Premodifying -ing participles in the parsed BNC

Turo Vartiainen and Jefrey Lijffijt

University of Helsinki, Aalto University

Abstract
In this article we will focus on premodifying -ing participles in English. By premodifying -ing participles we refer to NP-internal -ing forms, such as the ones found in *a charming man* or *a barking dog*. Our first goal is to see how these participles are used in four registers of Present-Day English: academic prose, newspaper articles, fiction and conversations. Furthermore, we will attempt to find corpus evidence for Vartiainen (forthcoming), where it was suggested that there are solid grounds for dividing the class of premodifying -ing participles into adjectival and verbal -ing participles. Our hypothesis is that if this division is evident in the syntactic behaviour of the participles, then it might also be reflected in the way -ing participles are used in different registers.

1. Introduction

Earlier research has indicated that premodifiers are used in very different ways in different registers. For example, Biber et al. (1999: 65) and Biber (2007: 135) found that adjectival premodifiers (including prenominal -ing participles) are used more often in newspaper texts and academic prose than in fiction or conversations. This tendency has been explained by the fact that attributive premodifiers are related to the "informational production" of the text (Biber 1998: 128–9). By this analysis, nominals with complex premodifiers are seen as densely packaged information units, the function of which is to allow the author to express their message in an economical way (Biber 2007, Biber and Clark 2002).

Although we find this idea sensible, we also believe that different kinds of premodifiers may have different functions and that these functions may not be observable if the modifiers are lumped together too coarsely. In the aforementioned literature, for example, the way word classes have been categorised is somewhat questionable. For instance, Biber and Clark (2002: 46) provide *detecting* in *detecting devices* as an example of a "participial adjective". In our opinion, *detecting* is not an adjective in this phrase; it is ambiguous between a noun and a verb ('devices for detecting' or 'devices that detect'). Similarly, in the workbook to *The Longman Student Grammar of Spoken and Written English* (Conrad, Biber and Leech 2002: 14) the authors argue that *coming* in the phrase *the coming weekend* is an adjective, because "it precedes and modifies the noun weekend, and the meaning is 'the weekend which is coming'." In our view, it does not seem satisfactory to claim that a participle is an adjective because in a closely synonymous clause the corresponding meaning is

expressed by *a finite verb*. Consequently, we would call *coming* in *the coming weekend* a verbal participle.

The problems in the categorisation of premodifying -ing participles have recently been discussed by Huddleston (1984), Laczkó (2001) and Vartiainen (forthcoming). From a morphosyntactic standpoint, it seems clear that the strategy of lumping all kinds of -ing forms into one "adjective" category is quite problematic indeed. For example, the -ing forms in *an interesting article* or *a fascinating experience* behave very similarly to central adjectives, whereas those in *the approaching aircraft* or *the laughing man* do not (see e.g. Vartiainen, forthcoming, for some distributional tests for categoryhood). Because of limitations of space we can only provide some cursory remarks to the matter here, but suffice it to say that in this paper we adopt the stance that premodifying -ing participles may be divided into adjectival and verbal -ing participles based on their morphosyntactic properties.

It should also be pointed out that the word class of the participle is not predetermined. Rather, the verbhood or the adjectivehood of the participle always emerges in context as a result of the relationship between the head and the modifier. Consequently, it is practically impossible to device scripts that would be automatically able to categorise participles into adjectival and verbal participles in corpora, which also means that each participle+head construction needs to be analysed by the linguist. Table 1 illustrates the way -ing participles can be used as adjectives and as verbs:

Table 1. Adjectival and verbal participles.

Adjectival participle	**Verbal participle**
An irritating habit	An irritating substance
A stimulating lecture	A stimulating electrode
A glowing review	Glowing coal
Freezing water ('very cold water')	Freezing water ('the water's temperature is decreasing towards its freezing point')
Sparkling results	Sparkling quartz

The relevance of the above discussion for our study is two-fold. First, the division of -ing participles into adjectival and verbal participles allows us to study the differences in the four registers in more detail (see section 3). Second, the problems in the categorisation of the -ing participle are not only theoretical; they also result in problems on a more practical level of linguistic research, the annotation of corpora and data retrieval. In the BNC-XML, for example, word classes are annotated with the CLAWS-5 tagger, which categorises nearly all premodifying -ing participles as adjectives. However, the parsed version of the BNC (see Andersen et al. 2008), which is also used in this study, applies a modified version of the CLAWS tagset, and consequently, there is much more variation in the way the premodifying -ing participles are annotated. Here, some

premodifying -ing participles are tagged as adjectives, while others have been tagged as verbs. We will discuss the specific details of the way these corpora are annotated below, but at this point it should be pointed out that the theoretical problems related to the categorisation of words into word classes often result in varying and messy POS annotation, which not only makes the retrieval of relevant forms more difficult but also complicates data comparison across corpora (and across different studies). Moreover, these theoretical issues imply that POS annotation may not be a very useful or reliable tool for retrieving relevant word forms from a corpus, as the annotation scheme always reflects the theoretical stance of its makers, and this stance may not be shared by the linguist using the corpus.

Let us conclude this section on a practical note. Even though we believe that there are good grounds for dividing -ing participles into two categories, adjectival and verbal participles, this fine-grained division may in fact turn out to be rather problematic for the corpus compiler. As it happens, if no parsing information is available, the retrieval of the premodifying -ing forms may actually be more efficient when the annotation scheme regards all premodifying -ing participles to be adjectives. However, this is not to say that all premodifying -ing forms *are* adjectives: it merely goes to show that the label *adjective* is very often used for words that occur in the typical *function* of the adjective, i.e. as a modifier to a term. This leads into a simplified situation where the POS category (i.e. the word class, *adjective*) in fact includes syntactic and functional information (modifier + head relations). Therefore, the confusion between the form class and the word's function may be convenient for the purpose of making corpus queries, but it nevertheless amounts to confounding different levels of analysis and including parsing information in the POS tag (see Pullum 2009 for an excellent discussion on the confusion of different levels of analysis in linguistics; also see Denison 2007 for the problems related to the POS annotation of categories that are undergoing change).

