Journal of the
Royal Statistical Society

SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2013)
176, *Part* 3, *pp.* 777–793

Statistical analysis of varieties of English

Christopher F. H. Nam,

University of Warwick, Coventry, UK

Sach Mukherjee

Netherlands Cancer Institute, Amsterdam, The Netherlands, and University of Warwick, Coventry, UK

and Marco Schilk and Joybrato Mukherjee

Justus Liebig Universität Giessen, Germany

[Received April 2010. Final revision May 2012]

Summary. Linguistic corpora are databases of text which are linguistically marked up or otherwise structured and designed to be representative of a specific language. The growing availability of such corpora has brought with it opportunities for statistical analysis. The paper develops and uses statistical approaches to address questions pertaining to an important linguistic phenomenon: the use of different syntactic alternatives. We present a model-selection-based approach for determining possible driving attributes affecting verb complementation for written sentence constructions using the verb ‘give’ in three varieties of English. We are interested in explaining the choice of alternatives in terms of a variety of sentence level linguistic features such as the meaning of the verb, in addition to the country of origin.

Keywords: Corpus linguistics; Dative alternation; Feature selection; Model selection

1. Introduction

Corpus linguistics is a discipline within linguistics that is concerned with the study of both written and spoken ‘real world’ text. Corpora are computer readable databases of text which are linguistically marked up or otherwise structured and designed to be representative of a specific language. The growing availability of such corpora has encouraged more quantitative research in linguistics and brought with it opportunities for statistical analysis in the field.

An area of on-going interest in linguistics is the analysis of complementation patterns of verbs. This includes the study of grammatical alternatives which result in different sentence constructions but convey similar messages. We refer to this as pattern selection in this paper for simplicity (see Mukherjee (2001)). Research into the varieties of English have shown that pattern selection may differ across varieties, and that they may be affected by a number of linguistic attributes (Bresnan and Hay, 2008). ‘Varieties’ are subtypes of English that differ with respect to a key factor such as social aspects, time or geographical region. We focus in this paper on regional varieties of English, which we refer to as varieties from now on. In particular, we focus on English as a second language regions such as South Asian post-colonial speech communities. English as a second language regions are defined in Görlach (1991).

Address for correspondence: Christopher F. H. Nam, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: c.f.h.nam@warwick.ac.uk

0964–1998/13/176777

A data-driven approach in understanding pattern selection is relatively new in contrast with the non-quantitative and traditional intuition-based approach where a set of deterministic rules founded on attributes of sentences would determine the resultant sentence construction (McArthur and McArthur, 2005). However, this deterministic set-up has been found to be inadequate in capturing the inherent uncertainty of language (Johnson, 2008). In light of this uncertainty, recent work has emphasized more quantitative approaches with Stefanowitsch and Gries (2003) analysing the interaction between words and constructions by using standard test statistics, and Bresnan *et al.* (2007), predicting a certain type of pattern selection by using logistic regression with respect to linguistic attributes to a high degree of accuracy. In addition to this, Bresnan and Hay (2008) used the same approach in determining the differences between spoken New Zealand and American English.

This paper, which builds on the basic framework discussed in Bresnan *et al.* (2007), Bresnan and Hay (2008) and Bresnan and Ford (2010), explores the particular complementation patterns of the verb ‘give’ in three varieties of the English language, namely British, Indian and Pakistani. We model the sentence construction, the dependent variable, by using a collection of linguistic attributes including the written sentence’s country of origin (variety), the independent variables, obtained from linguistically marked-up corpora. By taking a model-selection-based approach to linguistically marked-up corpora, we aim to determine potential influential linguistic attributes in driving the resultant sentence construction.

Current comparable approaches such as Bresnan and Ford (2010) have modelled the complementation patterns for ‘give’ with respect to a fixed collection of linguistic attributes, with no model selection procedure being performed. In these cases, differences between the varieties have been investigated by analysing the values of coefficients and their effect on the outcome. There is no reason to believe that all linguistic attributes that were considered in Bresnan and Ford (2010) are required: we use model selection to determine whether a simpler, more parsimonious model, with fewer linguistic attributes, can be used. We also consider whether differences between the varieties can be highlighted by stratifying the data with respect to each variety, and performing the model selection procedure on each of these stratified data sets separately. Differences may thus be highlighted by the selection of different linguistic attributes and how they interact with each other.

There is considerable interest within linguistics in better understanding the complexity of verb complementation. This is a scientific application in the context of understanding the production processes of languages but there are also some practical applications. Certain sentence constructions may be more commonly used in some places than others and, for this reason, verb complementation can potentially shed light on differences between varieties of the same language. In a similar way, the description of subtle grammatical differences can help in tracing the development of a language over time, on account of the changing popularity of certain constructions (Bresnan and Hay, 2008). In natural language processing, understanding syntactic choice can help in the design of computer programs which mimic the use of natural language by humans. Equally, in teaching and learning English as a foreign language, understanding grammatical variations can help to provide guidelines for learners. It can also highlight a particular writing style and authorship (say, for example, the differences between Shakespeare and J. K. Rowling) and may therefore be used in quantitative studies in English literature (Bresnan *et al.*, 2007).

We focus on the grammatical variation in written sentence samples of the verb ‘give’ in each of the three varieties, available from corpora. However, computerized linguistic mark-up remains limited in its ability to discern subtle linguistic and semantic features. We therefore manually marked up the data with respect to a number of attributes which are believed to have some influence on the syntactic outcome of the final written sentence.

The set-up of this problem leads quite naturally to a regression formulation where we treat the resultant sentence construction as the response and the linguistic attributes, in addition to variety, as explanatory variables. As the response is categorical, we use multinomial logistic regression, which is a generalization of standard binary logistic regression, as our model. As it is linguistically plausible that attributes may jointly influence pattern selection, this motivates a need to go beyond standard marginal statistics as in Stefanowitsch and Gries (2003) and to consider appropriate multiplicative interactions.

There are a variety of ways in which the model selection procedure can be performed. Motivated by the fact that our data set is relatively small, a result of the time and effort required to collate and tag individually the attributes in each sentence, we perform our model selection by using the Akaike information criterion AIC (Akaike, 1974). Furthermore, the simplicity of AIC aids exploration of thousands of models constructed from combinations of linguistic attributes.

