in English with special reference to Lancashire dialect. In: Hannay, M., van der Steen, G.J. (eds.) Structural-functional Studies in English Grammar: In Honor of Lachlan Mackenzie, pp. 83–102. John Benjamins, Amsterdam (2007)
18. West, B.T., Welch, K.B., Gałecki, A.T.: Linear Mixed Models: A practical guide using statistical software. Chapman & Hall/CRC, Boca Raton (2007)