A multifactorial corpus analysis of adjective order in English*

Stefanie Wulff
University of Bremen

This paper is concerned with the question of which factors govern prenominal adjective order (AO) in English. In particular, the analysis aims to overcome shortfalls of previous analyses by, firstly, adopting a multifactorial approach integrating all variables postulated in the literature, thereby doing justice to the well-established fact that cognitive and psychological processes are multivariate and complex. Secondly, the phenomenon is investigated on the basis of a large corpus, rendering the results obtained more representative and valid of naturally occurring language than those of previous studies. To this end, corpus-linguistic operationalizations of phonological, syntactic, semantic and pragmatic determinants of AO are devised and entered into a Linear Discriminant Analysis, which determines the relative influence of all variables (semantic variables being most important) and yields a classification accuracy of 78%. Moreover, by means of the operationalizations developed in this analysis, the ordering of yet unanalyzed adjective strings can be predicted with about equal accuracy (73.5%).

Keywords: adjective order, Linear Discriminant Analysis, multifactorial analysis

1. Introduction

This paper reassesses the question which factors govern the preference in the ordering of English prenominal adjectives. For example, (1a) is widely agreed to be preferable to (1b):

(1) a. *the big red ball*
    b. *the red big ball*

International Journal of Corpus Linguistics 8:2 (2003), **245–282.**
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

Although the number of analyses dealing with Adjective Order (henceforth AO) is substantial, most of them suffer from two methodological shortcomings undermining the validity of the results. Firstly, previous approaches mostly focused on a single variable only, i.e., it has never been argued in favor of a multifactorial approach comprising the various variables postulated so far. Particularly with respect to the fact that most analyses positively confirmed the stated influence of the different variables, the question arises how these variables actually work upon AO in combination. Do all of them have an effect of equal strength? Do their effects add up, or do some of them weaken or even annul the influence of others? Such questions, while often neglected in linguistic research proper, can be investigated within the field of corpus linguistics, where the explanatory improvement yielded by multifactorial analyses has been recognized, as the following quotation underscores.

> "[...] 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 & Wilson 1997: 82)

In this connection, a multifactorial analysis of AO is complicated in two ways. On the one hand, the variables proposed spread across the whole range of linguistic sub-branches: there are phonological, syntactic, semantic, as well as pragmatic variables. Some of these variables can be straightforwardly operationalized corpus-linguistically (consider, e.g., frequency), others do not lend themselves to such an operationalization easily: how, for instance, should one extract information about a word's 'nouniness' from a corpus? On the other hand, many of the variables have not been put forward in linguistics but in neighboring disciplines such as psycholinguistics and/or psychology, in which experimental designs are the predominant methodological device, from which corpus-linguistic operationalizations are, in many cases, difficult to derive. For example, it is difficult to see how subjects' responses to introspective variables such as the semantic closeness of particular adjectives and their corresponding head nouns could be operationalized corpus-linguistically.

The second methodological shortcoming of the majority of previous analyses is that data samples were often severely restricted so as to accommodate the researcher's goals, rendering these analyses unrepresentative of naturally-occurring language. Accordingly, the central aim of the present study is a methodological one, i.e. to carry out the first multifactorial and completely corpus-based analysis of AO. The focus will be on the following two questions:

– How can the variables be operationalized for a corpus analysis, in particular those which represent information that goes beyond word lengths, frequency values or the immediate context of the test item itself?
– What does a statistical analysis which takes all variables simultaneously into account reveal about the relative influence of these variables?

In the following section, the scope of the investigation will be specified. In Section 3, the generation of the data sample will be described. Section 4 will be devoted to a presentation of the variables postulated so far and how they were operationalized for my corpus analysis. Moreover, I decided also to test each variable individually to see whether the experimental results on AO in the previous literature are supported when investigated from a corpus-linguistic viewpoint; the results will be presented directly after the presentation of the corresponding variables. In Section 5, the results of a multifactorial statistical analysis will be presented and discussed; Section 6 concludes.

2. Scope of the investigation

2.1 Adjective order: what counts as an adjective?

Next to adjectives in attributive position such as the ones in (1), there is a variety of other pre-head modifiers in English for which a particular ordering may be established, yet they are beyond the scope of the present analysis. Some instances of these modifiers are exemplified in (2).

(2)	a.	*many flowers*	(logical qualifiers)
	b.	*these students*	(determiners)
	c.	*John's bicycle*	(possessive pronouns)
	d.	*the cleverest mouse*	(superlatives)
	e.	*the fourth grade*	(ordinal numbers)
	f.	*seven dwarfs*	(cardinal numbers)

There are semantic as well as syntactic criteria to distinguish between the two kinds of pre-head modifiers exemplified in (1) and (2). At the level of semantics, the pre-head modifiers in (2) fall into the class of limiting adjectives, whereas in the present study, only descriptive adjectives are taken into consideration (cf. Bloomfield 1933: 202; cf. also Teyssier's 1968: 227ff. analogous distinction between identifying and characterizing adjectives). Analogous to this semantic distinction, the two general adjective classes may also be differentiated on syntactic grounds: whereas limiting adjectives evoke a right-branching structure (i.e. all material standing on the right of them is modified as one entity), descriptive adjectives evoke a multi-branching structure (i.e., the adjectives independently modify the head noun; cf. Chomsky 1965: 165). Accordingly, various classification schemes of pre-head modifiers have been proposed, one of the most well-known is developed in Dixon (1977), which is also adopted here. He distinguishes three kinds of pre-head elements which can occur in a noun phrase: (i) pre-adjectival modifiers (such as qualifiers, determiners, etc.), (ii) adjectives as in (1), and (iii) post-adjectival modifiers (e.g., modifiers denoting the origin/composition or the purpose/beneficiary of the head noun).

There is a disagreement not only with respect to the number of classes necessary to describe all kinds of adjectives and their proper assignment to one of these classes; the crucial question indeed is: what exactly defines the word class adjective? To identify necessary and sufficient criteria which keep apart adjectives from other word classes is much more complicated than it might intuitively appear. Quirk et al. (1985) approach the problem by defining the class of adjectives not as an all-or-nothing class, but rather as a class with fuzzy boundaries, including central as well as peripheral adjectives. They propose four criteria for the identification of adjectives: (i) attributive use, (ii) predicative use after the copula *seem*, (iii) premodification by *very*, and (iv) gradability (in decreasing order of their significance for the definition of the class of adjectives; cf. Quirk et al. 1985: 402–404). Not all adjectives necessarily share all four criteria; words which do not meet any of the above-mentioned characteristics are excluded from the class of adjectives (e.g. *abroad*). As to the present analysis, I included only those adjectives into the data sample which satisfy Quirk et al.'s definition and are tagged as an adjective in the British National Corpus (BNC).[1]

2.2 Normal vs. contextually constrained adjective order

In several studies, it has been argued that AO is solely "determined by the pragmatic demands of the communication situation" (Danks & Glucksberg 1971: 66; cf. also Danks & Schwenk 1972, 1974), i.e. by a pragmatic communication rule which states that adjectives are primarily ordered according to the adjectives' discriminative potential: more discriminating adjectives precede less discriminating ones. However, Martin and Ferb (1973; cf. also Richards 1975) object to Danks and Glucksberg's view that the semantic and psychological factors influencing AO only constitute the high-frequency case of a purely pragmatic principle, arguing for the existence of two fundamentally different kinds of AO, normal AO (as investigated here) and contextually/pragmatically constrained AO, along the following lines: (i) whereas in normal preferred order the syntax is multibranching, in contextually constrained preferred order it is right-branching (cf. above); (ii) juncture between the adjectives is observable only with contextually constrained preferred order (cf. also Martin 1970: 382); (iii) in contextually constrained preferred AO, stress is on the first adjective, whereas in normal preferred order, stress remains constant or increases slightly from the first to the last adjective in a string (cf. also Teyssier 1968: 236); (iv) in contextually constrained preferred order, sequencing is "constrained by the contextually determined order of the sub-classification of the denotation of the noun" (Martin & Ferb 1973: 75), in normal preferred order, other factors like semantic closeness etc. influence the arrangement; (v) there is evidence from newly acquired words against the frequency-relation of normal and contextually constrained AO:

> "[I]f a person learns a new hue term, like *heliotrope*, then the new term is ordered appropriately immediately. In this case it cannot be argued that the speaker is responding to a simple index of frequency relating to the particular adjective. Clearly, it is at least necessary for Danks and Glucksberg to hold that the frequency index be associated with adjective classes rather than with individual adjectives." (Martin & Ferb 1973: 79)

2.3 Triples only

The present study takes adjective$_1$–adjective$_2$–noun combinations into consideration. Noun phrases containing more than two prenominal adjectives are relatively rare anyway: in the whole 10m-words spoken subcorpus of the BNC, 9,647 adjective pairs could be found, but only 426 of them (4.41%) are followed by a third adjective. On the other hand, 6,560 (68%) adjective pairs are immediately followed by a noun. So it is well justified to argue that a representative picture is already provided if only noun phrases containing two adjectives are taken into consideration.

2.4 Unbroken vs. broken adjective pairs

Finally, the adjective sequences investigated are all unbroken strings. There are indications (cf. Vendler 1968; Richards 1977) that the presence or absence of a conjunction influences the restrictiveness of the ordering preferences in prenominal AO: ordering constraints are apparently more restrictive for unbroken than for broken strings. However, the nature of this influence of conjunctions has not been clarified precisely yet and, for reasons of space, this interesting question can not be addressed here.

3. The data sample

All adjective$_1$–adjective$_2$–noun triples of the spoken part of the BNC which met the following conditions were included in the data sample:[2]

– to guarantee the representativity of the data sample, both adjectives constituting the triple in question had to belong to the class of adjectives which make up 90% of all adjectives in the whole BNC;
– the triples had to be neither left- nor right-branching structures; accordingly, it was checked whether a conjunction could be inserted or not; (3a)/(3b) and (4a)/(4b) exemplify a right-branching structure and a left-branching structure respectively.