The remainder of the paper is organized as follows. We first introduce some relevant background material in linguistics. We then discuss the key statistical methods that are used in our analysis. We go on to present the results obtained by using the corpus data and finally discuss the shortcomings of our work, highlighting key directions for further research.

2. Linguistics background

This section presents a self-contained introduction to the linguistics that is relevant to this study. For further details we refer the interested reader to Schilk (2011), Bresnan and Hay (2008) and McArthur and McArthur (2005).

2.1. Verb complementation

There are usually a variety of grammatically correct ways in which language users can convey a given message.

One area in which this is often so is verb complementation; grammatical patterns which accompany verbs. Here, language users often have several alternative ways of constructing a sentence. For example, the verb ‘like’ can be proceeded by the infinitive of another verb, or the continuous inflection, as in -*ing*, of the same verb. In effect, ‘I like to dance’ and ‘I like dancing’ both convey largely the same meaning. Although the meaning of the sentences ‘I like to dance’ and ‘I like dancing’ seem to be identical on a surface level, we assume that there are underlying reasons for the existence of these alternatives since languages usually lose redundancies over time.

We focus on the alternative complementations of the ditransitive verb ‘give’ because of its common use and relative simplicity. We define ditransitive verbs as verbs that potentially take both a direct and an indirect object, both of these objects being possibly realized by noun phrases (see Mukherjee (2005)). A particular property of ‘give’, and many other ditransitive verbs, which has previously been studied is dative alternation: the ‘existence of pairs of alternative paraphrases for dative verbs’ (Bresnan and Hay, 2008).

To fix ideas and terminology, we consider an example. Suppose that we want to convey the message that a person, named John, gave another person, Mary, a book. This message can be conveyed in six main ways as displayed in Table 1, with the numbers in parentheses denoting the frequency of the corresponding construction in our particular data set.

Each sentence can be decomposed into three common components, namely

(a) the *agent* who is performing the act of ‘giving’ (‘John’),
(b) the *recipient* who is on the receiving end of this ‘giving’ act (‘Mary’, the indirect object) and

Table 1. Complementation patterns of 'give'

Construction	*Voice*	
	Active	*Passive*
Double object DO	John gave Mary the book (217)	Mary was given the book by John (76)
Prepositional dative PD	John gave the book to Mary (142)	The book was given by John to Mary (37)
Monotransitive MT	John gave the book (130)	The book was given by John (34)

(c) the *patient*, the theme that is being transferred from the agent to the recipient ('the book', the direct object).

We observe that the constructions can firstly be separated by the active and passive voices which by definition are semantically synonymous. They can be seen as exchanging the subject of the sentences from the agent in the active voice, to the recipient or patient in the passive voice. Although the active voice is more commonly used in daily language, passive voices are used to emphasize the particular verb of action on the recipient or patient.

Within both voices, there are three possible constructions that one can use; the double object DO, the prepositional dative PD and the monotransitive MT. Dative alternation is defined more specifically between the DO- and PD-construction with the MT-construction not being included owing to the absence of the recipient and thus loss of information associated with it. The recipient can, however, often be deduced from the context. For example, in 'The student gave a presentation', it is naturally implied that the recipient is some sort of audience. Indian and Pakistani English users display a tendency to use this construction more frequently, particularly for other ditransitive verbs such as 'send' and 'offer', and thus its inclusion gives a fair representation of these two varieties (Schilk, 2011).

To identify which construction is being used, we can use the set of rules that is outlined in Table 2 which are founded on the location of the agent, recipient and patient relative to the verb within the sentence.

2.2. Linguistic attributes

We consider the following eight attributes of the message that we wish to convey and may be responsible in influencing the different syntactic constructions: our response. We use broadly the same attributes as those used in Bresnan and Hay (2008) and Bresnan and Ford (2010) with some modifications made to coding due to sparsity of the data set by collapsing some of the attribute category values.

(a) *Syntactic complexity, SC, recipient more complex*, compares the syntactic complexity of the recipient against that of the patient relatively. Syntactic complexity is measured by the length of each component (the number of graphemic words). This attribute can take two cases denoting whether the recipient is less or equally complex ($R \leqslant P$), or more complex ($R > P$) than the patient. For example, in the sentence 'John gave the book to Mary', the recipient syntactic complexity is 1 ('Mary'), whereas the patient syntactic complexity is 2 ('the book'). Thus this attribute would be coded for the recipient being less or equally

Table 2. Method of classifying constructions for ditransitive sentences†

Construction	*Active (subject ≡ agent)*		*Passive (subject ≡ patient or recipient)*
	Patient position relative to verb	*Recipient position relative to verb*	*Subject of sentence*
Double object DO	2nd	1st	Recipient
Prepositional dative PD	1st	2nd	Patient
Monotransitive MT	1st	—‡	Patient (no recipient)

†Noting that the position of the components agent, recipient and patient differ in each possible sentence structure, we provide a simple rule for identifying which sentence construction is being used.
‡Not applicable.

complex than the patient, $R \leqslant P$. In the case of MT-constructions where the recipient is not present, we code this attribute as $R \leqslant P$.

(b) *Animacy, P.Ani and R.Ani*, is coded identically for both recipient and patient and can take only two values; *animate* A and *inanimate* I. Animate includes humans, animals, humanoids (those having human-like forms and qualities) and groups without a collective purpose or voice, whereas inanimate covers all other categories.

(c) *Discourse accessibility, P.DA and R.DA*, is a binary attribute coded identically for patient and recipient. The attribute considers whether the patient or recipient is 'given' G (i.e. has been mentioned explicitly in the previous 10 lines of text or is an 'I' or 'You') or 'not given' NG.

(d) *Pronominality, P.Pro and R.Pro*, is coded identically for both patient and recipient; this binary attribute considers whether phrases are profiled by a pronoun or not (P and NP respectively). Pronouns are function words that are used to substitute for a noun or an object such as 'me, you, it, ...'.

(e) *Semantic class, SemC*, is coded once for a sentence; the attribute considers the specific use of the verb 'give'. We consider three classes of which the verb 'give' can be used in; *concrete* (Con, the transfer of a concrete patient, such as a 'book'), *informative* (Inf, the transfer of information, such as a 'lecture') and *abstract* (Abs, all other cases such as figures of speech, such as 'a hand', as in to help someone).

