VARIABLE SELECTION IN LOGISTIC REGRESSION: THE BRITISH ENGLISH DATIVE ALTERNATION

Daphne Theijssen
Department of Linguistics, Radboud University Nijmegen

Abstract. In this paper, we address 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 the help of 29 explanatory variables taken from the literature, we build two types of models: (1) with the verb sense included as a random effect (verb senses often have a bias towards one of the two variants), and (2) without a random effect. For each type, we build three different models by including all variables and keeping the significant ones, by sequentially adding the most predictive variable (forward regression), and by sequentially removing the least predictive variable (backward regression). Seeing that the six approaches lead to five different models, we advise researchers to be careful to base their conclusions solely on the one 'optimal' model they found.

1. Introduction

There are many linguistic phenomena that researchers have tried to explain on the basis of different partially explanatory features. Probabilistic modelling techniques can help in combining these 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 explanatory features should actually be included as variables in the model formula, (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.

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. Bresnan, et al. (2007)), (2) sequentially adding the most explanatory feature (forward), until no significant gain is obtained anymore (e.g. Grondelaers & Speelman (2007)), and (3) starting with a model containing all available features, and (backward) sequentially removing those that yield the lowest contribution (e.g. Blackwell (2005)). In general, researchers report on only one (optimal) model without giving clear motivations for their choices.

In this paper, we compare the three approaches in a case study: we apply them to a set of 915 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 variant (2).

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

The variables (explanations suggested in the literature) are taken from Bresnan et al's work on the dative alternation in American English.

Previous research (for example, Gries & Stefanowitsch (2004)) has indicated that the 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 variable in a regression model. Recently developed logistic regression models can handle these lexical biases by treating verb sense as a random effect (e.g. Bresnan et al. (2007)). 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. This leads to a total of six different models.

Our goal is to investigate the role of a random effect in a model of syntactic variation built with a medium-sized set of observations. In addition, we want to investigate whether it is justified to report only one 'optimal' regression model, if models can be built in three different ways. 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 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.

2. Related work

2.1. The dative alternation

Bresnan et al. (2007) 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 (Godfrey, et al. 1992). With the help of a logistic mixed-effect regression model with verb as a random effect, they were able to explain 95% of the variation. 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 prepositional dative constructions, the object first mentioned is the theme (*the book*), and the second object the recipient (*the student*). In double object constructions, 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 than the second object. The characteristics of the second object are generally the opposite: nongiven, nonpronominal, inanimate and indefinite.

According to Haspelmath (2007), 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.

Gries (2003) performed analyses with multiple variables that are similar to those in Bresnan et al. (2007), but applying a different technique (linear discriminant analysis or LDA) on a notably smaller data set consisting of only 117 instances from the British National Corpus (Burnard 2007). 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 & Stefanowitsch (2004) investigated the effect of verb biases in 1772 instances from the ICE-GB Corpus (Greenbaum 1996). When predicting the preferred dative construction for each verb, 82.2% of the constructions could be predicted correctly. It thus outperforms the majority baseline of 65.0% (always choosing the overall most frequent variant).

2.2. Variable selection in logistic regression

A recent and extensive textbook on modern statistical techniques is that by Izenman (2008). In chapter 5, he explains that variable selection is often needed to arrive at an interpretable model that reaches an acceptable prediction accuracy. Keeping too many variables may lead to overfitting, while a simpler model may suffer from underfitting. The risk of applying variable selection is that it optimizes the model for a particular data set. Using a slightly different data set may result in a completely different variable subset.

An approach to variable selection that is commonly used in linguistics is stepwise adding the most predictive variables to an empty model (e.g. Grondelaers & Speelman (2007)) or stepwise removing the least predictive variables from the full model (e.g. Blackwell (2005)). The main criticisms on 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 both methods may result in two different models that may not even be optimal, and (4) that each provides a single model, while there may be more than one optimal subset (Izenman 2008).

Another approach Izenman mentiones in the same section is to build all models with each possible subset and select those with the best results. An important objection to this approach is that it is computationally expensive to carry out. For this reason, we do not employ this method.

Instead, we follow Sheather (2009), who builds a model containing all variables that he expects to contribute to the model, and removes the insignificant ones (chapter 8). These expectations are based on plots of the variables that he made beforehand. Where desirable, he transformed the variables to give them more predictive power (e.g. by taking their log). As indicated by Izenman (2008), variable selection on the basis of a data set may lead to a model that is specific for that particular set. Since we also want to be able to compare our models to those found by Bresnan et al. (2007), we refrain from such preprocessing and use all variables they used in the variable selection process.

3. Data

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 (Greenbaum 1996).

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 (e.g. *Fold the short edges to the centre.*).

