Preposition Stranding in English: Predicting Speakers' Behaviour

Stefan Th. Gries
University of Southern Denmark at Sønderborg

1. Introduction

1.1 The phenomenon

In English PPs, the prepositions commonly precede their complements:

(1) He has paid [$_{PP}$ for the room].
(2) It is worth listening [$_{PP}$ to him].

There are cases, however, where this general word order preference is overridden in that the preposition is separated from its complement, In some instances, the choice of construction is optional:[1] either the preposition remains directly in front of its complement (i.e., the preposition is pied-piped; cf. the (a)-sentences) or it is stranded/deferred/orphaned after its complement has been moved away (the (b)-sentences; the examples are taken from Takami 1992:1):

(3) a. [$_{PP}$ To whom]$_i$ did John give the book t_i? (in VP
b. Who$_i$ did John give the book [$_{PP}$ to t_i]? or in S)
(4) a. [$_{PP}$ Of whom]$_i$ did you see a picture t_i? (in
b. Who$_i$ did you see a picture [$_{PP}$ of t_i] ? NP)

The (b)-sentences exhibit a phenomenon that has frequently been referred to as Preposition Stranding (henceforth PS).[2] From my point of view, there are three particularly interesting questions concerning PS:

1) When is it possible/grammatical to strand the preposition at all, and when is it not? This issue has been discussed in many studies. The approaches vary from purely syntactic ones (in which the argument-adjunct distinction, the notion of subjacency and the ECP have played a role; cf. Hornstein and Weinberg 1981; Chomsky 1981, 1986) over semantico-cognitive ones (Deane 1991, 1992; Kluender 1990) and discourse-functional ones (most notably Takami 1988, 1992) to psycholinguistic analyses (cf., e.g., Hawkins 1999 and the references cited therein).
2) Why does English offer the opportunity to strand prepositions at all? Given the following set of facts, it seems fairly strange that PS is possible and frequently found in English in the first place:
   - PS in interrogatives is prescriptively considered ungrammatical;
   - in general, English has a comparatively rigid word order allowing lit-

tle word order variation;

- filler-gap constructions are known for the processing load they impose on interlocutors compared to their pied-piped counterparts, which is why they are cross-linguistically quite rare: First, speakers need to process/produce the whole of the bridging structure while still having to produce the preposition. Second, hearers need to identify the gap to which the filler belongs (cf. Wanner and Maratsos 1978; Hawkins 1999): only after the final word of the sentence has been processed do they know that the sentence-initial NP is part of the PP (especially in the absence of overt case-marking). Moreover, hearers can sometimes choose one of several possible gap sites during online parsing: in *[$_{NP}$ Which student] did you ask t Mary about t?*, the hearer needs to relate the filler NP to one of possible gaps (indicated by the t's).

3) Which variables govern the choice of construction? More precisely, how important are these variables in determining the choice of construction? What is the reason for the distribution of constructions we find? On the basis of these variables, can we predict the constructional choices by native speakers of English?

It is question no 3 that I would like to focus on in this paper. But first it is necessary to introduce some terminology. In the remainder of this paper, the word order in the (a)-sentences is referred to as PPC (pied-piped construction) – the word order of the (b)-sentences is referred to as SC (stranded construction). Further, the utterance in which PS occurs is divided into several parts, as illustrated in (5) and (6).

(5) [$_{NP}$ Which posts]$_i$ | did you get | [$_{NP}$ an appointment [$_{PP}$ to t$_i$]]?
extracted phrase + head noun | bridging structure | extraction site

(6) [$_{NP}$ Which currency]$_i$ | would you prefer to trade | [$_{PP}$ in t$_i$]?
extracted phrase + head noun | bridging structure | extraction site

1.2 Hypotheses and Objectives

Various studies of word order alternations have shown that constructional choices are often influenced by the amount of processing that is necessary for the production of the utterance (cf. Gries 1999, 2000; Hawkins 1991, 1994, 1999; Arnold and Wasow 1996, 2000, to name but a few). While these theories share the idea that processing cost is an important determinant of constituent ordering, they also differ with respect to several parameters.