2.3. The data sets

Our data originally consist of 250 sentences containing the word 'give' for each of the three varieties, over all six possible sentence constructions. Sentences are collated from three different sources: the *British National Corpus* for British English, the *Times of India* for Indian English and the *Daily Times* for Pakistani English. The *British National Corpus* is a typical corpus; a 100-million-word collection of samples of written and spoken language from a wide range of sources (from newspaper articles to university and school essays, both published and unpublished) with the aim of best representing modern British English (late 20th century) (Burnard, 2000). Only the newspaper section of the *British National Corpus* has been used in forming our data set. The *Times of India* and *Daily Times* are popular English newspapers in their respective countries, and sentences have been collated from on-line articles. This 'Web page to mega-corpus' method as proposed by Hoffmann (2007) is necessary since, at present, no conventional balanced corpus exists for Pakistani English, and the international corpus of English which contains a corpus of both British and Indian English is considered too small to provide insight into the sentence constructions of interest. We denote the variable for variety as *Cou*, taking the three possible values GB, Ind and Pak.

Automated mark-up computer programs remain too approximate for a study of this kind, especially with respect to semantics. For this reason, we manually marked up the sentences. However, the manual nature of annotation means that our data set is small with a total of about 750 sentences: 250 per variety.

The 250 sentences containing 'give' for each variety are obtained via a stratified sampling procedure with respect to the different forms that 'give' can appear as. These are give, gave, given, giving and gives. We then randomly sample sentences containing each of these word forms exactly. The number of sentences that we sample for each form of 'give' is proportional to its occurrence in the entire corpus over the occurrence of all forms of 'give'.

Some details of the manual annotation procedure are as follows: 25 sentences containing the verb 'give' required approximately 60 min to annotate with respect to each of the aforementioned linguistic attributes and sentence constructions. The linguistic attributes that are associated with discourse accessibility were found to be the most time consuming. Only one individual performed the marking-up process for all 750 sentences, whose training consists of a Masters degree in linguistics and who has had involvement in comparable research projects. In addition, both linguistic authors double-checked any ambiguous cases. This thus provides a satisfactory level of consistency in the marking-up procedure.

The collected 250 sentences per variety are further reduced to 212, 222 and 202 for British, Indian and Pakistani English respectively. The reason for the further reduction in the data sets is because there are instances in the English language where 'give' is not used in its usual context. Examples include particle verb constructions such as 'give up' and 'give into' or when 'give' is used in a participle construction such as 'Given John's condition, he performed spectacularly'. We also note that derivative constructions also exist but have been collapsed under the six root constructions considered, owing to their sparseness in the data set and to maintain stability in our analysis. Table 3 displays the frequency breakdown of the data set for each linguistic attribute, stratifying with respect to the three varieties and when we consider a four-level and six-level sentence construction response.

3. Statistical methods

This section presents the key statistical methods that were used in our analysis. We begin with establishing notation, a preliminary association measure between a single explanatory variable and response, before discussing the model and the model selection technique that are used in our framework.

3.1. Notation

We denote our response by Y, which is influenced by some combination of subsets from the explanatory variables $\mathbf{Z}=(Z_1 \ldots Z_d)$; sample size is denoted by n and d explanatory variables are available. In the context of our linguistic problem, the resultant sentence construction is the response variable and the linguistic attributes that were marked up in the sentence and the variety are our explanatory variables. We note that both the response variable and the explanatory variables are of a categorical nominal nature and, consequently, the explanatory variables are factors. Factors can be denoted by several binary variables when used in generalized linear models. We denote the value settings for a given collection of explanatory variables in its binary

Table 3. Table of frequencies for the linguistic attributes values in the data set for both four and six response levels†

Attribute	*Value*	*Results for 4-level response*				*Results for 6-level response*			
		GB	*Ind*	*Pak*	*Pooled*	*GB*	*Ind*	*Pak*	*Pooled*
Total number of sentences		163	153	156	472	212	222	202	636
Construction, Cons	ADO	100	59	58	217	100	59	58	217
	APD	32	54	56	142	32	54	56	142
	PDO	23	28	25	76	23	28	25	76
	PPD	8	12	17	37	8	12	17	37
	AMT			—		40	54	36	130
	PMT			—		9	15	10	34
Syntactic complexity, SC	$R \leqslant P$	130	107	112	349	179	176	158	513
	$R > P$	33	46	44	123	33	46	44	123
Semantic class, SemC	Inf	10	9	12	31	28	41	36	105
	Abs	122	106	96	324	147	139	113	399
	Con	31	38	48	117	37	42	53	132
Recipient animacy, R.Ani	I	53	48	64	165			—	
	A	110	105	92	307			—	
Recipient discourse accessibility, R.DA	NG	76	79	62	217			—	
	G	87	74	94	255			—	
Recipient pronominality, R.Pro	NP	105	106	126	337			—	
	P	58	47	30	135			—	
Patient animacy, P.Ani	I	160	153	155	468	209	222	201	632
	A	3	0	1	4	3	0	1	4
Patient discourse accessibility, P.DA	NG	146	133	128	407	190	194	169	553
	G	17	20	28	65	22	28	33	83
Patient pronominality, P.Pro	NP	161	152	154	467	210	221	200	631
	P	2	1	2	5	2	1	2	5

†We display the frequencies with respect to each variety and over the entire data set, irrespective of the variety.

representation by $\mathbf{X}=(X_1 \ldots X_{d^*})$ where d^*, the number of binary variables that are used to denote the chosen explanatory variables settings, will vary with respect to chosen collection of explanatory variables and the number of levels that each factor can take.

Throughout this section, we make the distinction between the terms 'factors' and 'covariates'. We use the former to mean the linguistic attributes as they appear in the data set, and 'covariates' to mean the binary representation of the explanatory variables as they appear in a design matrix.