Following Bresnan et al. (2007), we ignored constructions with a preposition other than *to*, with a clausal object, with passive voice and with reversed constructions. 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 data sets

		nr of instances		
Type	Corpus	d.obj.	pr.dat.	Total
Spoken British English	ICE-GB	399	151	550
Written British English	ICE-GB	263	102	365
Total	ICE-GB	662	253	915

4. Method

4.1. Explanatory features

We adapt the explanatory features and their definitions from Bresnan et al. (2007) (Table 2), and manually annotate our data set following an annotation manual based on these definitions[1].

The table includes one new variable: *medium*. This tells us whether the construction was found in written or spoken text. It may well be that certain variables only play a role in one of the two mediums. In order to test this, we include the 14 (two-way) interactions between the variables taken from Bresnan et al. and the medium[2]. This leads to 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. The verb sense is the lemma of the verb together with its semantic class, e.g. *pay_a* for *pay* with an abstract meaning and *pay_t* when *pay* is used to describe a transfer of possession. In total, our data set contains 94 different verb senses

[1]The annotation manual is available online: http://lands.let.ru.nl/∼daphne/downloads.html.

[2]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 2: Features and their values (th=theme, rec=recipient). All nominal variables are transformed into binary variables with values 0 and 1. As a result, semantic verb class (*communication*, *abstract* or *transfer of possession*) is split into two effects: *verb=abstract* (0 or 1) and *verb=communication* (0 or 1). Cases with semantic verb class *transfer of possession* have value 0 for both variables.

Feature	Values	Description
rec = animate	1, 0	human or animal, or not
th = concrete	1, 0	with fixed form and/or space, or not
rec,th = definite	1, 0	definite pronoun, proper name or noun preceded by a definite determiner, or not
rec,th = given	1, 0	mentioned or evoked $\leq$20 clauses before, or not
length difference	-3.4-4.2	$\ln$(#words in th) $-$ $\ln$(#words in rec)
rec,th = plural	1, 0	plural in number, or not (singular)
rec = local	1, 0	first or second person (*I*, *you*), or not
rec,th = pronominal	1, 0	headed by a pronoun, or not
verb = abstract	1, 0	*give it some thought* is abstract, *tell him a story* is
verb = communication	1, 0	communication, *give him the book* is transfer
structural parallellism = present	1, 0	same variant used previously, or not
medium = written	1, 0	type of data is written, or not (spoken)

(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 & Stefanowitsch (2004), many verb senses show a bias towards one of the two constructions. The verb *pay* shows a clear bias towards the prepositional dative construction when it has an abstract meaning, but no bias when the literal transfer of possession is meant.

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. $\approx$ # 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	252	30	*do_a*	8	9	*pay_a*	2	12
give_c	65	10	*send_c*	9	7	*cause_a*	5	8
give_t	53	21	*lend_t*	8	7	*sell_t*	0	10
tell_c	67	1	*pay_t*	6	5	*owe_a*	2	6
send_t	41	15	*leave_a*	5	4	*explain_c*	0	6
show_c	37	9	*write_c*	4	5	*present_c*	0	6
offer_a	23	9				*read_c*	1	4
show_a	6	1						
offer_t	6	0						
tell_a	6	0						
wish_c	6	0						
bring_a	4	1						
bring_t	3	2						
hand_t	3	2						

4.2. Variable selection

Using the values of the variables (and the random effect), we establish a regression function that determines the log of the odds that the construction C in clause i (with verb sense j) is a prepositional dative. The prepositional dative is regarded a success (with value 1), while the double object construction is a failure (0). The regression function is defined as follows:

$$\ln odds(C_{ij} = 1) = \alpha + \sum_{k=1}^{29}(\beta_k V_{ijk}) \quad (+r_j). \tag{1}$$

The α is the intercept of the function. $\beta_k V_{ijk}$ are the weights β and values V_{ij} of the 29 variables k. The optional random effect r_j is normally distributed with mean zero ($r_j \sim N(0, \sigma)$). The optimal values for the function parameters α, β_k and r_j are found with the help of Maximum Likelihood Estimation[3]. The outcome of the regression enables us to use the model as a classifier: all cases with $\ln odds(C_{ij} = 1) \geq t$ are classified as prepositional dative, all with $\ln odds(C_{ij} = 1) < t$ as double object. The letter t is a threshold, which we set to 0. With this threshold, all instances for which the regression function outputs a negative log odds are classified as double object constructions, all other instances as prepositional dative.