For instance, Hawkins' studies focus on the processing cost of the hearer by postulating that particular constituent orders make online phrase structure recognition more efficient. Arnold and Wasow (1996, 2000), by contrast, emphasise the speaker's perspective and, in Arnold and Wasow (2000), argue convincingly that it can be very difficult to decide on whose processing effort (the speaker's or the hearer's) is relevant as the empirical evidence supports both points of view. In

Gries (2000), I tend towards assigning higher priority to the speaker's perspective on production, which I will also do in the present work.

A second major difference is concerned with the determinants (or manifestations) of processing effort. While earlier studies by Hawkins have exclusively relied on morphosyntactic determinants of processing, Hawkins (1999) also embraces lexico-semantic variables. Arnold and Wasow (2000) include morphosyntactic variables (heaviness) as well as discourse-functional ones (newness). In this study, I suggest (as in Gries 2000) that the processing cost of utterances differing only in terms of their constituent orderings is determined by (or, at least, correlates with) an even larger variety of variables, namely phonological, morphosyntactic, semantic, discourse-functional and other variables (such as structural priming or speed of lexical retrieval).

Given the fact that filler-gap dependencies generally involve a large amount of processing cost, I propose that the choice of construction in the case of PS will also be sensitive to the processing cost incurred by the planning and production of the utterance. Since, the SC involves more processing cost I propose that the SC will be avoided in situations where its processing cost would add to an already high amount of processing effort. In such cases, the PPC would be chosen in order to minimise the overall processing effort. More succinctly, I propose that

- the PPC will be used in instances where the processing cost of the utterance is already high;
- the SC will be used in instances where the processing cost of the utterance is not too high.

Additionally, on a methodological level, I would also like to support my claim (cf. Gries 2000) that instances of syntactic variation are best analysed

(i) on the basis of naturally-occurring corpus data and

(ii) by using multifactorial statistics such as the General Linear Model (GLM), Linear Discriminant Analysis (LDA) and Classification and Regression Trees (CART).

As a basis for my analysis, I used a concordance program to search the British National Corpus (BNC) for instances of the two constructions; the following set of data was obtained:

	Written	Spoken	Row totals
PPC	122 (49.39%)	0 (0%)	122 (40.53%)
SC	125 (50.61%)	54 (100%)	179 (59.47%)
Column totals	247 (100%)	54 (100%)	301 (100%)

Table 1: Analysed Data from the BNC (Raw Frequencies + Column Percentages)

2. Previous Analyses

Previous analyses have shown that different groups of variables are relevant to whether PS is possible or not and the choice of construction; consider Table 2.

Value for PPC	Variable	Value for SC
dominant	dominance of extracted phrase (Erteschik-Shir and Lappin 1979)	
high	attention attraction of extracted phrase (Deane 1992)	
high	topicality of extracted phrase (Kuno 1987)	
high	semantic barrierhood[3] of the extracted phrase (Kluender 1990)	low
high	entrenchment of the extracted phrase (Deane 1992)	
low	semantic barrierhood of the bridging structure (Kluender 1990)	high
short	syllabic length of the bridging structure (Quirk et al. 1985)	long
high	relation between light verb and extraction site (Deane 1992)	
low	attention attraction of the bridging structure (Deane 1992)	
VP-final	position of extraction site (Deane 1992)	
newer/more important than rest of S	cognitive status of extraction site (Takami 1992)	
high	attention attraction of extraction site (Deane 1992)	
low	entrenchment of the extraction site (Deane 1992)	
attribute or characteristic part	referent/denotatum of extraction site (Bolinger 1972)	
indefinite	definiteness of the extraction site (Deane 1992)	
	semantic case role of the extraction site (Deane 1992)	agent / subject
non-specific	specificity of the extraction site (Deane 1992)	
formal	formality of register (Quirk et al. 1985)	low / neutral
complex	syll. length of preposition (Quirk et al. 1985)	short
	frequency of preposition (Quirk et al. 1985)	frequent
temporal/abstract	meaning of preposition(al phrase) (Quirk et al. 1985)[4]	spatial, instrum., reason
passive	voice of the verb	active
strong	relation between preposition and its complement (Quirk et al. 1985)	loose
loose	relation between preposition and its verb (Quirk et al. 1985, Biber et al. 1999)	strong/close (prep. verbs)[5]

Table 2: Variables That Are Argued to Govern PS

The following comments on this inventory of variables are called for: First, the

analyses are commonly only based on intuitive and introspective examples and acceptability judgements: sometimes this is explicitly mentioned (cf. Takami 1992:5f.) – sometimes we are simply intended to follow the author's claims (cf., e.g., Deane 1992). Correspondingly, naturally-occurring data have hardly ever been used to validate prior analyses.