3.2. Uncertainty coefficient: measure of association

We measure the possible association between variables A and B say, independent of all other variables, by the uncertainty coefficient, which is defined as

$$U = -\frac{\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \mu_{ab} \log(\mu_{ab}/\mu_{a+}\mu_{+b})}{\sum_{b\in\mathcal{B}} \mu_{+b} \log(\mu_{+b})} \tag{1}$$

where A and B can take values from sets $\mathcal{A}$ and $\mathcal{B}$ respectively with $|\mathcal{A}|, |\mathcal{B}| \geqslant 2$, $\mu_{ab} = P(A=a, B=b)$ and $\mu_{a+} = \Sigma_b \, \mu_{ab}$ denotes the marginal probability of A ($\forall a \in \mathcal{A}; \forall b \in \mathcal{B}$). The coefficient is well defined when $\mu_{+b} > 0$ for more than one b.

$U \in [0, 1]$ with $U=0$ suggesting that the two considered variables, A and B, are independent, whereas $U=1$ suggests dependence and thus association, such that $\forall a \in \mathcal{A}$, and $\exists b \in \mathcal{B}$ such that $P(B=b|A=a)=1$, meaning that we can perfectly infer on variable B if we know the value of A. This also indicates that there is a lack of conditional variance. To deal with empty cells, we add a small constant (0.001) to all cell counts (Everitt, 1992). Calculating the uncertainty coefficient between the response sentence construction and each linguistic attribute may therefore highlight a preliminary indication of influential attributes.

3.3. Multinomial logistic regression

As our response can take several categories, the number of responses which fall into each response category follows a multinomial distribution. This naturally leads to modelling the response via multinomial logistic regression, which is a generalization of standard binary logistic regression. As our study concerns only nominal responses, we focus solely on the multinomial logistic regression for the nominal case.

We implement multinomial logistic regression by using the baseline category logit models method (Hosmer and Lemeshow, 2000). Without loss of generality, suppose that our response Y takes values from the set $\mathcal{S}=\{1,2,\ldots,S\}$. We denote $y_{ij}=1$ if observation i's response falls in category $j \in \mathcal{S}$; otherwise $y_{ij}=0$. We set $R \in \mathcal{S}$ to be the baseline category response level with which we compare all other response levels via the log-odds-ratio. R can be set to any of the response levels with common practice setting R to be the most common response category in the data.

We define $\pi_{ij}=\pi_j(\mathbf{x}_i)=P(Y_i=j|\mathbf{x}_i)$ for fixed covariates $\mathbf{x}_i=(1,x_{i1},x_{i2},\ldots,x_{id^*})^{\mathrm{T}}$ for observation i such that $\Sigma_{j=1}^{S}\pi_{ij}=1$. We thus consider the counts in each of the S categories as multinomial with probabilities $\pi_{i1},\pi_{i2},\ldots,\pi_{iS}$. We construct $S-1$ simultaneous logit models, comparing each response category with the baseline category R. Thus

$$h_j(\mathbf{x}_i)=\log\left\{\frac{\pi_j(\mathbf{x}_i)}{\pi_R(\mathbf{x}_i)}\right\}=\beta_j^{\mathrm{T}}\mathbf{x}_i, \tag{2}$$

$$\pi_{ij}=\pi_j(\mathbf{x}_i)=\frac{\exp(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i)}{1+\sum_{j\in\mathcal{S}\backslash R}\exp(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_i)} \tag{3}$$

$\forall j \in \mathcal{S}\backslash R$, $i=1,2,\ldots,n$. These describe the simultaneous effects of $\mathbf{x}_i$ on these $S-1$ logits and $\beta_j=(\beta_{j0},\beta_{j1},\ldots,\beta_{jd^*})^{\mathrm{T}}$. We note that we are in effect constructing $S-1$ binary logistic regression models for each possible pair with the base reference level. The multinomial response probabilities $\{\pi_j(\mathbf{x}_i)\}$ are calculated from equation (3):

$$\log\left\{\prod_{i=1}^{n}\left(\prod_{j=1}^{S}\pi_{ij}^{y_{ij}}\right)\right\}=\sum_{j\in\mathcal{S}\backslash R}\left(\sum_{k=0}^{d^*}\beta_{jk}\sum_{i=1}^{n}x_{ik}y_{ij}\right)-\sum_{i=1}^{n}\log\left\{1+\sum_{j\in\mathcal{S}\backslash R}\exp(\beta_j^{\mathrm{T}}\mathbf{x}_i)\right\}. \tag{4}$$

This is a concave function, and thus the maximum likelihood coefficients $\boldsymbol{\beta}_j$ can be estimated via Newton–Raphson iteration.

3.4. Model selection

This subsection outlines how a restricted set of models was enumerated and scored.

3.4.1. Model enumeration

Let $\mathcal{M}$ be the model space where a model M is defined from a subset of factors $\{Z_1,Z_2,\ldots,Z_d\}$ and the interactions between them. Given the nature of the problem and the data, we consider a subspace of the complete model space restricted to relatively parsimonious models. Our model subspace is defined as follows.

(a) Consider all subsets of factors, with $1 \leqslant k \leqslant d$, of maximum cardinality k from $\{Z_1, Z_2, \ldots, Z_d\}$.
(b) For each subset of factors, consider the set of possible terms generated by it, considering up to two-way interaction terms.
(c) Construct all possible models by inclusion–exclusion of elements from the set of terms.

For example, for $\{A, B\} \subseteq \{Z_1, Z_2, \ldots, Z_d\}$, we can construct the following models to model the response variable Y:

$$Y \sim 1, \qquad Y \sim A, \qquad Y \sim B,$$
$$Y \sim A + B, \qquad Y \sim A*B$$

where $A*B = A + B + A{:}B$, with $A{:}B$ denoting a multiplicative term.

Linguistically, these restrictions correspond to the belief that no more than k linguistic attributes will influence the resultant sentence construction, and large interaction terms of order greater than 2, will have little effect on the response. These restrictions increase stability in the fitted models since our explanatory variables are categorical and thus some factor configurations may not be realized in the data, particularly at higher orders of interaction. These restrictions also concur with obtaining a parsimonious sparse model with which we can interpret and communicate with ease. The restricted model subspace should therefore lead to models which are linguistically interpretable.

Under the restrictions that are outlined above, all models considered can be enumerated and scored by exhaustive enumeration. For less restricted model spaces where an exhaustive approach is not feasible, Monte Carlo methods could be employed (for example, see Dellaportas *et al.* (2002)).