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

For the second approach, being forward sequential regression, we start with an empty model and sequentially add the variable that is most predictive. As Izenman (2008) warns us, deciding which variable to keep is not trivial. We decide to keep the variable that yields the highest area under the ROC (Receiver Operating Characteristics) curve. This curve is a plot of the correctly classified positive instances (prepositional dative) and the incorrectly classified positive instances. The area under it (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:

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

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[4]. We add the most predictive variables to the model as long as it gives an improvement over the AUC of the model without the variable. An interaction of variable A with *Medium* is only included when the resulting AUC is higher than that reached after adding the single variable A[5]. Two AUC are considered different when the difference is higher than a threshold. We set the threshold to 0.002.[6]

For the third approach (backward sequential regression), we use the opposite procedure: we start with the full model, containing all 29 features, and sequentially leave out the variable A that yields the model with the highest AUC that is not lower than the AUC for the model with A. When the AUC of a model without variable A does not differ from

[3]We use the functions `glm()` and `lmer()` (Bates 2005) in `R` (R Development Core Team 2008).

[4]We use the function `somers2()` created in `R` (R Development Core Team 2008) by Frank Harrell.

[5]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 (Rietveld & van Hout 2008).

[6]The threshold value has been established experimentally.

the AUC of the model without the interaction of A with Medium, we remove the interaction. Again, AUCs are only considered different when the difference is at least the threshold (again set to 0.002).

We evaluate the models with and without random effects by establishing the model quality (training and testing on all 915 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 the 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 915 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 915 instances.

5. Results

5.1. Mixed models

Table 4 gives the model fit 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 NP-NP) in all settings, also when testing on new data ($p < 0.001$ for the three models, Wilcoxon paired signed rank test).

Table 4: Model fit and prediction accuracy of the regression models with *verb sense* as a random effect

			model fit (train=test)		*10-fold cv*
selection	#variables	baseline	AUC	accuracy	aver. accuracy
1. significant	5	0.723	0.979	0.936	0.825
2. forward	4	0.723	0.980	0.938	0.832
3. backward	4	0.723	0.980	0.938	0.832

When training and testing on all 915 instances, the mixed models reach a considerable 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 well to previously unseen data.

The significant effects in the models resulting from the three approaches are presented in Table 5. The directions of the main effects are the same as those for American English (Bresnan et al. 2007), as presented in Section 2.1.

The forward (2) and backward (3) selection approaches lead to the same regression model. The differences between this model and the one obtained by keeping the significant variables (1) may be caused by the fact that the information contained in the variables shows considerable overlap. For instance, pronominal objects are also typically discourse given. A significant effect for the one variable may therefore decrease the possibility of

Table 5: Coefficients of significant effects in (mixed) regression models with verb sense as random effect, trained on all 915 instances, *** p<0.001 ** p<0.01 * p<0.05. The last column indicates the direction towards the prepositional dative (p.dat.) and double object construction (d.obj.).

Effect	1. significant		2. forward		3. backward		
th=pronominal, medium=written	-2.01	*					
length difference	-2.52	***	-2.41	***	-2.41	***	d.obj.
rec=local	-2.68	***	-1.86	***	-1.86	***	↑
rec=given			-1.48	***	-1.48	***	|
th=definite	1.67	***					|
th=pronominal	2.15	***					↓
th=given			2.32	***	2.32	***	p.dat.
(intercept)	1.27	**	2.53	***	2.53	***	

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, while it is the other way around for the model with the significant variables from the full model.

Only the model obtained by keeping the significant variables in the full model contains an 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). Whether there really exists a difference in the effect of the pronominality of the theme in speech and writing is not clear, since only one model shows this difference.

What also remains unclear, is which of the two models is more suitable for explaining the British English dative alternation. Seeing the differences between the significant effects found in the two 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 915 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 fit 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 fit (train=test)		*10-fold cv*
selection	#variables	baseline	AUC	accuracy	aver. accuracy
1. significant	5	0.723	0.934	0.882	0.882
2. forward	5	0.723	0.941	0.883	0.870
3. backward	8	0.723	0.945	0.882	0.875

The model fit figures AUC and accuracy are considerably lower than the figures reached with the mixed models. 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 fit accuracy (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 forward/backward mixed model following the Wilcoxon paired signed rank test). Apparently the simpler models (those 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 as expected, but the three models disagree on the significance of the variables. Four variables have significant effects in two of the three models, one (the definiteness of the theme) only has an effect in the stepwise forward selection model. Only the concreteness of the theme is selected in all three approaches, as opposed to the mixed-effect approach of the previous section, where it was not selected at all. According to all three models, speakers tend to use the double object construction when the theme is longer than the recipient, and when the recipient is pronominal. The backward selection model (3), however, shows that the effect of length difference is especially strong in speech, while the effect of the pronominality of the recipient is particularly strong in writing. As in the previous section, where the one significant interaction (medium with pronominality of theme) was only found in model 1, it is not clear whether this difference really exists.