(3) a. *some red balls*
    b. **some and red balls*

(4) a. *medium sized tank*
    b. **medium and sized tank*

– compound forms like *polar bear* or *High Commissioner* were filtered out because they are conceptualized as a single entry by the speaker so that the order is fixed and any adjective adding to the compound will precede it rather than divide it in half. Items were counted as compound forms if they have a separate entry in the Collins Cobuild on Compact Disc, Version 1.2 (1995) dictionary next to the entries found for their component words;
– triples containing *young*, *little*, or *old* were excluded from the sample (cf. Note 1).

The final sample size yielded 3,234 adjective$_1$–adjective$_2$–noun triples. There is, to my knowledge, no investigation of AO with such a large database. In her analysis of semantically incongruent adjectives, for instance, Richards (1977) created a large variety of examples of semantically incongruent adjective pairs.

However, having a look into the present data sample, it becomes obvious that semantically incongruent adjective combinations constitute roughly 14 out of 3,234 (0.5%) instances. Thus, even though Richards's (1977) analysis of these adjective combinations may be valid in its predictions, it is – quantitatively – not representative of naturally-occurring oral communication. Similarly, other previous analyses which consider only color and shape adjectives are not representative either: if one counts the triples in the present data sample in which only one of the two adjectives belongs either to the class of color or shape adjectives, only about 10% of the sample are accounted for (328 out of 3,324 tokens in the sample contain a shape and/or a color adjective). Thus, the present analysis considerably exceeds previous analyses of AO regarding the representativity of the data analyzed.

4. Corpus-linguistic operationalizations and monofactorial results

4.1 Phonological factors

4.1.1 *Length (*Length*)*

Given equal importance of both adjectives, longer adjectives have been found to follow shorter adjectives (cf. Behaghel 1930), reflecting a general tendency in language for the longer member of a word pair to follow the shorter member (cf. Cooper & Ross 1975; Bock 1982). For instance, Goyvaerts (1968: 13) states that (5a) is preferable to (5b).

(5) a. *the long intelligent book*
    b. *the intelligent long book*

Although length constraints have been subjected to empirical analysis for a variety of different syntactic variation phenomena (e.g. Cooper & Ross 1975; Wasow 1997), no empirical results are available for AO. For the present analysis, length is measured in letters. One might argue that as the data analyzed here all stem from the oral subcorpora of the BNC, one should operationalize length on the basis of phonemes rather than letters. However, given the very strong correlation between length in letters and length in phonemes ($r = .93$; $p < .001$ in a test sample of 2,362 words, Gries, p.c.), the gain in efficiency by having software count the letters of the adjectives clearly outweighed the perhaps more precise but much more cumbersome manual analysis of phonemic lengths.

In order to determine the influence of each individual variable on AO, a so-called PRE (Proportional Reduction of Error)-measure was calculated. I will explain this measure in some detail here, taking the example of LENGTH; for the subsequent variables, only the results will be provided. PRE-measures are measures of correlation which estimate how much the prediction accuracy for a dependent variable can be improved by knowing one (or more) independent variables (or, to put it the other way around, how much the rate of wrong predictions can be reduced). The logic behind such measures of correlation is the following. If one wants to predict which value a dependent variable will have, one would yield at least some correct predictions already by mere guessing (generally, given a sufficiently large number of trials, 50% of the predictions would be correct, the other 50% would be false). However, if one knows about some independent variable which correlates with the dependent variable in question, and if one also knows which values of the independent and the dependent variables correlate which each other, the prediction accuracy can be improved. Taking the example of AO, trying to predict whether an adjective will prefer to take the position of $adjective_1$ or $adjective_2$ could be based on mere guessing. Given a sufficiently large number of trials, one would make a considerable number of wrong as well as correct predictions. However, if one also knows that LENGTH influences AO, one can improve the prediction accuracy by predicting that all short adjectives should prefer to appear as $adjective_1$, while all long adjectives should prefer the position of $adjective_2$. The number of wrong predictions made should decrease in proportion to the explanatory power of LENGTH.[3] With respect to AO, the prediction accuracy could be improved by 15.66% if the adjectives' length was taken into account, yielding an overall prediction accuracy of 57.83%. Accordingly, the mean lengths of $adjective_1$(6.18) and $adjective_2$(6.72) differ highly significantly ($t_{Welch} = 8.528$; $df = 6402$; $p < .001$***).[4]

4.2 Syntactic factors[5]

4.2.1 *Nominal character (*NOMCHAR*)*[6]

According to Biber et al. (1999:599), there is an "overall tendency for the most nounlike modifiers to occur closest to the head noun." Posner (1986:316), comparing color and size adjectives, terms this factor Nouniness Principle; he also gives several criteria for determining the nominal character of the adjective:

– color words are more easily used as objects (*I like white* vs. ?*I like round*);
– color words take adjective attributes more easily (*This is an interesting white* vs. ?*This is an interesting round);*
– color words can more easily be used in the oblique case (*The homogeneity of this white is remarkable* vs. ?*The homogeneity of this round is remarkable)*.

However, this operationalization is inadequate in two respects. Firstly, the applicability of Posner's gap-filling sentences is all too sensitive to the kind of adjective under investigation. Inserting, for instance, the adjective *bottom* into Posner's gap-filling sentences results in the following sentences: *I like bottom*, *This is an interesting bottom*, and *The homogeneity of this bottom is remarkable* respectively. Although the exact location of these sentences on an acceptability scale is unknown, it is clear that they are far from being unequivocally acceptable. So in the case of *bottom*, one would infer that it is very adjectival, since it (more or less) refuses insertion into any of the three gap-filling sentences. However, *bottom* intuitively appears to be an atypical adjective. The second major shortfall of Posner's operationalization of NomChar is that, contrary to his own claim that nouniness is a gradual phenomenon, he does not explicitly explain in what way his operationalization can be interpreted as gradual: how to interpret successful insertion into one, two, or all three of the gap-filling sentences types? Is successful insertion into any of these sentences to be weighted as equally strongly indicating nouniness, or is insertion into the first more strongly indicating nouniness than in the other two sentences? And how are varying degrees of acceptability that supposedly result in judgments of the various sentences to be interpreted?

For the present analysis, the following procedure was adopted. Since Posner's test sentences apparently imply a test for the probability of occurrence of the adjectives' zero-derived nouns, the operationalization is based on the assumption that the more often an adjective is used as a zero-derived noun, the more nouny it is in general – purely adjectival concepts, on the other hand, will refuse zero-derivation. Accordingly, for all adjectives in the present data sample, I determined how often they are tagged as adjectives and as nouns in the whole BNC. On the basis of these frequencies, an index was calculated ranging between 0 (very adjectival) and 1 (very nouny). Consider the formula in (6) and Table 1 below for examples from the present data sample.

(6) $$\text{NomChar} = 1 - \frac{\text{frequency}_{\text{tagged as adjective}}}{\text{frequency}_{\text{tagged as adjective}} + \text{frequency}_{\text{tagged as noun}}}$$

Table 1. Examples of adjectives and their corresponding NomChar index values

adjective	frequency with adjective-tag	frequency with noun-tag	index
woolen	509	0	0
white	18,701	3,726	0.166
round	3,188	1,575	0.331
secret	2,047	1,478	0.419
bottom	1,126	3,979	0.779

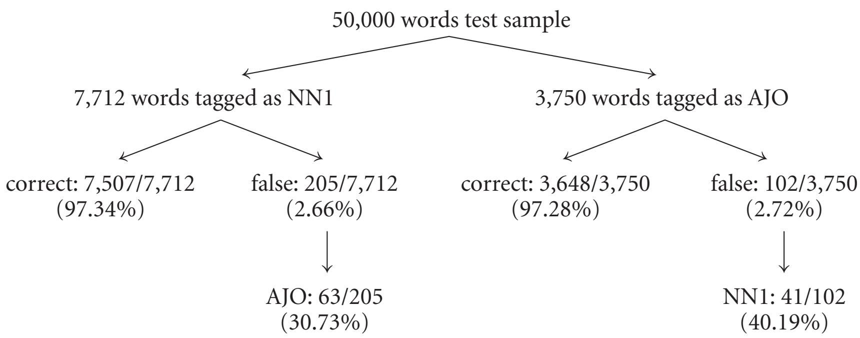

Figure 1. Tagging accuracy in the BNC (adapted from Leech & Smith 2000)

Two points require more detailed discussion here. Firstly, it has to be acknowledged that to calculate the NomChar values by counting the frequencies of noun (NN1) and adjective (AJO) tags is highly dependent on the correctness of the tags in the corpus, which also brings about some disadvantages. With respect to NomChar, one problem was that instances of polysemy were not automatically filtered out. For example, the core meaning of the noun *round* is not zero-derived from the polysemous shape adjective, but the procedure is not sensitive towards this matter of fact and treats all polysemous instances alike. Even though these polysemous items have been filtered out manually,[7] the general reliance on the correctness of the tags prevails. Fortunately, there is good reason to do so: as Leech and Smith (2000) point out, the automatic word class tagging for the second version of the BNC (employed here) is indeed highly accurate; consider Figure 1 for an overview of correct and wrongly classified nouns and adjectives in a 50,000 words test sample.

As can be seen in Figure 1, the overall error rate for noun as well as adjective tags is negligible (2.66% and 2.72% respectively) so, in this respect, the NomChar index values are valid.