The model subspace considered is as follows: for each possible subset of factors, we consider the set of terms which can be generated by this subset. We then consider all possible models generated by the inclusion or exclusion of terms from this set. At most, 23983 distinct, non-equivalent models are considered when seven factors are available ($d = 7$), with maximum cardinality set at 5 ($k = 5$).

3.4.2. *Model selection methods*

Having defined the model subspace that we explore, we need to consider the most appropriate model to use. We proceed by using a penalized likelihood to consider not only the fit of the model to the data (via the likelihood), but also model complexity.

Let $\boldsymbol{\theta}$ be a vector of parameters for the associated model M. Let p denote the complexity that is associated with this model, such as the effective degrees of freedom associated with $\boldsymbol{\theta}$. Then we define the likelihood under the model M with parameters $\boldsymbol{\theta}$ to be $\mathcal{L}_M(\boldsymbol{\theta}) = \mathbb{P}(Y|\boldsymbol{\theta}, M)$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimates for the model. We use the Akaike information criterion AIC (Akaike, 1974) as our model selection criterion. AIC is defined as follows:

$$\text{AIC} = -2 \log\{\mathcal{L}_M(\hat{\boldsymbol{\theta}})\} + 2p; \tag{5}$$

selected model,

$$M^*_{\hat{\boldsymbol{\theta}}} = \arg \min_{M_{\hat{\boldsymbol{\theta}}}} \{\text{AIC}(\hat{\boldsymbol{\theta}})\}. \tag{6}$$

Models with scores within a ±2-range are considered to be equivalent in their suitability. The term $2p$ acts as our penalization term in this penalized likelihood equation, taking into consideration the complexity of the model. Related model selection criteria include the Bayesian information criterion, which has a harsher penalty of $p \log(n)$, but we prefer AIC here because of the nature of the problem and the size of the data set. Müller and Welsh (2010) have provided a good review of different possible penalties for penalized likelihood model selection criteria.

4. Results

We perform two sets of analysis: the first has four response levels corresponding to the active and passive voices of DO and PD. We thus in turn analyse dative alternation specifically and consider in our model selection exploration technique the aforementioned linguistic attributes and variety as explanatory variables. As the frequency levels for P.Ani and P.Pro are sparse at some factor levels, we have excluded them as potential influential explanatory variables from our model selection procedure. Their inclusion may lead to increased instability of fitted models, and they are unlikely to be influential on the resultant sentence construction by taking dominant values. We thus consider modelling the response sentence constructions taking four possible values {ADO, APD, PDO, PPD} in terms of subsets of the explanatory variables {SC, SemC, R.Ani, R.DA, R.Pro, P.DA, Cou}. We consider the maximum cardinality of subsets, k, to be 5, which allows for influential attributes and their interactions to be identified where possible.

The second set of analyses concerns six response levels for the active and passive voices of DO, PD and MT. As no recipient is stated in the MT constructions, we do not include the attributes featuring the recipient {R.Ani, R.Pro, R.DA} in our model exploration process. P.Ani and P.Pro have been excluded for the same reasons as in the four-level response case. We thus consider modelling the response sentence construction taking values {ADO, APD, AMT, PDO, PPD, PMT} from subsets of {SC, SemC, P.DA, Cou}. In this instance, we set the maximum cardinality to be 4, i.e. $k = d = 4$, meaning that we consider all possible models that can be constructed from all possible linguistic attributes.

All results were performed using the software R via the package `nnet` (Venables and Ripley, 2002) where a multinomial logistic function can be fitted via the function `multinom`.

4.1. Four-level response analysis (ADO, APD, PDO, PPD)

4.1.1. Pooled analysis

This set of results corresponds to analysis on the entire data set with variety Cou as a possible explanatory variable in the models constructed, in addition to the marked-up linguistic attributes.

Fig. 1(a) displays the uncertainty coefficient between each linguistic attribute and the response sentence construction. There are no attributes which have a particularly high degree of association with the response, although recipient pronominality R.Pro and syntactic complexity SC are more associated with the response. Although this has highlighted pairwise association between attributes and the response, this approach fails to highlight influential attributes when they are working in combination with each other. This thus motivates our particular model selection procedure.

Table 4 shows the AIC model selection results as explained in Section 3.4.1. We display the models achieving AIC-scores within the +2-range of the minimum (the 'best' model). Our results highlight that, under the current restriction of using at most five linguistic attributes in constructing models, all the sentence level features that are included in our framework are considered influential. This thus suggests that these attributes all have some part to play in explaining the sentence response construction, validating selection of these features for linguistic mark-up. Table 4 also highlights how some attributes interact with each other in explaining