Table 7: Coefficients of significant effects in regression models (without random effect), trained on all 915 instances, *** p<0.001 ** p<0.01 * p<0.05. The last column indicates the direction towards the prepositional dative (p.dat.) and double object construction (d.obj.).

Effect	1. significant		2. forward		3. backward		
length difference, medium=spoken					-2.29	***	
rec=pronominal, medium=written					-2.07	***	d.obj.
length difference	-1.75	***	-1.97	***			↑
rec=pronominal, medium=spoken					-1.71	***	|
length difference, medium=written					-1.49	***	|
rec=definite			-1.18	***	-1.20	***	|
rec=local	-1.13	***	-1.20	***			|
rec=pronominal	-1.15	***	-1.20	***			|
th=definite			1.12	***			|
(intercept)					1.37	***	↓
th=given			0.95	**	1.44	***	p.dat.
th=concrete	1.50	***	1.52	***	1.27	***	

Surprisingly enough, excluding the verb sense as a random effect has not resulted in a significant effect for semantic verb class in any of the models. Given the high model fit we found for the models with verb sense as a random effect, and the predictive quality of verb sense found in previous research (Gries & Stefanowitsch 2004), one would expect that having information about the semantic verb class would be useful in the models without this random effect. Apparently, the effect is not strong enough.

6. 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 915 instances taken from the ICE-GB Corpus, and took the explanatory factors suggested by Bresnan et al. (2007). 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 sequentially adding the most predictive variables, and (3) starting with a model with all 29 features and backwardly sequentially 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 (Bresnan et al. 2007): pronominal, relatively short, local, discourse given, definite and concrete objects typically precede objects with the opposite characteristics. Contrary to the observations in Haspelmath (2007), we have no reason to believe that the dative alternation in British English differs from that in American English. We have found no clear indications of differences between the dative alternation in speech and writing either: only three variables were selected in interaction with medium, and they occurred in only one model.

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 an artefact of lack of data instances. In the near future, we therefore aim at extending our data set, employing the British National Corpus (Burnard 2007). Since manually extending the data set in a way similar to that taken to reach the current data set of 915 instances is too labour-intensive, we aim at automatically extending the data set (in an approach similar to that taken in Lapata (1999)), 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 methods can lead to different models (Izenman 2008). Unsurprisingly, the six approaches we took have led to five different models. How can we decide which is the optimal model for our purpose? Of course, the approach depends on your goal. For a researcher building a machine translation system, the goal will probably be to reach the highest prediction accuracy on previously unseen data. For linguists, however, the goal is more complex. 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 the explanatory features and what kind of effect they show, but it is unclear how they behave in combination with the others. Also, we want a model that is interpretable and, ideally, reflects the processes in our 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 statistical characteristics that play a role in syntactic variability. But contrary to what is common in linguistics, researchers should be careful in choosing a single approach and drawing conclusions from one model only.

References

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

A. A. Blackwell (2005). 'Acquiring the English adjective lexicon: relationships with input properties and adjectival semantic typology'. *Child Language* (32):535–562.

J. Bresnan, et al. (2007). 'Predicting the Dative Alternation'. In G. Bouma, I. Kraemer, & J. Zwarts (eds.), *Cognitive Foundations of Interpretation*, pp. 69–94. Royal Netherlands Academy of Science, Amsterdam, the Netherlands.

L. Burnard (ed.) (2007). *Reference Guide for the British National Corpus (XML Edition)*. Research Technologies Service at Oxford University Computing Services, Published for the British National Corpus Consortium.

J. J. Godfrey, et al. (1992). 'Switchboard: Telephone speech corpus for research and development'. In *Proceedings of ICASSP-92*, pp. 517–520, San Fransisco, U.S.A.

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

S. T. Gries (2003). 'Towards a corpus-based identification of prototypical instances of constructions'. *Annual Review of Cognitive Linguistics* (1):1–27.

S. T. Gries & A. Stefanowitsch (2004). 'Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations''. *International Journal of Corpus Linguistics* (9):97–129.

S. Grondelaers & D. Speelman (2007). 'A variationist account of constituent ordering in presentative sentences in Belgian Dutch'. *Corpus Linguistics and Linguistic Theory* **3**(2):161–193.

M. Haspelmath (2007). 'Ditransitive alignment splits and inverse alignment'. *Functions of Language* **14**(1):79–102.

A. J. Izenman (2008). *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*. Springer, New York, USA.

M. Lapata (1999). 'Acquiring lexical generalizations from corpora: a case study for diathesis alternations'. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics (ACL)*, pp. 397–404, Morristown, USA.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

T. Rietveld & R. van Hout (2008). *Statistical Techniques for the Study of Language and Language Behavior*. Mouton de Gruyter, Berlin, Germany.

S. J. Sheather (2009). *A Modern Approach to Regression with R*. Springer, New York, USA.