Secondly, note that although NomChar index values may theoretically range between 0 and 1, they cannot be expected to be distributed across the whole range. Obviously, a NomChar value approaching one would mean that the adjective is hardly ever tagged as an adjective, undermining its membership in the word class adjective. So although it is expected that the mean NomChar values for adjective$_1$ and adjective$_2$ differ significantly, the significance of this difference need not necessarily correlate with a high effect size. For the investigation of adjectives, relatively small effect sizes are expected as adjectives constitute a word class the members of which should display similar features (even though these features need not apply rigidly and uniformly well to all members of a word class). This, however, is not incompatible with the assumption underlying the present analysis that within the limits of this word class, adjectives display more or less systematic and therefore predictable differences in their behavior as, e.g., with respect to their preferred position in a multi-adjectival phrase as investigated here. Accordingly, the examples in Table 1, while illustrating the possible range of NomChar values, are actually not representative of the distributional variance of the NomChar values for all adjective types in the data sample: out of 1,154 adjective types, the large majority of 1,028 types (89.1%) take on a NomChar value between 0 and 0.1. Nevertheless, the examples in Table 1 also show how the more objective operationalization adopted here actually validates one's intuition in two ways: on the one hand, it assigns adjectives like *round* or *white* relatively low nouniness values (thereby supporting Posner's analysis of these adjectives); on the other hand, it assigns *bottom* an extremely nounlike value (where Posner's analysis would have produced counter-intuitive results).

The PRE-measure for NomChar yielded an error reduction of only .47%, the overall prediction accuracy amounts to 50.23%. The mean NomChar index values of adjective$_1$(.047) and adjective$_2$(.038) differ highly significantly ($t_{Welch} = 3.500$; df = 5981; p <. 001 ***), yet the directionality of this influence is diametrically opposed to the one that has been hypothesized in the literature: according to these results, nouny adjectives do not directly precede the nouns they modify. Consider (7) for examples of preferred orderings and their NomChar values. I have to admit here that I cannot offer a straightforward answer to this puzzling result; in Section 4.5, possible methodological shortcomings are discussed.

(7) a. *nice* (.038) *clean* (0) *plaster*
b. *red* (.091) *big* (0) *flowers*

c. *white* (.166) *fluffy* (0) *cat*
d. *plain* (.327) *white* (.166) *trousers*

4.3 Semantic factors

4.3.1 *Semantic closeness (*SemClose*)*

In a sequence of adjectives, those denoting non-inherent qualities precede inherent adjectives (cf. Whorf 1945; Biber et al. 1999: 599). Probably the best-known analogue to this factor is one of Behaghel's Laws which states that things belonging closely together in mind are also put closely together in communication. This concept of semantic interplay and dependence between words as reflected in this variable is given theoretical support by various analyses (cf., e.g., Kilgarriff 1997). With respect to adjectives, even more direct support is provided by Stubbs (2001: 32f.), who refers to semantically overlapping words as delexicalized words; he concludes:

> "We now have several cases where units of meaning do not coincide with individual words. Taken separately, they look like minor exceptions to the idea that individual words have individual meanings, but taken together, they start to throw considerable doubt on the status of words as the normal units of meaning." (Stubbs 2001: 34)

Accordingly, Stubbs distinguishes so-called selecting and focussing adjectives. Selecting adjectives share characteristics such as being independent, being separately chosen and adding separate meaning whereas the latter are characterized as being dependent, co-selected with the noun, and repeating part of the meaning of the noun (Stubbs 2001: 33). To conclude, selective adjectives are another term for what is referred to here as semantically less close adjectives while the term focussing adjective is a paraphrase of adjectives which are semantically close to their head nouns. Richards (1975: 201) has related this concept of semantic overlap between adjectives and head nouns to Ziff's (1960) concept of the adjectives' different "privilege of occurrence": adjectives differ in the degree to which they may occur in different contexts. Her line of argumentation is the following: since semantically close adjectives semantically depend on the particular head nouns they modify and are (most often) co-selected with particular nouns only, it follows that these adjectives will collocate with a small number of different nouns only; on the contrary, adjectives which are semantically not close to their head noun(s) may provide information which is not already (partially) captured by the meaning of the noun, so they are semanti-

Table 2. Examples of adjectives and their corresponding SemClose values

adjective	corpus frequency	number of different noun collocates	SemClose value
new	114,655	10,016	.087
red	11,605	1,777	.153
brave	1,571	342	.218
nasty	1,787	461	.258
slight	2,830	997	.352

cally not bound to particular nouns only, but are good candidates for modifying a wide range of different head nouns. As, according to Behaghel's Law, closely related items also stand close together, semantically close adjectives will directly precede their head nouns. On the other hand, adjectives which have a low or even no semantic overlap with their head nouns are further removed from them in position.

Accordingly, the semantic closeness of the adjectives in the present data sample was measured via the number of different head nouns that the adjective in question collocates with. For all 1,154 adjectives in the present data sample, it was checked in the whole BNC how often they occurred with any noun. More precisely, three concordances and corresponding frequency lists were produced, as not only the adjective in its positive form, but also in its comparative and superlative forms had to be included to achieve a representative picture of the span of nouns the adjective collocates with. The resulting frequency lists had to be checked manually for potential double counts of nouns, i.e. cases where a noun collocates with an adjective in its positive as well as with any/both of its compared forms. The resulting number of different noun collocates was relativized against the corpus frequency of the adjective in question because adjectives which are generally more frequent than others will automatically have a greater number of different noun collocates. Consider Table 2 for examples from the data sample.

Comparing the mean SemClose values of adjective$_1$ (.195) and adjective$_2$ (.19), adjectives occurring in the first position of an adjective pair have slightly more different noun collocates than adjectives occurring in the second position. This tendency is marginally significant ($t_{Welch} = 1.953$; df = 6440; $p \approx .051$). The PRE-measure yielded an error reduction of 21.76%, which results in an overall prediction accuracy of 60.88%. Consider (8) for examples of the preferred order (the SemClose values given in brackets).

(8) a. *light* (.280) *brown* (.218) *leather*

b. *ordinary* (.217) *married* (.057) *woman*
c. *cheap* (.333) *unskilled* (.150) *labor*
d. *solid* (.252) *stainless* (.027) *steel*

4.3.2 *Independence from comparison (*IndComp*)*

Adjectives which are less dependent on comparison are put nearer to the head noun (cf. Martin 1969b; Posner 1986). Compare, for instance, the adjectives *round* and *heavy*: while one can identify a round object in a given set of other objects only by looking for the round object itself, the selection of a heavy object necessitates the comparison of the weights of – at least some of – the other objects, as one does not know by lifting only one object whether it may be called heavy in comparison to the others.

The operationalization of this variable was motivated by the obvious relationship between an adjective's independence from comparison and its gradability: if an adjective denotes reference-point dependent attributes, this is reflected at the level of grammar via gradability (cf. Ertel 1971). Translating this into more technical terms, it is argued here that comparison-independent adjectives are primarily used in the positive form, but much less so in their comparative or superlative forms. Comparison-dependent adjectives, however, should occur significantly more often in the comparative and superlative forms than in their positive form. This is based on the assumption that the number of objects in the non-linguistic, real world that have to be compared in order to decide on the appropriate adjective finds its corresponding expression in language in form of the frequency with which forms of degree are used. Accordingly, the frequencies of analytic and synthetic comparatives/superlatives of every adjective in the sample were summed and relativized against the frequency of the adjective in all its forms in the 100 million words of the whole BNC,[8] i.e. the adjective's lemma frequency because general adjective frequency naturally correlates with the number of compared forms. The lemma frequencies of the adjectives were calculated by summing up the adjective's corpus frequency and the number of synthetic comparatives/superlatives found for the adjective in question.[9] The operationalization process for IndComp is represented in (9); Table 3 provides examples from the data sample.

(9) $$\text{IndComp} = \frac{\sum_{\text{compared forms}}}{f_{\text{lexeme}}} = \frac{f_{\text{analytic forms}} + f_{\text{synthetic forms}}}{f_{\text{adjective}} + f_{\text{synthetic forms}}}$$

Table 3. Examples of adjectives and their corresponding IndComp values

adjective	$f_{analytic}$	$f_{synthetic}$	$\sum$ compared forms	$f_{adjective}$	$\text{IndComp}_{adjective}$
brown	3	32	35	3,908	.009
successful	998	0	998	10,803	.092
thick	1	464	465	3,173	.128
good	55	46,203	46,258	74,839	.382
high	24	19,725	19,749	28,698	.408

Note that the examples in Table 3 also intuitively conform to the above-mentioned description of an adjective's independence from comparison: it is relatively easy to identify a brown object in a group of objects; likewise, it requires very special circumstances to imagine a situation in which one would tend to speak of different degrees of brownness. Accordingly, the IndComp value for *brown* is very small (.009), meaning that it is mostly independent from comparison. With respect to *high*, on the other hand, its relatively high IndComp value of .408 matches one's intuitive feeling that height is an attribute that is used very often in the context of comparing things. The mean IndComp values for adjective_1 (.088) and adjective_2 (.0045) differ highly significantly (t_{Welch} = 16.81; df = 5945; p < .001***). Adjectives standing further from their head nouns occur with more forms of degree than adjectives directly preceding the head noun, which supports the results from previous works. Knowing the adjectives' IndComp values improves the prediction accuracy by 35.78% (total prediction accuracy: 67.89%). Examples of preferred orderings with IndComp values are given in (10).

(10) a. *pleasant* (.063) *elderly* (.005) *doorman*
b. *pale* (.067) *pink* (.008) *pants*
c. *awful* (.012) *American* (.001) *word*
d. *small* (.163) *red* (.008) *file*

4.3.3 *Dixon's semantic classes (*Dixon*)*

In his analysis of inter-language class correspondences, Dixon identifies seven subclasses of adjectives, arguing that "each semantic type has, in a particular language, its own particular norm and extensional grammatical properties" (Dixon 1977: 30), the latter also including its position in a string of adjectives. In accordance with Dixon's original proposal, all adjectives in the data sample were classified as belonging to one of Dixon's seven adjective classes (increasing

Table 4. Dixon's (1977) semantic classes and their hypothesized order

sequence	$adj_{Value} > adj_{Dimension} > [\ldots] > adj_{Age} > adj_{Color} > noun$ →						
semantic class	Value	Dimension	Physical Property	Speed	Human Propensity	Age	Color
code	7	6	5	4	3	2	1
Dixon's examples	*good*, *bad*	*big*, *long*	*hard*, *sweet*	*quick*, *slow*	*happy*, *kind*	*new*, *young*	*red*, *white*

numbers indicating increasing distance from the noun). Consider Table 4 for an overview.