Second, most variables were investigated in isolation only so (i) no weightings of variables are offered, i.e. we cannot assess/quantify the degree of importance of any particular variable, and (ii) no interactions of variables can be considered.

Finally, let us turn to what are generally claimed to be the objectives of scientific research, namely description, explanation and prediction. As to description, no satisfactory data-based description has been offered so far. As regards explanation, with few exceptions (most notably Deane 1992, Hawkins 2000, Takami 1992), no explanatory account incorporating several analyses has so far been proposed. Finally, the prediction of native speakers' constructional choices has never been attempted although it is plausible to assume that prediction would be the most rigorous way of putting one's own analysis or that of others to the test.

3. Results (for Selected Variables Only)

So far, not all of the above variables have been investigated: the results still must be taken with a grain of salt. The following is a list of variables (and possible levels) entering into the analysis; the dependent nominal variable is of course the choice of construction (where PPC and SC are coded as 0 and 1 respectively).

- MODALITY: *spoken, written*;
- VERB: *transitive, intransitive, prepositional, copula, phrasal-prepositional*;
- VOICE: *active, passive*;
- PREP_SEM: prepositional semantics: *abstract, metaphorical, spatial, temporal*;
- AGENT_HEAD: *agent, non-agent*;
- CONCRETE_HEAD: *abstract, concrete*;
- FREQ_HEAD: *infrequent, frequent*;
- ENTRENCH_HEAD: entrenchment of the head noun according to Deane's (1992) entrenchment hierarchy;
- FREQ-PREP: frequency rank of the preposition (in each modality);
- LENGTH_BS: syllabic length of the bridging structure;
- LENGTH_PREP: syllabic length of the preposition;
- BARRIER_BS: barrierhood of the bridging structure;
- LENGTH_EP: syllabic length of the extracted phrase;
- BARRIER_EP: barrierhood of the extracted phrase.

3.1 Monofactorial Results

As a first and simple step, one can start by (i) calculating means of the ordinal/interval variables and (ii) crosstabulating the nominal variables for both con-

structions. For instance, the means (and standard deviations) of Length_BS of the PPC and the SC are 13.3 (8.7) and 4.5 (2.3) respectively. This difference is highly significant (t_{Welch}=10.95; df=133; $p_{2\text{-tailed}}$<0.001 ***), showing that longer bridging structures result in a preference for PPC whereas shorter bridging structures are more likely to license SC; this result can be summarised using a simple coefficient of correlation (r_{pb}=-0.6; t=-12.92; p<0.001 ***). Analogous calculations can be done for all measurement variables. Consider, e.g., Table 3.

	Transitive	Intransitive	Prep.	Phrasal-prep.	Copula	Totals
PPC	73	24	4	0	21	122
SC	38	65	14	6	56	179
Totals	111	89	18	6	77	301

Table 3: Distribution of Constructions Relative to VERB

For such a table, a Chi-square value and a corresponding coefficient of correlation can be computed in order to determine whether VERB contributes to the choice of construction. In this case, the results also deviate highly significantly from the (according to H_0) expected results (χ^2=48.33; df=4; p<0.001 ***).[6] In order, however, to avoid going through all individual results at such a tiring level of specificity, the following table summarises the results for all variables investigated (sorted according to strength of impact of the variables).

Variable	Correlational Strength with PS
LENGTH_BS	r_{pb}=-0.6; p<0.001 ***
BARRIER_BS	r_{pb}=-0.594; p<0.001 ***
VERB	ϕ=0.4; p<0.001 ***
MODALITY (written=0; spoken=1)	ϕ=0.386; p<0.001 ***
VOICE (act.=0; pass.=1)	ϕ=-0.28; p<0.001 *
LENGTH-PREP	r_{pb}=0.246; p<0.001 ***
ENTRENCH_HEAD	τ=0.14; p<0.001 ***
CONCRETE_HEAD (abstract=0; concrete=1)	ϕ=0.14; p<0.016 *
BARRIER_EP	r_{pb}=0.13; p=0.029 *
AGENT_HEAD (no agent=0; agent=1)	ϕ=0.115; p=0.054 ns
PREP_SEM	ϕ=-0.1103; p=0.301 ns
FREQ_HEAD (rare=0; frequent=1)	ϕ=-0.096; p=0.107 ns
FREQ-PREP	τ=0.035; p=0.362 ns
LENGTH_EP	r_{pb}=-0.003; p=0.959 ns