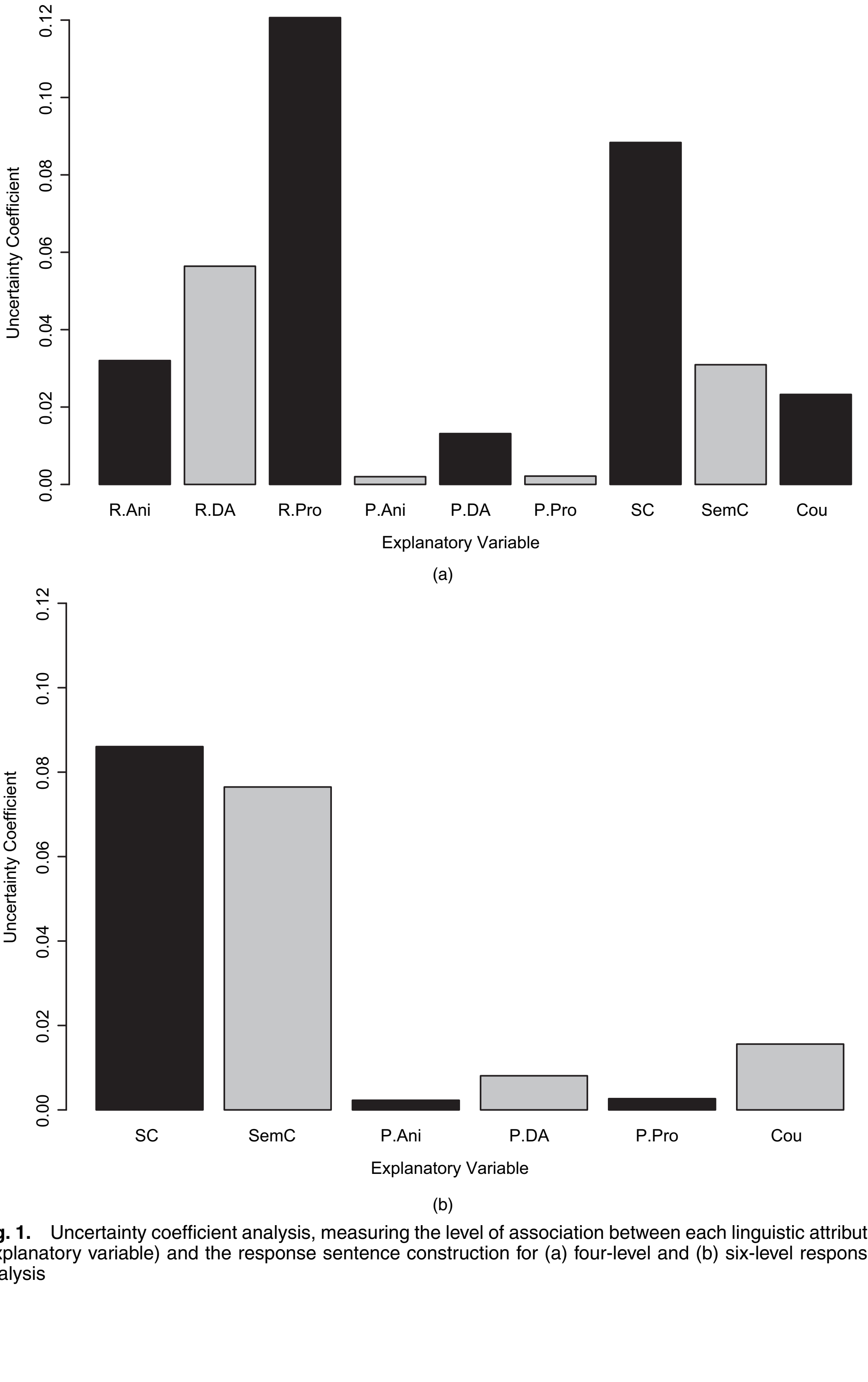

Fig. 1. Uncertainty coefficient analysis, measuring the level of association between each linguistic attribute (explanatory variable) and the response sentence construction for (a) four-level and (b) six-level response analysis

Table 4. Model selection results for four-level {ADO, APD, PDO, PPD} by using the set of explanatory variables {SC, SemC, R.Ani, R.DA, R.Pro, P.DA, Cou}†

Model	*AIC-score*
Cons ∼ SC∗R.Pro + SemC∗P.DA + R.Ani∗P.DA	929.71
Cons ∼ SC∗R.Pro + SemC∗P.DA + R.Ani∗P.DA + R.Pro∗P.DA	930.78

†We display the models achieving AIC-scores within the +2-range of the minimum. $A*B$ denotes all equivalent models containing the multiplicative term A:B.

the response (e.g. SC∗R.Pro and SemC∗P.DA), but not all two-way interactions are necessary in explaining the response sentence construction. As five attributes feature in all the top selected models, equal to the maximum cardinality set, this suggests that there are no smaller subsets of influential attributes under the current settings. Various orders of maximum cardinality could thus be further investigated in determining whether the set of influential attributes that are presented differs with respect to the cardinality, which may also provide information in ranking influential attributes.

The linguistic attribute corresponding to variety, Cou, has not been selected in these top models, which suggests that it is not considered as influential in explaining the resultant sentence construction under the current model selection condition, compared with the sentence level features. This suggests that variety is not an important explanatory variable of construction, but it leaves open the possibility that subsets of other explanatory variables that are influential may change under stratification by variety, which is a possibility that we explore below.

We consider the model Cons ∼ SC∗R.Pro + SemC∗P.DA + R.Ani∗P.DA for some further inference. We stress that this is by no means the best model that can possibly be used, with the goodness-of-fit test proposed in Goeman and Le Cessie (2006) being potentially applicable here.

Table 5 displays the information regarding the assumed fit, namely the estimated coefficient values, standard error and their significance. The majority of the coefficients appear to be well estimated and significant with reasonably low standard errors. Instances where the standard errors are undefined, namely when SC $= (R > P)$:R.Pro $= P$, represent when this particular configuration is not realized in the data. This is not particularly surprising as pronouns are short phrases and thus it is seldom that the recipient is a pronoun, but the recipient is more syntactically complex than the patient. This may be a natural representation of the English language.

Model inference generally concurs with linguistic behaviour and natural use of the English language. Active constructions appear to be more probable over passive constructions. The DO-construction also appears to become more probable when SC $= R \leqslant P$, and PD being more probable in the other case. As an example of the former scenario, consider ‘John gave her a very fascinating book about cats’ (ADO) *versus* ‘John gave a very fascinating book about cats to her’ (APD), where ADO appears more natural to use.

4.1.2. *Stratifying the data set by variety*

The current analysis considers variety Cou as an explanatory variable in modelling the resultant sentence construction. This is found not to be particularly influential as it has not been

Table 5. Estimated coefficients values and standard error for the model considered, Cons ∼ SC:R.Pro + SemC:P.DA + P.DA:R.Ani in the case of the four-level response analysis†

	PPD		*ADO*		*APD*	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Intercept	4.02†	0.32	6.65†	0.25	2.69†	0.26
SC = ($R \leqslant P$):R.Pro = NP	1.44†	0.34	2.28†	0.20	1.60†	0.26
SC = ($R > P$):R.Pro = NP	2.79†	0.37	1.15†	0.29	2.361†	0.30
SC = ($R \leqslant P$):R.Pro = P	−0.22	0.60	3.22†	0.24	−1.27†	0.57
SC = ($R > P$):R.Pro = P	0.00	—‡	0.00	—‡	0.00	—‡
SemC = Inf:P.DA = NG	−5.47†	0.65	−9.08†	0.60	−2.43†	0.56
SemC = Abs:P.DA = NG	−7.59†	0.43	−7.88†	0.32	−3.44†	0.29
SemC = Con:P.DA = NG	−6.59†	0.51	−8.55†	0.41	−2.90†	0.38
SemC = Inf:P.DA = G	24.56†	0.60	20.76†	0.60	−8.47†	0.00
SemC = Abs:P.DA = G	−11.11†	0.00	6.02†	0.55	10.03†	0.60
SemC = Con:P.DA = G	10.21†	0.54	5.37†	0.49	9.91†	0.54
P.DA = NG:R.Ani = A	0.28†	0.55	0.04	0.34	−0.88†	0.34
P.DA = D:R.Ani = A	−15.57†	0.72	−14.63†	0.59	−13.53†	0.60

†Significant coefficient at a 5% level.
‡Not applicable.