As Dixon claims that adjectives denoting origin, composition, purpose or beneficiary of the head noun are post-adjectival modifiers, these adjectives were coded with 0. Even though these adjectives are not part of Dixon's definition of proper adjectives, they can well be predicted along Dixon's lines always to be the ones closest to the noun. For the majority of adjectives, classification was straightforward because the classes that Dixon distinguishes are relatively clear-cut and Dixon provides various examples for each class. However, for a considerable number of adjectives, the classification scheme proved incomplete: e.g., for all adjectives denoting date and time (*pre-war*, *annual*, etc.), position (*inner/outer*, *left/right*) and many forms in *-ic*/*-ical* which simply mean 'pertaining to', 'of', or 'relating to' (*mathematical*, *physical*, *financial*, etc.), Dixon neither makes any predictions about the position of these adjectives in a sequence, nor does he explicitly exclude them from the class of proper adjectives. Consequently, 1,701 adjective tokens (458 adjectives types) had to be excluded from the analysis as they could not be assigned (unambiguous) membership in either of the seven classes. Roughly 39% of these 1,701 adjectives are forms ending in *-al*, the three most common of which were *local* (100), *financial* (57), and *political* (50).[10] For an overview of the monofactorial results of the Dixon variable, consider Table 5.

The overall distribution of data in the table is highly significant (χ^2 = 492.29; df = 7; p < .001***). The cells which are responsible for this significance (i.e., significant contributions to Chi-square) are printed in bold letters; all significant contributions to Chi-square are significant at the .05% level of significance. That is, the frequencies of adjective$_1$ and adjective$_2$ deviate significantly from the expected frequencies for the following adjective classes: as to classes 1 (Color) and the extra-class 0, one finds significantly more adjectives directly in front of the noun than further away. In contrast, adjectives which

Table 5. Distribution of adjective$_1$ and adjective$_2$ with respect to Dixon's (1977) semantic classes

semantic class	0	1	2	3	4	5	6	7	row totals
adjective$_1$	**203**	**62**	122	**1,254**	7	**156**	365	**376**	2,545
adjective$_2$	**379**	**227**	105	**858**	11	**237**	365	**40**	2,222
column totals	582	289	227	2,112	18	393	730	416	4,767

Table 6. Examples of preferred and non-preferred orderings with respect to Dixon

semantic class	preferred ordering	non-preferred ordering
0	*thick woollen material*	*Indian blue dress*
	flimsy wooden building	*woollen soft material*
1	*clean white sweater*	*red round table*
	big pink elephant	*black heavy dirt*
3	*brave new world*	*new bold student*
	pleasant elderly doorman	*white heterosexual men*
5	*creamy white bourbon*	*black porous substance*
	light blue paper	*red hot coke*
7	*good clean teeth*	*round good idea*
	important new publication	*speedy good ball*

belong to classes 3 (Human Propensity), 5 (Physical Property) and 7 (Value) significantly more often occur further from the noun than directly in front of it. Taking the Dixon variable into account raises the prediction accuracy by 19.17%; the overall prediction accuracy is 62.32%. Consider Table 6 for examples of preferred orderings, as well as examples which, although attested, were found significantly less frequently with respect to the adjectives' semantic class.

4.3.4 *Subjectivity-objectivity gradience (SubObj)*

Quirk et al. (1985: 1341) argue for a subjective-objective gradience determining AO such that "modifiers relating to properties which are (relatively) inherent in the head of the noun phrase ... will tend to be placed nearer to the head and be preceded by modifiers concerned with what is relatively a matter of opinion." Hetzron (1978) has elaborated on this suggestion, emphasizing the non-discrete character of this scale; he worked out 13 classes (cf. Table 7 below), which were used for the operationalization of this variable. As with Dixon's scale, each adjective in the present sample was assigned membership in one of the 13 classes; again, the higher the number, the further the adjective (class) is positioned from the head noun. It has to be noted here that, ac-

Table 7. Hetzron's (1978) subjectivity-objectivity gradience and hypothesized order of adjectives

sequence	$adj_{Epistemic\ qualifier}$>$adj_{Evaluation}$>[…]>$adj_{Composition}$>$adj_{Purpose/Destination}$>noun →												
semantic class	Epistemic qualifier	Evaluation	Static permanent property	Sensory contact property	Speed	Social property	Age	Shape	Color	Physical defect	Origin	Composition	Purpose/Destination
code	13	12	11	10	9	8	7	6	5	4	3	2	1
example	*famous*	*good*	*wide*	*sweet*	*fast*	*cheap*	*young*	*square*	*blue*	*deaf*	*Asian*	*wooden*	*ironing*

Table 8. Distribution of $adjective_1$ and $adjective_2$ with respect to SubObj

SubObj	1	2	3	4	5	6	7	8	9	10	11	12	13	row totals
$adjective_1$	6	60	249	10	**62**	18	122	872	8	**78**	440	**663**	21	2,609
$adjective_2$	15	31	321	20	**227**	36	104	905	11	**153**	412	**142**	4	2,381
column totals	21	91	570	30	289	54	226	1777	19	231	852	805	25	4,990

cording to Hetzron, adjectives denoting origin, purpose/destination as well as composition are explicitly included in the class of proper adjectives (classes 1 and 2 respectively); similarly, Hetzron does not explicitly exclude any adjectives or adjective classes. Accordingly, no item was coded with 0. However, as with Dixon's classification scheme, many adjectives could not be assigned (unambiguous) membership and therefore had to be excluded.[11] Consider Table 7 for an overview.

As with the Dixon factor, a cross-tabulation displays the monofactorial results for SubObj; consider Table 8.

Again, the overall distribution in this table is highly significant (χ^2 = 492.89; df = 13; p < .001***). Significant contributions to Chi-square include

Table 9. Examples of preferred and non-preferred orderings with respect to SubObj

semantic class	preferred ordering	non-preferred ordering
5	*beautiful black skin*	*yellow round head*
	huge black cloud	*green skinny legs*
10	*stiff white collars*	*white fluffy cat*
	smelly black bananas	*red big flowers*
12	*wonderful new system*	*fat ugly mother*
	tremendous Irish wit	*red gorgeous colour*

classes 5 (Color), 10 (Sensory Contact Properties), and 12 (Evaluation). That is, adjectives which belong to the class of evaluating adjectives occur significantly more often at the first position of an adjective pair rather than at the second, noun-closer position, whereas adjectives belonging to any of the other three classes occur significantly more often at the position directly preceding the head noun. Note here that the fact that not all classes exhibited a significant distribution should not be interpreted as weakening the explanatory power of this approach, but rather as giving further support of, e.g., Hetzron's (1978) suggestion that the classes established are ordered on a continuous scale. According to the PRE-measure for this variable, knowing the SubObj values for the adjectives improves the prediction accuracy by 16.17%; the overall prediction accuracy amounts to 60%. Table 9 provides examples from the data sample; again, both preferred as well as non-preferred orderings are given.

4.3.5 *Semantic congruity and affective load (*SemCon/AffLoad*)*

Richards (1977) compared order preferences for semantically congruent (i.e. adjectives which may be linked by the conjunction *and*) and incongruent adjectives (i.e. adjectives potentially linked with *but*), examples for which are given in (11) and (12), respectively.

(11) a. *The poor, wretched child begged on the street*
b. *The poor and wretched child begged on the street*

(12) a. *The poor, happy child begged on the street*
b. *The poor but happy child begged on the street*

Richards also identified each adjective with a sign (plus or minus) on its dominant semantic dimension to indicate its affective load.[12] As to the above examples, *happy* would get a positive sign, *wretched* and *poor* negative ones. Richards reports the following results: semantically congruent adjectives have (i) significantly more stable ordering preferences in attributive as well as predicative contexts, (ii) significantly subjectively stronger prenominal ordering preferences, and (iii) are judged significantly more acceptable (cf. Richards 1977: 494–499). With semantically incongruent adjectives, the factor of affective loading comes into play: Richards argues the ordering of incongruent adjectives is governed by a 'first the good news, then the bad news' principle, i.e. positively signed adjectives precede adjectives carrying a negatively signed adjectives. Boucher and Osgood's (1969) "Pollyanna Hypothesis" provides more general support for this hypothesis, which states that there is a universal human tendency to report the positive rather than the negative aspects of life (cf. also Bock 1982). For instance, (13a) was preferred to (13b) in 80% of all cases (cf. Richards 1977: 498).