Table 4: Monofactorial Results

Less technically, in the monofactorial analysis the bridging structure seems to be the most important determinant of the constructional choice. Given the high correlation between LENGTH_BS and BARRIER_BS (r=0.92; p<0.001 ***), the closeness of the morphosyntactic length and the semantic barrierhood is little surpris-

ing. Equally obvious is that the preposition does not seem to too relevant to the constructional choice contrary to what was suggested by some authors.[7] On the whole, the following overall ranking of variables is found: bridging structure – verb – head noun – preposition.

3.2 The Problem of Interactions

While the preceding investigation goes beyond many previous studies (by precisely measuring the importance of the variables for the first time), it is still far from complete. Knowing monofactorial preferences for constructions does not necessarily enable us to predict speakers' choices since in many (if not most) discourse situations, we will find conflicting preferences of variables. For instance, we know that transitive verbs prefer PPC while concrete head nouns prefer SC. How do speakers, then, decide in the cases given in (7) (transitive verb + a concrete head noun) and (8) (intransitive verb and abstract head noun)?

(7) a. Which half do you want the marmalade on?
b. On which half do you want the marmalade?

(8) a. Which sport, apart from rowing, could you do that in?
b. In which sport, apart from rowing, could you do that?

This is a difficult question, since

1) in monofactorial analyses, interactions of variables cannot be identified;
2) for purely mathematical reasons, the absolute values of the correlation coefficients must not be compared directly.

Thus, two possible strategies are proposed: one can resort to truly multifactorial procedures (cf. section 3.3) or one can use multidimensional crosstabulation to determine the frequencies of the two constructions in all cases of conflicting variable values/levels. For instance, multidimensional crosstabulation shows that of all 301 cases, there are 30 cases like (7) (i.e. where VERB: *transitive* contrasts with CONCRETE_HEAD: *concrete*), of which 19 exhibit PPC and 11 exhibit SC (this distribution is not significant: $p_{\text{binomial test}} \approx 0.1$). In other words, in a direct comparison, VERB: *transitive* wins out in getting its constructional preference recognised, but fails to do so significantly.[8] This can be done for all contrasting pairs in order to determine a ranking of variable strengths. Since this (i) is quite a laborious task and (ii) still does not enable us to predict speakers' choices, however, an analysis using multifactorial techniques is probably more rewarding.

3.3 Multifactorial Results

One might wonder how many variance one's present state of the art can account for and, at the same time, how the variables' influence is altered once they are all considered simultaneously (the only cognitively realistic avenue of research). 'The General Linear Model (GLM) answers exactly these questions. The multiple correlation coefficient (with correction for shrinkage according to Wherry) for all above variables without interactions is quite high and highly significant: $R_c=0.635$; $F_{18, 273}=17.01$; $p<0.0001$ ***).[9]

More interesting for our present purposes, however, is to try to predict speakers' choices. A linear discriminant analysis (LDA) takes as input a set of independent variables and produces as output a categorial choice of the level of the dependent variable (STRUCTURE). Using cross-validation, *a priori* predictions of speakers' choices in one's analysis can be tested for accuracy while, at the same time, the analysis as a whole can be subjected to the most rigorous test conceivable, namely whether it enables the researcher to actually predict what native speakers do. The results of the LDA for our data set can be summarised as follows.

The set of variables entering into the analysis discriminates highly significantly between the two constructions (canonical R=0.746; χ^2=219.48; df=19; p<0.001 ***). More interestingly, the constructional choices can be classified correctly (*post hoc*) in 89.7% of all cases. The most essential result, however, is that the *a priori* prediction accuracy (as determined by the so-called leave-one-out method) is 86.1%, i.e. 86.1% of the constructional of native speakers in actual discourse choices can be predicted correctly.[10] What is more, the predictions are arrived at by assigning to each variable a numerical weighting/loading, which can be interpreted as reflecting the importance of a variable in discriminating between PPC and SC. Table 5 provides the weightings resulting from the present analysis.