Table 6. Model selection results for the four-level response when we stratify the data set with respect to the variety and perform the model selection procedure on the three separate data sets, independently of each other†

Model	*AIC-score*
GB	
Cons∼SC + R.Ani∗R.DA + R.Ani∗P.DA	284.91
Cons∼SemC + SC∗R.DA + R.Ani∗R.DA	286.15
Cons∼SC + P.DA + R.Ani∗R.DA	286.87
Ind	
Cons∼SemC + R.Pro	312.28
Cons∼SemC + R.Ani + R.Pro	313.26
Cons∼SC + SemC + R.Pro	313.31
Pak	
Cons∼P.DA + SC∗R.DA + SC∗R.Pro	353.53
Cons∼P.DA + SC∗R.DA + R.DA∗R.Pro	353.71
Cons∼P.DA + SC∗R.Ani + SC∗R.Pro	354.62
Cons∼SC∗R.Ani + SC∗R.Pro + SemC∗R.Ani	354.94

†This may highlight the difference between varieties of English in terms of the selected influential explanatory variables and terms. $A*B$ denotes all equivalent models containing the multiplicative term A:B.

included in any of the top models in Table 4. We consider stratifying the complete data set with respect to variety and repeating our model selection procedure on each variety data set independently. This approach may highlight any influential explanatory variables which are specific to the variety. This would not be possible when variety is considered as an explanatory variable. The explanatory variables that are considered are {SC, SemC, R.Ani, R.DA, R.Pro, P.DA, SC,

Table 7. Model selection results for six-level English by using the set of explanatory variables {SC, SemC, P.DA, SC, SemC, Cou}†

Model	*AIC-score*
Cons∼SC + Cou + SemC∗P.DA	1754.76
Cons∼SC + SemC + P.DA + Cou	1755.05
Cons∼SC + SemC + Cou	1755.36

†We display the models achieving AIC-scores within the +2-range of the minimum. $A*B$ denotes all equivalent models containing the multiplicative term $A{:}B$.

SemC}, and we set the maximum cardinality of subsets considered as 4.

Table 6 displays the model selection results, stratifying with respect to variety. Some differences are observed in the selected influential terms; for example, R.Ani∗P.DA prominently features in many of the top selected models for British English, but rarely in the other two cases. In addition, multiplicative terms do not feature in the models for Indian English, which suggests that multiplicative interactions may not be as influential for this variety. The selection of different terms, and the way that they interact, thus suggests that there are some subtle differences between the varieties of English that could not be identified under the original pooled analysis and a stratified approach may be necessary.

However, we stress that this a tentative finding as the data leave open the possibility that the apparent differences between varieties are spurious.

4.2. Six-level response analysis (ADO, APD, AMT, PDO, PPD, PMT)

We do not discuss in as much detail the results for six-level response data. Fig. 1(b) displays the uncertainty coefficient between the new reduced set of explanatory variables that are considered and the response construction. Again, there are no particularly strong associations, although syntactic complexity SC and semantic class SemC have the strongest association with the response. Table 7 displays the model selection results, noting that all possible linguistic attributes have been highlighted as being influential. We choose Cons $\sim$ SC + Cou + SemC:P.DA as the model to perform some preliminary inference on. Table 8 displays information regarding the fitted model. The variety explanatory variable has been selected as an influential variable in explaining the response sentence construction. This tentatively suggests that there are some differences between the varieties of English that are considered in this particular analysis, although we note that this is under the least restrictive conditions of being able to construct a model from all available linguistic attributes.

We observe that the AMT-construction becomes more probable when semantic class is informative. This seems intuitive as 'The lecturer gave a presentation' (AMT) is completely natural to use in place of 'The lecturer gave the students a presentation' (ADO) or 'The lecturer gave a presentation to the student' (APD), with the recipient implicitly assumed from the informative context. This is not so when the object is the other possible alternatives: abstract (e.g. 'an advantage') or concrete ('a book').

5. Discussion and conclusions

This paper has presented an analysis, rooted in a model selection framework, where similarities

Table 8. Estimated coefficient values and standard error for the model considered, Cons ~ SC + Cou + SemC:P.DA in the case of the six-level response analysis

	PPD		*PMT*		*ADO*		*APD*		*AMT*	
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Intercept	−1.15†	0.42	−0.56	0.40	3.03†	0.26	−1.26†	0.29	2.49†	0.30
SC = (R>P)	1.65†	0.48	−15.32†	0.00	−1.50†	0.39	1.23†	0.33	−3.41†	1.05
Cou = Ind	0.21	0.57	0.30	0.52	−0.71†	0.33	0.30	0.37	0.04	0.38
Cou = Pak	0.44	0.55	0.00	0.56	−0.57	0.34	0.49	0.38	−0.32	0.41
SemC = Inf:P.DA = NG	0.56	0.62	1.36†	0.55	−2.11†	0.57	1.75†	0.52	0.35	0.48
SemC = Abs:P.DA = NG	−1.32†	0.38	−0.34	0.36	−1.21†	0.24	1.03†	0.25	−2.03†	0.27
SemC = Con:P.DA = NG	−0.35	0.45	−0.56	0.53	−1.75†	0.33	1.31†	0.33	−2.62†	0.42
SemC = Inf:P.DA = G	13.85†	0.62	14.66†	0.43	11.55†	0.41	−7.69†	0.00	12.02†	0.40
SemC = Abs:P.DA = G	−15.05†	0.00	−0.89	0.95	−1.40†	0.52	1.28†	0.56	−1.95†	0.58
SemC = Con:P.DA = G	1.16†	0.48	−14.80†	0.00	−2.06†	0.49	1.07†	0.48	−3.28†	0.72

†Significant coefficients at a 5% level.

and differences between varieties of written English were investigated by highlighting influential linguistic attributes driving a sentence construction. We can also determine which two-way multiplicative interaction terms are influential in explaining the sentence construction. This is not possible under existing marginal statistical approaches which consider only pairwise association.