(13) a. *strong dangerous*
b. *dangerous strong*

(Note in passing that the preference of (13a) over (13b) can likewise be explained in terms of the LENGTH variable, which drives home the point made earlier that only a multifactorial analysis will help to clarify the variables' exact impact). Richards operationalized affective load as a dichotomous variable, i.e. an adjective can either be positively or negatively loaded. With respect to her data sample, this might have been a feasible operationalization because all of her adjective pairs are prototypical antonyms (*happy-sad, hot-cold*, etc.), which clearly enhances the probability that these adjectives constitute two extremes on one semantic dimension. Likewise, the adjectives Richards investigated are core members of the semantic dimensions she considered. However, for the present data, such a priori restrictions would mean to restrict the representativity of the data sample. Consequently, there are many items which do not fit well into Richards's operationalization. Firstly, for many adjectives, it is rather difficult to decide which of the three semantic dimensions offered by Richards is most appropriate, if not to render the dimension of Potency more or less a wastebasket-category (for instance, to classify *silky*, *grammatical*, and *brown* as belonging to the domain of Potency seems counterintuitive). A further consequence of this is that many (if not most) adjectives are not affectively loaded in any direction (i.e. negatively/positively), but rather constitute purely neutral terms. This assumption is supported by Deese (1964: 355) who notes that "[s]ome adjective pairs do not define scalar dimensions", one of the most striking exceptions being color adjectives. Moreover, Richards's assignment of positive and negative affective loadings, affective load being defined in logical terms, does not do justice to the context-dependence of many adjectives,

Table 10. Corpus examples of semantically congruent and incongruent adjective pairs

corpus example	affective load	semantic congruity value	semantic congruity status
nice cosy glow	+1/+1	1	semantically congruent
silly daft woman	–1/–1	1	semantically congruent
big red mark	0/0	1	semantically congruent
fantastic new world	+1/0	0	semantically congruent
terrible personal problem	–1/0	0	semantically congruent
awful funny feeling	–1/+1	–1	semantically incongruent

even of those she herself investigates. The interpretation of *hot* and *cold* with respect to their position on a semantic dimension largely depends on the referent they modify: in the case of *hot tea*, e.g., *hot* clearly has a positive affective load, yet the same adjective in *hot room* does not (in most cases) lend itself to such an interpretation; rather, it seems that here, the polarity of the dimension is reversed.

Consequently, Richards's (1977) operationalization was slightly modified so as to overcome the above-mentioned problems. With respect to affective load, three values could be assigned to an adjective: 1 (positive affective load), 2 (neutral), and 3 (negative affective load). Assignments are based on the adjectives' definition in a dictionary, where affective load is commented upon with phrases like "used showing (dis)approval".[13] Accordingly, semantic congruity was measured including three levels, the values depending on the combination of the adjectives and their respective values regarding their affective loading; Table 10 provides an overview.

As can be seen from Table 10, the specification of the semantic congruity value has been made to distinguish between semantic congruity value 1 (adjective combinations with identical affective values, be they positive, negative, or neutral), semantic congruity value 0 (adjective combinations with non-identical affective values, i.e. one adjective having a neutral affective load, the other being either positively or negatively loaded), and semantic congruity value –1 (adjective combinations with opposing affective values). Consequently, one can immediately tell apart 'mixed' adjective pairs from the first group, which might be useful if it should actually show that these two combinatory types behave in different ways with respect to order restrictions. As to the adjective combinations with semantic congruity value 0, for instance, it should be interesting to see whether the 'first the good news, then the bad news' principle also leads to a preference for 0/–1 and +1/0, respectively, or

Table 11. Combinations of AffLoad values with respect to the position in a pair

adjective$_1$	adjective$_2$			row totals
	–1	0	1	
–1	17	79	4	100
0	63	2,328	62	2,433
1	10	647	24	681
column totals	90	3,054	90	3,234

whether this principle only holds for clearly loaded adjectives. A dichotomous variable would not allow for such comparisons. Consider Table 11, which provides the numbers for the nine possible combinations of positively, neutrally, and negatively loaded adjectives pairs.

The crucial aspect to note is that positively loaded adjectives$_1$ marginally significantly more often precede negatively loaded adjectives$_2$ than vice versa according to a one-tailed binomial test of the distribution of 10:4 ($p_{binomial}$ = .089).[14] However, positively-loaded adjectives significantly precede neutrally-loaded adjectives more often than vice versa since the distribution of 647:62 is highly significant ($p_{binomial} < .001$ ***). Finally, the distribution of neutral and negatively loaded adjectives does not deviate from expected frequencies significantly (79 : 63, $p_{binomial} = .104$).

4.4 Pragmatic factors

4.4.1 *Noun-specific frequency (*NSpecFreq*)*[15]

Lockhart and Martin (1969) were able to demonstrate that those adjectives that tend to stand closest to a noun are the ones which are remembered most easily upon the occurrence of the noun. Referring back to work by Shapiro, who "has established that less reaction time in memory experiments indicates greater strength of association" (Posner 1986: 330) and Martin's (1969b: 479) assumption that strength of association is a function of frequency and degree of proximity of two words, Posner (1986: 330) concludes:

> "It is frequent use that makes an attribute a stereotype and it is the character of being a stereotype that increases the frequency of a word in everyday communication still further. Thus, the conclusion is obvious that an attribute is put nearer to its head noun the more often it is used together with it. Local proximity to the head noun indicates greater noun-specific frequency."

Table 12. Examples of adjective-noun pairs and corresponding NSpecFreq values

adjective+noun	p(Adj+N)	p(N)	p(Adj|N)
murky area	3	35,145	.000
big boat	15	5,403	.002
true story	154	13,677	.0113
blue dress	55	3,679	.015
serious damage	101	5,779	.0175

For the present analysis, noun-specific frequency was operationalized by calculating for every adjective-noun combination the conditional probability p(Adj|N), i.e. the probability to know the adjective(s) given the knowledge of the noun; cf. the formula as represented in (14).

(14) $$p(\mathrm{Adj}|\mathrm{N}) = \frac{p(\mathrm{Adj}+\mathrm{N})}{p(\mathrm{N})} = \frac{f(\mathrm{Adj}+\mathrm{N})}{f(\mathrm{N})}$$

As to the collocational probability of adjectives and nouns, a frequency list comprising all adjective-noun combinations in the whole BNC was conducted, amounting to about 1.4m combinations. So for every adjective being part of an adjective$_1$–adjective$_2$–noun triple in my data sample (e.g. adjective$_1$ *big* and adjective$_2$ *red* in *big red house*), it could be checked how often this adjective collocates with a particular noun (i.e. how often do we find *big house* and *red house* respectively?). The calculation of p(N) was measured in terms of general corpus frequency (cf. also Lapata et al. 1999). Table 12 provides examples from the data sample.

Comparing the mean NSpecFreq values of adjective$_1$ (.004) and adjective$_2$ (.023) in the present data sample, adjectives occurring closer to the head noun have higher NSpecFreq values than those further from the noun. This tendency is highly significant ($t_{Welch} = -11.684$; df = 3371; $p < .001$***). Accordingly, knowing the adjectives' NSpecFreq values reduces the prediction error by 43.91% (overall prediction accuracy: 71.95%). For examples of preferred orderings and their NSpecFreq values, consider (15).

(15) a. *normal* (0) *free* (.023) *tickets*
b. *plain* (.001) *hard* (.009) *facts*
c. *nice* (.001) *blonde* (.018) *hair*
d. *excellent* (.002) *new* (.013) *book*

4.4.2 *Frequency (*GenFreq*)*

Finally, the general frequency of the adjective has been claimed to influence its position in a string (cf., e.g., Bock 1982) such that more frequent adjectives precede less frequent ones. Ney (1983) argues that if two adjectives are equally frequent, the difference in familiarity of the two adjectives would determine the ordering (cf. Ney 1983: 101). However, I would rather argue that frequency and familiarity correlate positively in that the former is strongly influenced by the latter and vice versa. That is, speakers are assumed to make use of words they are familiar with, and the more familiar they are with these words, the more often they will employ them to express what they want to say. Findings from research in language production also support this claim: frequently used words have a higher probability of winning out against "competing" synonymous words because frequency of usage increases the word's resting level (cf., e.g., Bock 1982). Methodologically, Lapata et al. (1999) provide further support, as they also measure familiarity by means of corpus frequency. Therefore, frequency of adjectives will also be operationalized in this way and be equated with the speakers' familiarity with this adjective. Comparing the mean frequencies of $\text{adjective}_1 (18,671)$ and $\text{adjective}_2 (14,460)$, adjectives occurring in the first position of an adjective pair are highly significantly more frequent than those adjectives which occur directly before the noun ($t_{Welch} = 7.374$; df $= 6307$; $p < .001$***). That is, knowing the adjectives' corpus frequency reduces the prediction error by 17.92%; the overall prediction accuracy raises to 58.96%. Examples of preferred orderings are provided in (16), together with their corpus frequency in the BNC.

(16) a. *strong* (15,898) *brown* (3,908) *bags*
b. *damp* (1,571) *salty* (178) *margins*
c. *big* (24,684) *cold* (6,438) *lakes*
d. *lively* (1,472) *vibrant* (264) *conference*

4.5 Summary of monofactorial results

By and large, the monofactorial analyses of the various factors that have been proposed in the literature were confirmed to have a significant influence on AO. Moreover, the hypothesized directionality of the variables' impact was also supported in the present analysis except for two factors, NomChar and AffLoad, which I want to discuss in some more detail in the following.

Firstly, NomChar had a highly significant impact, yet one diametrically opposed to its hypothesized directionality: on average, more nouny adjectives are placed further away and less nouny ones closer to the noun rather than the other way around. How can this result be explained? It cannot be ruled out, admittedly, that the way in which the variable was operationalized simply is not adequate reproduction of this factor. However, although I have already explained why the fact that 89.1% of all adjectives having NomChar values between 0 and .1 does not threaten the methodological validity of the procedure per se, it may hint to the conclusion that the operationalization of NomChar applied here is not inadequate, but probably incomplete in the sense that NomChar should be more adequately conceived of as a multifactorial construct of which the tendency towards nominalization is just one aspect.[16] Another possible argument could be that the tagging accuracy yielded for the BNC is not yet sufficiently high to ensure correct NomChar values. As can be deduced from Figure 1, the overall error rate for adjectives and nouns is negligibly small; however, the number of nouns misclassified as adjectives and vice versa actually is substantial (40.19% and 30.73% respectively), suggesting that the two word classes which enter into the computation of NomChar values are particularly prone to be confused by automatic taggers. However, as the absolute number of errors involved (63 + 41 = 104) is still so small, I would place the burden of proof on those who wanted to claim that those (and only those) 104 misclassified items account for the surprising result for NomChar as it stands. To conclude, as nouniness is one of the few factors not subjected to corpus-linguistic analysis before, it is legitimate to argue that the present results indicate that NomChar has no impact on AO, or if at all, then in the opposite direction (even though I cannot offer an explanation for this directionality at present).