Variable	Factor Loading	Choice of Construction
barrierhood of the bridging structure	-0.701	high values for these variables ⇒ PPC low values for these variables ⇒ SC
length of the bridging structure	-0.69	
transitive verbs	-0.426	
voice of the verb	-0.258	
temporal meaning of the preposition	-0.089	according to the low factor loadings (-0.223 ≤ loading ≤ 0.223),[11] these variables do not discriminate significantly between the two constructions
frequency of the head noun	-0.087	
metaphorical of the preposition	-0.009	
abstract meaning of the preposition	0.014	
length of the extracted phrase	0.036	
spatial meaning of the preposition	0.04	
agentivity of the head noun	0.104	
phrasal-prepositional verbs	0.114	
frequency of the preposition	0.115	
barrierhood of the extracted phrase	0.119	
prepositional verbs	0.126	
concreteness of the head noun	0.132	
copula as verb	0.153	
entrenchment of the head noun	0.165	
intransitive verbs	0.165	
length of the preposition	0.218	
modality	0.382	high/low value ⇒ SC/PPC

Table 5: Factor Loadings of the Discriminant Analysis

It is obvious that, of all variables investigated, the bridging structure, the verb and the modality influence PS most strongly. The hypothesis of the influence of processing effort on the choice of construction seems to be borne out since the length and the barrierhood of the bridging structure relate straightforwardly (along the lines discussed in section 1.1) to the morphosyntactic and semantic processing effort respectively necessary for the production of the utterance.

As to the influence of transitive verbs on PS, one might wonder whether this finding supports the role of processing put forth, but there is an obvious explanation for that, too: as opposed to all other kinds of verbs investigated here, transitive verbs require a direct object, i.e. at least an additional NP. This NP will obligatorily add to the length and the barrierhood of the bridging structure as in, say, *To whom did John give [$_{NP}$ the book]?* or *Who did John give [$_{NP}$ the book] to?* and thereby yield a preference for the PPC. A look at our data supports this hypothesis; consider Table 6.

	Transitive (111 sentences)	Not transitive (190 sentences)	Total
LENGTH_BS: Mean (Std. dev.)	10.9 (7.7)	6.5 (6.4)	8.1 (7.2)
BARRIER_BS: Mean (Std. dev.)	4 (2.9)	2.5 (2.7)	3 (2.9)

Table 6: The Effect of Transitivity on LENGTH_BS and BARRIER_BS

The average length and barrierhood of the bridging structure is much higher for transitive verbs than for non-transitive verbs; the differences are, according to Welch's *t* test, highly significant and the influence of transitive verbs can, thus, be explained in terms of processing effort.

The effect of verb voice on PS is more difficult to relate to processing cost: when the main verb is in the passive, we find SC significantly less than expected. At this preliminary stage, I can only suggest somewhat tentatively that the non-canonical passive is more difficult to process than the canonical active so that both passive and SC is avoided by speakers. Admittedly, compared to the other more solid arguments, this is fairly vague and requires further investigation.

The strong influence of the modality, however, is most probably not due to a causal influence on processing – rather, it is more likely due to writers' prescriptive knowledge/awareness (never use a preposition to end a sentence with!).

4. Summary / Conclusions

We have seen how the analysis of syntactic variation can benefit from the use of rigorous corpus-based and (multifactorial) statistical investigation. While such techniques to analysing variation data were quite common in the 70s (cf. the notion of variable rules employed by Cedergren, Labov, Sankoff and others), nowadays the analysis of variation does not (at least to my mind) utilise the power of these techniques frequently enough. This is all the more surprising since even introductory textbooks (!) to corpus linguistics as well as other publications

have argued time and again that monofactorial studies often do not suffice:

> […] straightforward significance or association tests, although important, cannot always handle the full complexity of the data. The multivariate approaches […] offer a way of looking at large numbers of interrelated variables and discovering or confirming broader patterns within those variables. (McEnery and Wilson 1997:82)

> Although linguists … typically do not use statistical techniques, the approach just illustrated fits conceptually with correlational models using multiple regression analyses … [i.e.,] with a more complex design we can obtain information that is not readily available by armchair analysis. (Bates and McWhinney 1982:181)

In this respect, I would thus argue that, methodologically at least, there is a great deal that we as linguists can learn from other behavioural sciences as far as data collection, hypothesis testing and exploratory statistical techniques are concerned. I would also hope that a shift to more rigorous testing of the sort detailed above would render linguistic findings more objective and reliable than has been the case in the preceding 40 years of predominantly intuitive/introspective analyses of acceptability/grammaticality judgements (cf. Schütze 1996 for a similar line of reasoning, though not in the direction of multifactorial corpus analyses).