We conclude that there are overall similarities between the varieties, although subtler differences can potentially be detected via stratification of the data. We conclude this from the exclusion of the variety explanatory variable in the final models presented for four-level analysis, suggesting that it is not salient in explaining the response in relation to the other linguistic attributes. However, if we perform our model selection procedure on the data set stratified with respect to variety, some acute differences are suggested by the influential attributes and terms chosen, although these cannot be said to be statistically significant. Under a six-level response analysis, the inclusion of the variety attribute suggests that there are some subtle differences with the AMT-construction being slightly more probable for the south Asian varieties that were considered. This also concurs with existing linguistic studies (Schilk, 2011). We also generally conclude that the models considered agree with natural usage of the English language.

We have also demonstrated that not all linguistic attributes that were recorded in the data may be needed in explaining and modelling the sentence construction. For example, attributes corresponding to patient, namely patient pronominality and animacy, may not be required as they are dominant in taking a single factor level over all response constructions. This paper thus highlights that model selection should be incorporated in future linguistic studies as not all linguistic attributes may be needed in modelling the response. Syntactic complexity also appears in all sets of analysis (four level or six level, pooled or stratified), perhaps highlighting how influential it is in explaining the response sentence construction.

As a single simple model has not been reached under our analysis, this may indicate that such a simple model and a simple structure in explaining the response sentence construction do not exist. Given the numerous candidate top models, and the appearance of almost all marked-up linguistic attributes, it is likely that they all explain something different about the response. Model stability approaches such as those proposed in Meinshausen and Bühlmann (2010) and Müller and Welsh (2010) may be applicable in identifying a stable structure and the stable explanatory variables influencing the response.

The statistical analysis approaches that are presented are, however, not without shortcomings. The inclusion of two-way interaction terms with respect to a small data set has led to some problems in the fitted models. There are several instances where certain attribute combinations are realized infrequently or not at all, which lead to poor estimates of coefficients with large standard errors associated with them (see Tables 5 and 8). This consequently leads to some instabilities in the estimated models and the conclusions that we can draw from them. However, such instances may also be a reflection of the language itself (consider the pronoun, syntactic complexity scenario as mentioned in Section 4.1.1). This also suggests that the solution may not be as simple as having larger data sets.

We have not discussed in detail how the maximum cardinality has been determined in our analysis. This is decided by the user and, ultimately, how parsimonious they desire the model and results to be. Smaller values of maximum cardinality will pick out the influential attributes at a more restrictive level. Indeed, this is so if we perform our model selection procedure on a maximum cardinality that is smaller than 5, with less complex models appearing as the top models. The current maximum cardinality used in this paper also suggests that the majority of the attributes used in the marking-up process do actively contribute in the pattern selection response.

There are several possible extensions with respect to this study and problem. The possible association between the linguistic attributes themselves could be investigated. This may shed additional light on why particular attributes and terms are chosen over others in our model selection framework. This extension could be implemented by using the uncertainty coefficient between all attributes as performed previously in our study. Secondly, Bresnan and Hay (2008), and many others, simply consider the similarities and differences by fitting a generalized linear model constructed from all available linguistic attributes and variety as explanatory variables, with no model selection being performed. Inference is based on analysing the coefficient that is associated with variety and its effect on the probability of sentence constructions. A more effective approach may be to exploit the fact that there is a possible hierarchical structure when considering varieties of English. It may be more informative to build this into the modelling procedure explicitly, and thus to highlight the similarities and differences between varieties of English.

Acknowledgements

The authors thank the referees, the Joint Editor and John Aston for constructive comments that improved the paper.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Bresnan, J., Cueni, A., Nikitina, T. and Baayen, R. (2007) Predicting the dative alternation. In *Cognitive Foundations of Interpretation* (eds G. Boume, I. Kraemer and J. Zwarts), pp. 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, J. and Ford, M. (2010) Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language*, **86**, 168–213.

Bresnan, J. and Hay, J. (2008) Gradient grammar: an effect of animacy on the syntax of give in New Zealand and American English. *Lingua*, **118**, 245–259.

Burnard, L. (2000) *The British National Corpus Users Reference Guide*. Oxford: Oxford University Computing Services. (Available from `http://www.natcorp.ox.ac.uk/corpus/`.)

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002) On Bayesian model and variable selection using MCMC. *Statist. Comput.*, **12**, 27–36.

Everitt, B. (1992) *The Analysis of Contingency Tables*. London: Chapman and Hall.

Goeman, J. and Le Cessie, S. (2006) A goodness-of-fit test for multinomial logistic regression. *Biometrics*, **62**, 980–985.

Görlach, M. (1991) *Englishes: Studies in Varieties of English, 1984-1988*. Amsterdam: Benjamin.

Hoffmann, S. (2007) From web-page to mega-corpus: the CNN transcripts. In *Corpus Linguistics and the Web* (eds M. Hundt, N. Nesselhauf and C. Biewer). Amsterdam: Rodopi.

Hosmer, D. and Lemeshow, S. (2000) *Applied Logistic Regression*. New York: Wiley-Interscience.

Johnson, K. (2008) *Quantitative Methods in Linguistics*. Hoboken: Wiley–Blackwell.

McArthur, T. and McArthur, R. (2005) *Concise Oxford Companion to the English Language*. New York: Oxford University Press.

Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.

Mukherjee, J. (2001) Principles of pattern selection. *J. Engl. Ling.*, **29**, 295–314.

Mukherjee, J. (2005) *English Ditransitive Verbs: Aspects of Theory, Description and a Usage-based Model*. Amsterdam: Rodopi.

Müller, S. and Welsh, A. H. (2010) On model selection curves. *Int. Statist. Rev.*, **78**, 240–256.

Schilk, M. (2011) *Structural Nativisation in Indian English Lexicogrammar*. Amsterdam: Benjamin.

Stefanowitsch, A. and Gries, S. (2003) Collostructions: investigating the interaction of words and constructions. *Int. J. Corp. Ling.*, **8**, 209–243.

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