Secondly, AffLoad yielded only marginally significant results: the Pollyanna Hypothesis did not receive strong support when applied to contrasting adjectives, but fared rather well when applied to pairs of positive and neutral adjectives. This result is at least in part due to the rarity of pairings of positively and negatively loaded adjectives (14 cases out of 3,234 adjective pairs).

For all remaining factors, however, and especially for those factors which have been empirically tested in previous analyses so that a comparison of the results is possible, we see that the corpus-based operationalizations are adequate. This is a non-trivial issue since it demonstrates that, to the extent that previous hypotheses which found support in my analysis are correct, corpus data can actually be used to investigate mechanisms hitherto only identifiable via experimental research. Put differently, we have obtained prima facie evidence that corpus data can also tap into those processes that traditional linguistic analyses on the basis of judgment data etc. have failed to illuminate objectively.

Still though, I have already argued that this way of analysis is not yet sufficient since, whenever speakers subconsciously decide how to order the adjectives in question in a particular situation, they of course do not consider the value of one factor only – rather, in actual discourse, all of the factors are given at the point of time where a speaker finally has to decide for either order. Moreover, examples such as in (13) clearly necessitate a multifactorial approach, because otherwise, it is unclear which variable is more likely to determine the ordering preference, in this case LENGTH or AFFLOAD. Thus, in the subsequent section I will subject the data sample to a multifactorial analysis using a Linear Discriminant Analysis (LDA).

5. Testing the variables multifactorially

A Linear Discriminant Analysis tests how well we can discriminate between different values of a dependent or so-called grouping variable, given some set of independent variables. In the present analysis, the grouping variable has two possible values, namely either the position adjective$_1$ or adjective$_2$; the independent variables are LENGTH, NOMCHAR, etc. as presented and discussed above. In other words, for every adjective in the data sample, we test whether the values that the adjective takes on with respect to the independent variables enable us to predict whether the adjective will generally prefer to occur as either adjective$_1$ or adjective$_2$.

Before presenting the results, it has to be mentioned that the data sample had to be modified in some respects in order to yield interpretable results. Firstly, the DIXON factor was excluded, which, as the monofactorial results have shown, makes no predictions that are fundamentally different from the SUBOBJ factor. Secondly, the problem that LDA does not permit ordinal variables had to be overcome. For this reason, the ordinal SUBOBJ factor had to be converted into a categorial factor with two levels; accordingly, the variable was split into two halves: the first group comprises the adjective classes 1 to 7, thus representing adjectives which denote objective properties of their head noun's referents, while the second group includes the adjective classes 8 to 13, representing adjectives which denote subjective properties. The choice of the turning point between classes 7 (Age) and 8 (Social Property) is motivated theoretically insofar as the truth value of adjectives belonging to classes 1 (Purpose/Destination) to 7 (Age) is objectively falsifiable, yet the truth value of adjectives belonging to classes 8 (Social Properties) to 13 (Evaluation) may not be objectively falsified; rather, these adjectives express individual opinions or beliefs (cf. Hetzron 1978: 179). Thirdly, the factor SemCon was excluded from the multifactorial part of the analysis as this variable makes no direct predictions about the position of an adjective in a string – rather, SemCon states that ordering restrictions are expected to be more rigid with semantically congruent adjectives than with semantically incongruent ones. Finally, the factor AffLoad was re-coded because the monofactorial results have clearly shown that the only level of AffLoad which actually contributes to the variable's impact is the presence or absence of positive load. In addition, dummy recoding of variables with more than two levels regularly inflates intercorrelations and increases the difficulty of staying above the minimum tolerance level required for the analysis. Thus, restricting the attention to the most important predictor level of AffLoad also avoids purely mathematical problems. Accordingly, for the LDA, this variable was re-coded as a categorial variable with two levels (1 = presence of positive load, 0 = absence of positive load).

Let us now turn to the results of the LDA. Consider Table 13 for the overall results.

As Table 13 shows, the output of an LDA is, first of all, a Wilks's Lambda value and a canonical correlation value both of which, roughly speaking, represent the fit of the discriminant function. A Chi-square test is used to check whether the overall discriminant analysis yields a significant result. Table 13 shows that on the basis of the variables that were included into the present analysis, the two positions in an adjective$_1$–adjective$_2$–noun sequence can be distinguished clearly, i.e. the analysis supports that these variables have a considerable impact on which position an adjective will preferably take.

Moreover, the discriminant analysis also weighs each variable with respect to how well it can discriminate between the two groups, adjective$_1$ and adjective$_2$ (producing the smallest degree of error). Therefore, it assigns a factor

Table 13. Overall results of the Linear Discriminant Analysis (LDA) I

Wilk's Lambda	.861
canonical correlation	.373
Chi-square (df = 8)	740.23
p	< .001***

loading to each variable which can be interpreted as reflecting the variable's relevance for the discrimination between adjective$_1$ and adjective$_2$.[17] These factor loadings range between –1 and 1; the closer the loadings approach either extreme, the more important is the variable; loadings approaching zero characterize unimportant variables. In the present analysis, negative factor loadings mean that there is a negative correlation between the variable values and the adjective's position in a pair such that the higher the variable value, the more likely the adjective will occur as adjective$_1$. Positive factor loadings indicate the reverse. For example, NSpecFreq would be expected to have a positive factor loading, as an adjective with a high noun-specific frequency is hypothesized to occur preferably as adjective$_2$, whereas an adjective with a low noun-specific frequency should rather occur as adjective$_1$. Generally, factor loadings of $\pm$.22 can be assumed to have an impact on AO worth mentioning because these values can account for 5% of the variance.

Interpreting Table 14, the most important predictor of AO in the LDA is AffLoad. Recall that the monofactorial results did not provide direct support for the "First the good news, then the bad news" principle, as oppositely loaded adjectives were not distributed significantly. However, my suggestion that positively loaded adjectives nevertheless influence AO so that speakers tend to be preoccupied with positive rather than negative qualities receives support from the multifactorial analysis: if an adjective is positively loaded, there is a strong tendency for it to occur as adjective$_1$ rather than as adjective$_2$. Except for SemClose, the LDA neatly weighs the variables according to the linguistic subbranch to which they belong: semantic variables are most important, followed by the pragmatic variables, which in turn are followed by Length, which is taken here to represent a phonological factor. The factor loading of the only

Table 14. Factor loadings of variables in the LDA I

variable	factor loading
AffLoad	–.821
IndComp	–.555
SubObj	–.362
NSpecFreq	.348
GenFreq	–.203
Length	.174
SemClose	–.083
NomChar	–.022

syntactic variable investigated here, NomChar, is so small that it may even be denied a substantial impact on AO. So it seems plausible to argue that this variable's influence, be it significant in isolation, is overridden by the impact of the other variables in a multifactorial setting. For instance, recall examples like (7b) and (7c), repeated here as (17a) and (17b) respectively.

(17) a. *red big flowers*
b. *white fluffy cat*

Both triples are examples of preferred NomChar orderings as *red* and *white* yield higher NomChar index values than their partners *big* and *fluffy*. However, at the same time, these constitute examples of non-preferred orderings of the variable which the LDA considers the second most important, namely IndComp. Indeed, examples like these where NomChar wins out against IndComp are significantly rare (580 out of 3,234 triples in the present data sample; χ^2 = 13.168; df = 1; p < .01**).

Next to the factor loadings, the LDA also provides a classification function which calculates how many adjectives were classified correctly by the LDA as adjective$_1$ or adjective$_2$ using the discriminant function.[18] Table 15 shows how many adjectives were correctly and falsely classified.

As can be deduced from Table 15, the discriminant function correctly classified 1,304 + 1,832 = 3,136 adjectives (i.e. 63.3%). However, this figure has to be corrected because the LDA's classification of an adjective's position cannot simultaneously take into account the properties of the other adjective in the pair under investigation. That is, the position of a particular adjective is derived from its own discriminant score alone (cf. above) irrespective of the discriminant score of the other adjective. Thus, the wrongly classified adjective positions are cases where the LDA has assigned the same position to both adjectives, a result that is, of course, somewhat counterintuitive. Fortunately, however, the LDA provides enough information to allow for a straightforward correction of this problem: rather than just considering the final classification of the LDA, we can also turn to the individual adjectives' discriminant scores

Table 15. Classification accuracy of LDA I

	observed adjective$_1$	**observed adjective$_2$**	**row totals**
expected adjective$_1$	**1,304**	1,272	2,576
expected adjective$_2$	549	**1,832**	2,381
column totals	1,853	3,104	4,957

and determine whether the difference between the scores is in the predicted direction. The results are dramatic: the number of misclassified cases drops from 1,821 to 1,093 (i.e. by 40%), thereby at the same time increasing the rate of correct classifications to 78%. Again, this classification accuracy is impossible to achieve by chance ($p < .001^{***}$). To clarify the procedure, consider the expression *vital national importance*. Looking at the discriminant scores of both adjectives reveals that they are far from being equally likely to occur in the predicted second position: the discriminant score of *national* is 1.6 times higher than the one for *vital*, indicating that *national* is much more likely to occur in second position than *vital*. This likelihood can in fact be given precisely: *national* is 12.3% more likely to be adjective$_2$ than *vital*, as can be inferred from the posterior probabilities of the LDA that figure in the classification (cf. Backhaus et al. 1996: 129–134). Summing up, on the basis of the independent variables that have entered the LDA, a very satisfactory classification accuracy is achieved.

However, it has to be noted that the classification accuracy gained by the LDA cannot be equated with a prediction accuracy as the LDA has classified all adjectives a posteriori. Indeed, one would like to know how accurately the independent variables postulated serve to predict the order of unknown adjective combinations because this reflects more adequately the actual situation which a speaker has to face when – subconsciously – planning an utterance. Therefore, a "leave-one-out" cross-validation was conducted, which predicts every individual adjective's preferred position on the basis of all 4,956 remaining adjectives in the sample. The prediction accuracy of this cross-validation was 63.2%, ranging only slightly below the one of the first overall LDA. Again, this percentage could not have been achieved by guessing ($p < .001^{***}$). Concluding, by means of the independent variables investigated here, adjective strings may not only be classified correctly with significant accuracy; what is more, yet unanalyzed adjective sequences can be predicted equally accurately.