In the case at hand, the most crucial determinants of PS seem to be the processing effort associated with the two word orders and the knowledge of prescriptive grammar rules. On a more general note, the findings concerning processing effort lend themselves to being integrated into psycholinguistic theories based on interactive activation networks such as Bates and MacWhinney's (1982, 1989) Competition Model, where variables with different constructional preferences compete with each other: the notion of interaction as dealt with in section 3.2 operationalises the notion of conflict validity, the prior probabilities of the two constructions in the LDA/CART analyses correspond to resting levels / baseline activations, and the variables' weightings could readily be interpreted as association strengths between variables and the constructional choice. However, further research is necessary to integrate more of the previous findings into psycholinguistic theory.

5. Notes

[1] Here and in the rest of the paper, the expressions *choice of construction* or *speakers' decisions* are not to be understood as implying that there is always a conscious choice on the part of the speaker.

[2] In the psycholinguistic literature, PS is just one instance of what is frequently referred to as filler-gap dependencies. However, this paper is only concerned with PS in interrogatives; I will leave aside instances of pseudo-passives (such as *The problem had been accounted for.*), *Tough*-Movement (such as *Last night was difficult to sleep through*) and relative clauses (*They ate what they had paid for*).

[3] Barrierhood is an index accounting for open/closed-class words and frequency.

[4] Biber et al. (1999:106) provide a list of prepositions that can usually be stranded (*about, after, at, by, for, from, in, like, of, on, to, with*) while some others are only rarely attested (*against, around, into, near, off, through, under, up*). However, on the whole, Quirk et al's (1985) generalisation seems to hold as many of these prepositions are indeed used to denote spatial configurations or to introduce an instrument. Note also that there are some prepositions that are hardly ever deferred: *since*, *during*, *until* (Quirk et al. 1985:817).

[5] Unfortunately, the identification of intransitive prepositional verb is far from straightforward. So far, no clear-cut tests have been devised to distinguish intransitive prepositional verbs (as in *John asked for some details*) from verbs that are simply followed by a PP (*John left before noon*). One test that has been proposed (cf. Collins Cobuild on CD-ROM) is that only prepositional verbs license the SC, but of course this test could not be used here since it is not independent of the focus of the present paper. For traditional treatments of this question, cf. Quirk et al. (1985:1165ff.) and Biber et al. (1999:406, 414). The from my point of view most promising approach is illustrated in Hawkins (2000:241ff.).

[6] Note however, that the overall significant deviation mainly results from the effects found for transitive verbs as can be inferred from the individual cells' contributions to Chi-square.

[7] LENGTH_PREP has resulted in a significant effect, but the actual difference is so small as to be meaningless (mean LENGTH_PREP for PPC: 1 syllable; mean LENGTH_PREP for SC: 1.2 syllables).

[8] This strategy is very similar to the operational definition of the notion of conflict validity as proposed by Bates and MacWhinney (1989).

[9] With interactions the model results in a multiple correlation coefficient larger than 1 (not defined), so problems of multicollinearity still need to be addressed.

[10] There are researchers who might object to the application of an LDA to my data since the data do not meet the requirement of a multivariate normal distribution, which is why distribution-free techniques such as CART should have been used. However, while many researchers tend to emphasise the importance of distributional assumptions, there is also a number of scholars who argue that, in practice, these assumptions are not as essential as they might seem on a purely mathematical basis (cf. Winer et al. 1991:5). Second, it has even been claimed that there is no test that reliably identifies multivariate normal distributions (cf. Bortz 1999:435). Lastly, CART and LDA differ in that the former includes all variables in a sequential fashion whereas the latter does so simultaneously (and, thus, more cognitively realistically). Nevertheless, it might very well be the case that these reasons do not satisfy truly mathematically-oriented researchers. I have, therefore, also analysed my data using the CART module of Statistica 5.5; the algorithms used therein are based on CART by Breiman et al. (1984). The results are very similar: the classification accuracy obtained is 90.4%, the prediction accuracy for a small part of the corpus data is 87.5%, and the six most important variables are BARRIER_BS, LENGTHBS, FREQ_PREP, MODALITY, LENGTH_EP and VOICE. Thus, even a distribution-free technique does not invalidate the result of the LDA.