So far, the LDA only served to identify and predict one adjective's (let us refer to it as x) most typical position regardless of the particular second adjective it co-occurs with (let us refer to it as y); i.e. whether the adjective x will generally prefer to occur as adjective$_1$ or adjective$_2$ in some triple. In a more technical parlance, we were concerned with $p\ (x = \text{adjective}_1)$and $p\ (x = \text{adjective}_2)$.However, one might object that it remains unclear how x behaves if we know y. In an again more technical way of expression: the objection might be whether we should not rather be concerned with $p\ (x = \text{adjective}_1|y)$, i.e. the conditional probability that x occupies the first/second position within the pair,

Table 16. Overall results of the LDA II

Wilk's Lambda	.657
canonical correlation	.585
Chi-square (df = 16)	1651.331
p	< .001***

given that the other adjective to be ordered is y. In order to provide an answer to this question, a second LDA was calculated which analyzed the adjectives in a pairwise fashion, namely which position does x take if it is accompanied by y? Table 16 reports the overall results.

The (ranking of the) factor loadings of the independent variables differ only marginally (GenFreq and Length swap places). The classification accuracy gained by the second LDA is even higher than for the first: 73.8%. What is more, the cross-validation also raises to 73.5% ($p < .001$*** for both values). Concluding, on the basis of the independent variables included here, it is also possible to predict adjectives' preferred position in yet unanalyzed triples with high accuracy.

5.1 Summary of multifactorial results

As the LDA has shown, the semantic variables included in the present analysis are by far the most important ones in predicting AO. The only exception is SemClose, the influence of which could not be supported by any of the analyses. Although Martin (1969a) reported similar results in his analysis, I still hesitate to neglect a potential influence of this variable. In their analysis of measures of semantic similarity, Miller and Charles (1991: 2) express the conflict most concisely:

> "[S]emantic similarity has become one of those ubiquitous and important variables, like familiarity or frequency of occurrence, that are often used to explain psychological phenomena, but are seldom seen as being in need of explanation. This degree of acceptance for semantic similarity is remarkable in view of the theoretical complexity of the judgment involved. Yet subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty."

Moreover, as reported above, Martin's (1969a) methodology has not been without criticism. So how can this result be accounted for? It cannot be ruled out, of course, that the operationalization of SemClose adopted in the present study is not a felicitous one. Indications for this are also provided by Miller and Charles (1991). Considering semantic similarity as a function of the contexts in which words are used, they criticize contextually-based estimates of contextual similarity (as my operationalization of SemClose can also be labeled). Their argument is the following:

> "The problem with co-occurrence measures is not merely that they dismember the contexts they are supposed to represent. A more serious problem is that they do not approach these tasks the way people do – whatever a word's contextual representation may be, it is certainly not a collection of other words. If the argument advanced here is correct, people's knowledge of how to use a word is organised to enable them to recognise rapidly the contexts it goes into." (Miller & Charles 1991: 23)

However, Miller and Charles also acknowledge that for the alternative estimate of semantic similarity based on substitutability, "there is no quick and easy computer algorithm for calculating" (Miller & Charles 1991: 23f.); what is more, Gries (2001, 2003b) has successfully used a context-based asymmetric measure of semantic similarity. So although substitutability-based estimates are the theoretically more attractive alternative, to date, contextually-based estimates are the only technically fully workable technique for a corpus analysis like the present one.

A possible argument against the validity of the present study is to say that a correct prediction in 73.5% of all cases actually does not seem overly accurate. However, besides that fact that this result is statistically significant, there are also linguistic arguments to support my claim that this prediction accuracy is indeed valid. First of all, in a real production situation, many more (potentially non-linguistic) variables will enter into the production process which have not been considered here (as they cannot be accounted for in a corpus analysis and are even difficult to control in experimental designs) such as, e.g., the attentional state of the speaker, his physical condition, subconscious stimuli he/she receives from his/her environment, etc. Adding up the potential impact of these variables, a prediction accuracy of 73.5% may already by interpreted in a different light. However, I am aware of the fact that if misclassifications are not interpreted as being indicative of a sub-optimal model, but are rather explained away as having been expected (or even intended according to the theoretical commitments underlying the model), one could easily object that (potential) falsification is no longer possible. With respect to the present analysis, however, I do not want to claim that a prediction accuracy of 73.5% is the best possible result; indeed it is indicative of the fact that there are probably other variables which exert influence on AO which have yet to be identified.

6. Conclusion

To conclude, the present analysis has shown that AO, although superficially a purely syntactic phenomenon, was shown to be determined by a variety of variables from different levels of linguistic analysis. Moreover, the corpus-based methodology adopted here was demonstrated to be a widely applicable tool for linguistic analysis; the corpus-linguistic operationalizations found for some of the variables which were included may be employed for the analysis of various other phenomena. They could also be relatively easily applied automatically to other texts, as they – with the exception of the semantic classes – exclusively rely on information that can be directly extracted from the corpus (annotations). Moreover, the present approach has combined frequency information and symbolic data in a way that invites implementation into natural language generators (cf. Malouf (2000) on combining memory-based learning and positional probabilities in predicting AO). Finally, these operationalizations as well as the statistics used for the empirical analysis could be shown to yield similar explanatory and predicting power as reported from experimental studies, which means that this combination of methodology is an attractive complement to experimental designs.

Notes

* I am indebted to (in alphabetical order) Thomas Berg, Stefan Th. Gries, Klaus-Uwe Panther, Ulrich Schade and two anonymous reviewers for extensive comments on and discussion of an earlier version of this paper. It has to be acknowledged here that methodologically, the present paper heavily falls back on Gries (2003a), who demonstrated how the multifactorial approach and the statistical tools applied here can be fruitfully employed for the analysis of syntactic variation. Naturally, all remaining errors are the author's.

1. Although the adjectives *young*, *little*, and *old* would meet these criteria, they were not included into the data sample because several authors have attributed these adjectives a special status (cf., e.g., Goyvaerts 1968; Martin & Ferb 1973; Hetzron 1978). Vendler (1968: 132) has argued that "these three adjectives tend to form a sort of petrified compound with certain nouns". Similarly, Taylor (1992) distinguishes absolute and synthetic adjective senses: whereas adjective-noun pairs including the former adjective type are analyzable in terms of compositionality, the meaning of phrases including the latter adjective senses rather "emerges from a subtle interaction between the meaning of the noun and the meaning of the adjective" (Taylor 1992:2). As phrases including *young*, *little*, or *old* could involuntarily distort the picture, these special adjectives were not investigated here.

2. I decided to restrict the data sample to oral data so as not to involuntarily distort the picture due to register-specific differences. However, even if it cannot be ruled out that ordering constraints differ, it is plausible to assume that oral language production is influenced by the speaker's knowledge of both registers. Accordingly, whenever it was necessary to determine frequencies, ranges of collocates etc. which enter into the calculation of factor values, the whole BNC was used.

3. It has to be noted here that the measure used here is conceptually identical with a PRE-measure called Lambda (cf. Bortz, Lienert & Boehnke 1990:340–341). Although Lambda is normally only applied to $k \times m$ crosstabulations, it can also be applied in this context because the logic behind the measure remains the same.

4. A t_{Welch}-test is employed for comparing the mean values of variables on interval level in two independent samples, given homogeneity of variances; cf. Bortz (1999:137–140).

5. In many formalist approaches to AO, it has been argued that the adjectives' different transformational investment may account for ordering preferences. However, there is overwhelming evidence from linguistic, psycholinguistic and psychological research for AO to be constrained by principles not residing exclusively in the domain of syntax; this is incompatible with the fundamental commitments of (most) formalist approaches to language. What is more, the equivalence of attributive and predicative structures as well as the derivation of the former from the latter via transformational operations poses unsolvable problems even within the boundaries of a transformational-generative approach (cf., e.g., Bolinger 1967). Therefore, the variable of transformational investment will not be included into the present analysis.

6. After their first mention, variables will be referred to using the abbreviated forms given alongside the introduction of the variables in order to provide coherence with the output of the statistical analyses.

7. In a processing-based account of AO, however, it could be argued that the polysemous nouns should actually be included into the calculation of the NomChar value: as they are co-activated whenever the intended zero-derived noun is activated, they contribute to the NomChar value of an adjective as (subconsciously) perceived by the speaker.

8. As many adjectives undergo slight orthographic changes when forms of degree are attached to them (**fuzzyer*, **safeer*, etc.) or have irregular forms of degree (*good/better/best*), the synthetic comparatives of all adjectives were also checked manually to ensure correct numbers.

9. For those adjectives which have analytic compared forms, these are already included in the corpus frequency because the searching procedure is not sensitive towards the fact that the adjective may be preceded by *more/most* and therefore is part of a comparative/superlative.

10. It cannot be ruled out, admittedly, that some of the adjectives I excluded native speakers would feel able to assign clear membership. However, given that Dixon defined his semantic classes in a universal manner such that they could be employed for a cross-linguistic study, the adjectives that finally were included into the analysis could be assigned clear membership without further reliability tests.

11. Although Hetzron's classification scheme was used here without any restrictions or modifications, it has to be acknowledged that Hetzron's definition of several classes is not without problems, which rendered unambiguous classification difficult for many adjectives. For example, the class Speed is defined as a property which "shows only when the entity is in motion" (Hetzron 1978: 180). My objection here is that I consider it well possible to call, say, a car a *fast car* even if it is not moving at the moment I am making this judgment, as fast cars have various characteristics which also show when the car is not driving, e.g. the maximum speed displayed by the speedometer, the car's aerodynamic look, etc. Similarly, Hetzron argues for the distinction between Evaluations and Epistemic Qualifiers saying that the former are more ad hoc judgments which are often accompanied by an affective value, whereas Epistemic Qualifiers are used "where the speaker's earlier experiences, accumulated knowledge are the conditioning factors" for making the judgment (Hetzron 1978: 181). Given this definition, I fail to see how the influence of the speaker's experience and knowledge should render the judgment less objective than an ad hoc evaluation; rather, I would argue that experience and knowledge normally tend to enhance the objectivity of a statement. A related problem (which also holds for many ambiguous cases in the other classes) is that even thorough analysis of the larger context may not provide sufficient information for unambiguous class assignment (which actually is not surprising given the gradual character of the SubObj scale). Concluding, both adequate definitions of semantic classes as well as automatic semantic class assignment remain a notoriously difficult aspect (not only) in corpus-linguistic analyses. Future research may proceed along the lines of Malouf (2000), who calls for the use of distributional clustering techniques to extract semantic classes directly from corpora themselves.

12. Referring to Osgood's (1964) work in semantic differential research, she assigned each adjective a value on three dimensions of meaning (Evaluation, Potency, and Activity), the dominant semantic dimension being the one on which the adjective yields the highest factor loadings. To give an example, the dominant semantic dimension of *happy* would be Evaluation, *strong* would yield highest loadings in the domain of Potency, and *slow* is an adjective the most dominant semantic trait of which is in the domain of Activity. However, some of the assignments Richards employed appear somewhat unmotivated: why, for instance, is Activity the underlying semantic dimension of the adjective pairs *wet-dry* and *young-old*? Even though one may argue that research with semantic differential scales falls back on association tests with native speakers, the results may also be a consequence of the fact that only three semantic dimensions were offered overall; completely free association tests, on the contrary, might have provided different results (cf. also Deese's (1964) classic analysis of the associative structure of English adjectives).

13. The dictionary used was Collins Cobuild on Compact Disc, Version 1.2 (1995). Note here that this way of identifying the adjectives' affective load is not context-sensitive, either; therefore, the sample had to be checked manually for ambiguous items like *hot* in *hot tea* or *hot room*.

14. A binomial test was used here because it is distribution-free and not impaired by small cell frequencies (cf. Gries 2003b).

15. The assignment of the variables to one of the sub-branches of linguistics is not always straightforward; this is particularly true for the frequency-related variables presented under the label of pragmatics, which might well be argued to be positioned more adequately elsewhere. However, I decided to take over the categorization as proposed throughout the literature.

16. One potential shortcoming of the operationalization of NomChar via the adjectives' preparedness for zero-derivation was pointed out to me by Thomas Berg (p.c.): adjectives in contemporary English (as opposed to other languages) show a tendency to be relatively resistant towards nominalization. However, this should primarily have quantitative consequences, and it still remains to be seen whether this general reluctance of English adjectives to be nominalized touches upon my qualitative expectation that more adjectival adjectives should be (even) less often nominalized than more nouny adjectives.

17. There is an ongoing discordance whether a discriminant function should rather be interpreted on the basis of the so-called standardised discriminant function coefficients. In this analysis, I follow, among others, Bortz (1999: 588, 595–596) and rely on the factor loadings instead.

18. This classification is based on a discriminant function resulting in a discriminant score for each adjective: if this discriminant score is lower or higher than a to-be determined threshold value (in this case .015), then the LDA suspects the adjective to occur as adjective$_1$ and adjective$_2$ respectively.

References

Bache, C. (1978). *The Order of Premodifying Adjectives in Present-day English*. Odense: Odense University Press.

Backhaus, K., B. Erichson, W. Plinke, & R. Weiber (1996). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. 8th ed. Berlin, Heidelberg, New York: Springer.

Behaghel, O. (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde, 44*, 81–89.

Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan (Eds.) (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bloomfield, L. (1933). *Language*. Chicago, London: Chicago University Press.

Bock, J. K. (1982). Toward a Cognitive Psychology of Syntax: Information Processing Contributions to Sentence Formation. *Psychological Review, 89*, 1–47.

Bolinger, D. L. (1967). Adjectives in English: Attribution and Predication. *Lingua, 18*, 1–34.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler*. 5th edition. Berlin, Heidelberg, New York: Springer.

Bortz, J., G. A. Lienert, & K. Boehnke (1990). *Verteilungsfreie Methoden in der Biostatistik*. Berlin, Heidelberg, New York: Springer.

Boucher, J., & C. E. Osgood (1969). The Pollyanna Hypothesis. *Journal of Verbal Learning and Verbal Behavior, 8*, 1–8.
Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Cobuild on CD-ROM (1995). Glasgow: Harper Collins.
Cooper, W. E., & J. R. Ross (1975). World Order. In R. E. Grossman, J. L. San, & T. J. Vance (Eds.), *Chicago Linguistic Society: Papers from the Parasession on Functionalism* (pp. 63–111). Chicago: Chicago Linguistic Society.
Danks, J. H., & S. Glucksberg (1971). Psychological Scaling of Adjective Order. *Journal of Verbal Learning and Verbal Behavior, 10*, 63–67.
Danks, J. H., & M. A. Schwenk (1972). Prenominal Adjective Order and Communication Context. *Journal of Verbal Learning and Verbal Behavior, 11*, 183–187.
Danks, J. H., & M. A. Schwenk (1974). Comprehension of Prenominal Adjective Orders. *Memory and Cognition, 2*, 34–38.
Deese, J. (1964). The Associative Structure of Some Common English Adjectives. *Journal of Verbal Learning and Verbal Behavior, 3*, 347–357.
Dixon, R. M. W. (1977). Where Have all the Adjectives Gone? *Studies in Language, 1*, 19–80.
Ertel, S. (1971). Pränominale Adjektivfolgen und semantische Tiefenstruktur. *Studia Psychologica, 13*, 127–137.
Goyvaerts, D. L. (1968). An Introductory Study on the Ordering of a String of Adjectives in Present-day English. *Philologica Pragensia, 11*, 12–28.
Gries, St. Th. (2001). A Corpus-linguistic Analysis of *-ic* and *-ical* Adjectives. *ICAME Journal, 25*, 65–108.
Gries, St. Th. (2003a). *Multifactorial Analysis in Corpus Linguistics*. London and New York: Continuum.
Gries, St. Th. (2003b). Testing the Sub-test: A Collocational-overlap Analysis of English *-ic* and *-ical* Adjectives. *International Journal of Corpus Linguistics, 8* (1), 31–61.
Hetzron, R. (1978). On the Relative Order of Adjectives. In H. Seiler (Ed.), *Language Universals* (pp. 165–184). Tübingen: Narr.
Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*, 91–113.
Lapata, M., S. McDonald, & F. Keller (1999). Determinants of Adjective-Noun Plausibility. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 30–36.
Leech, G., & N. Smith (2000). *Manual to Accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. <http://www.comp.lancs.ac.uk/ucrel/bnc2postag_manual.htm>.
Lockhart, R. S., & J. E. Martin (1969). Adjective Order and the Recall of Adjective-Noun Triples. *Journal of Verbal Learning and Verbal Behavior, 8*, 272–275.
Malouf, R. (2000). The Order of Prenominal Adjectives in Natural Language Generation. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 85–92.
Martin, J. E. (1969a). Some Competence-Process Relationships in Noun Phrases with Prenominal and Postnominal Adjectives. *Journal of Verbal Learning and Verbal Behavior, 8*, 471–480.

Martin, J. E. (1969b). Semantic Determinants of Preferred Adjective Order. *Journal of Verbal Learning and Verbal Behavior, 8*, 697–704.

Martin, J. E. (1970). Adjective Order and Juncture. *Journal of Verbal Learning and Verbal Behavior, 9*, 379–383.

Martin, J. E., & T. E. Ferb (1973). Contextual Factors in Preferred Adjective Ordering. *Lingua, 32*, 75–81.

McEnery, T., & A. Wilson (1997). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Miller, G. A., & W. G. Charles (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes, 6*, 1–28.

Ney, J. W. (1983). Optionality and Choice in the Selection of Order of Adjectives in English. *General Linguistics, 23*, 94–128.

Osgood, C. E. (1964). Semantic differential technique in the comparative study of cultures. *American Anthropologist, 66*, 171–200.

Posner, R. (1986). Iconicity in Syntax. The Natural Order of Attributes. In P. Bouissac, M. Herzfeld, & R. Posner (Eds.), *Iconicity. Essays on the Nature of Culture* (pp. 305–337). Tübingen: Stauffenburg Verlag.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (Eds.) (1985). *A Comprehensive Grammar of the English Language*. London: Longmans.

Richards, M. M. (1975). The Pragmatic Communication Rule of Adjective Ordering: A Critique. *American Journal of Psychology, 88*, 201–215.

Richards, M. M. (1977). Ordering Preferences for Congruent and Incongruent English Adjectives in Attributive and Predicative Contexts. *Journal of Verbal Learning and Verbal Behavior, 16*, 489–503.

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

Sichelschmidt, L. (1986). Optionality and Choice in the Analysis of Adjective Order: Comments on Ney. *Studia Linguistica, 40*, 135–148.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Cornwall: Blackwell.

Taylor, J. R. (1992). Old Problems: Adjectives in Cognitive Grammar. *Cognitive Linguistics, 3*, 1–35.

Teyssier, J. (1968). Notes on the Syntax of the Adjective in Modern English. *Lingua, 20*, 225–249.

Vendler, Z. (1968). *Adjectives and Nominalizations*. The Hague and Paris: Mouton.

Wasow, T. (1997). Remarks on Grammatical Weight. *Language Variation & Change, 9*, 81–105.

Whorf, B. L. (1945). Grammatical Categories. *Language, 21*, 1–11.

Ziff, P. (1960). *Semantic Analysis*. Ithaca, New York: Cornell University Press.