[11] The question may arise as to what is the motivation for the cut-off point of ±0.223. Basically, the choice of a cut-off point is in general an arbitrary one – I have chosen ±0.223 because this rules out factor loadings contributing less than 5% to the variance ($0.223^2 \approx 0.05$).

6. References

Arnold, Jennifer E., and Thomas Wasow. 1996. *Production Constraints on Particle Movement and Dative Alternation*. Poster presented at the CUNY Conference on Human Sentence Processing.

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. 2000. "Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering", *Language*, 76:28-55.

Bates, Elizabeth, and Brian MacWhinney. 1982. "Functionalist Approaches to Grammar", in: Wanner Eric, and Lila R. Gleitman (eds.). *Language Acquisition: The State of the Art*. Cambridge: Cambridge University Press, p. 173-218.

Bates, Elizabeth, and Brian MacWhinney. 1989. "Functionalism and the Competition Model", in: MacWhinney, Brian, and Elizabeth Bates (eds.). *The Crosslinguistic Study of Sentence Processing*. Cambridge: Cambridge University Press, p. 1-73.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow, Essex: Pearson Education.

Bolinger, Dwight D. 1972. “What Did John Keep the Car that Was in?”, *Linguistic Inquiry* 3:109-114.

Bortz, Jürgen. 1999. *Statistik für Sozialwissenschaftler*. 5th ed. Berlin, Heidelberg, New York: Springer.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software.

COBUILD on CD-ROM. 1994. HarperCollins Publishers Ltd.

Deane, Paul D. 1988. “Which NPs Are there Unusual Possibilities for Extraction From?”, in: Macleod, Lynn et al. (eds.). *Proceedings of the Twenty-fourth Annual Meeting of the Chicago Linguistics Society*. Chicago: Chicago Linguistics Society, p. 100-111.

Deane, Paul D. 1991. “Limits to Attention: A Cognitive Theory of Island Phenomena”, *Cognitive Linguistics* 2:1-63.

Deane, Paul D. 1992. *Grammar in Mind and Brain*. Berlin, New York: Mouton de Gruyter.

Erteschik-Shir, Nomi, and Shalom Lappin. 1979. “Dominance and the Functional Orientation of Island Phenomena”, *Theoretical Linguistics* 6: 41-86.

Gries, Stefan Th. 2000. *Towards Multifactorial Analyses of Syntactic Variation: The Case of Particle Placement*. PhD Dissertation, University of Hamburg.

Hawkins, John A. 1999. “Processing Complexity and Filler-Gap Dependencies across Grammars”, *Language* 75:244-285.

Hawkins, John A. 2000. “The Relative Ordering of Prepositional Phrases in English: Going Beyond Manner-Place-Time”, *Language Variation and Change* 11:231-266.

Hornstein, N., and Amy Weinberg. 1981. “Case Theory and Preposition Stranding”, *Linguistic Inquiry* 12:55-92.

Kluender, Robert. 1990. “A Neurophysiological Investigation of *Wh*-Islands”, in: Hall, Kira et al. (eds.). *Proceedings of the Sixteenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society. p. 187-204.

Kuno, Susumo. 1987. *Functional Syntax: Anaphora, Discourse, and Empathy*. Chicago: University of Chicago Press.

McEnery, Tony, and Andrew Wilson. 1997. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Radford, Andrew. 1997. *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.

Ross, John Robert. 1967. *Constraints on Variables in Syntax*. PhD Dissertation, M.I.T, Cambridge, MA.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*. Chicago: University of Chicago Press.

Takami, Ken-Ichi. 1988. “Preposition Stranding: Arguments against Syntactic Analyses and an Alternative Functional Explanation”, *Lingua* 76: 299-335.

Takami, Ken-Ichi. 1992. *Preposition Stranding: From Syntactic to Functional Analyses*. Berlin, New York: de Gruyter.

Winer, B.J., Donald R. Brown, and Kenneth M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. New York: McGraw-Hill.