We will now turn to the description of the corpora we have used in our research. The details of the corpora and the script that we have used to extract the relevant -ing forms are discussed in section 2. The results of our study are introduced in section 3, which is followed by a discussion of our findings in section 4.

2. BNC-XML and the parsed BNC

The data for our study comes from the British National Corpus (BNC), both from its XML version (BNC-XML) and a parsed version which is currently under development at the University of Cambridge (see Andersen et al. 2008). The BNC contains almost a hundred million words of Present-Day English, with ca. 90 million words from the written and 10 million words from the spoken domain. All word tokens in the BNC-XML have been automatically annotated using the

CLAWS-5 tagger.[1] The tagger uses a *hidden Markov model* for deciding the class of a word. The result is accurate, but not perfect, because the word class cannot always be decided on using the adjacent words only. Indeed, there may even be cases where human interpretations may differ. In the parsed BNC, a different POS-tagger has been used, which is incorporated in the *Robust Accurate Statistical Parser* (*RASP*). The set of tags is also slightly different. For brevity, we do not discuss the parser in detail. For our purposes it is enough to know that there are 22 relations. Each represents a relationship between a head word and a dependent word, except for the *passive*, which takes only one argument. The most relevant relation for us is *ncmod*, which states that one word is a non-clausal modifier of another word. Accuracy estimates for CLAWS-5 tagger can be found in the BNC manual[2] and accuracy estimates for RASP in Briscoe et al. (2006). To summarize, the following data is available to us in the parsed BNC: each word has both a CLAWS-5 tag and a RPOS (RASP POS) tag and each sentence has a variable number of grammatical relationships.

2.1 Instance retrieval

To find all instances of premodifying -ing participles in the parsed BNC, we need to write a query on the database, similar to writing a query in BNCweb. The problem is that we do not know in advance how the premodifying -ing participles have been annotated, and there are probably many other words matching such a straightforward query. For example, the BNCweb query **ing* gives about 2.75 million words, of which very few are premodifying participles. Using the more restrictive query *{*ing/ADJ}|{*ing/VERB}*, we can cut this down to about 400 thousand, but still the percentage of premodifying participles is very low. Clearly, a more sophisticated method to retrieve the relevant information is desirable.

The problem of retrieving words that are subject to intrinsic and grammatical constraints (premodifying -ing participles) can be viewed as a problem of building a classifier that *predicts* which words are interesting. We shall base this classifier on a set of rules. In this view, {**ing/ADJ}* is a classifier which predicts that all words tagged as adjective and ending with -ing are positive instances and all other words are negative instances. If the parser works perfectly, it will be very straightforward to construct such a classifier. A rule such as "*the POS tag is adjective AND the word is a non-clausal modifier of a later word"* would give us a perfect prediction if both the POS tag and the parser were flawless. Unfortunately, we know that they are not (see Briscoe et al. 2006 and Burnard 2007).

We will construct a model that consists of a set of rules that accurately predict the class. We restrict our search space to words that end with -ing. To choose the final set of rules, we need an estimate of the precision and recall of those rules. To do so, we have constructed a data set that we can use to *train* the model. We picked three texts from the written part of the BNC at random and

[1] http://ucrel.lancs.ac.uk/bnc2/bnc2autotag.htm
[2] http://www.natcorp.ox.ac.uk/docs/URG/posguide.html

extracted all words ending with -ing, including the full sentence as context. We then annotated all words as being a premodifying -ing participle or not. Table 2 lists the statistics of the training data.

Table 2: Statistics for the training data set based on texts A6G, B06 and C8A.

#words	#-ing words	#ambiguous cases	#premodifiers	#other forms
110990	2902	12	351	2539

We have checked the tag distributions for the C5 and RPOS tags and observe that the tags are quite similar for *-ing* words in general, but the CLAWS-5 tagger categorizes marginally more words as adjectives. Alarmingly, for the premodifying -ing participles the two taggers produce completely different results. The CLAWS-5 tagger assigns nearly all (93%) to the class of adjectives (AJ0), but the RASP tagger assigns roughly two thirds of the premodifying -ing participles as -ing verbs (VVG), while one sixth of the participles are tagged as adjectives (JJ) and one sixth as nouns (NN1).

The approach we take to construct a set of rules is simple. We choose the variables that our model is allowed to use and then we run a simple algorithm to incrementally add rules to the model:

(1) Compute for each possible rule an estimate of precision
(2) Sort all rules according to their estimated precision
(3) Start with an empty set of rules
(4) For *i = 1 to number of rules*
(5) Include the rule with highest precision
(6) Compute and store current precision and recall
(7) End

The set of possible rules is spanned by a conjunction over all variables, such that each rule covers a most specific set of instances and the rules have no overlap. Because there are only a few variables, the number of possible rules is quite limited. Hence, the computation is fast and straightforward.

We now have everything we need to construct a good set of rules for retrieving premodifying -ing participles from the parsed BNC. We run our algorithm for combinations of the variables corresponding to the part-of-speech tags and the grammatical relations. Figure 1 gives the results in terms of precision and recall for different sets of variables. The best possible performance would be to return all 351 premodifying participles and no other -ing forms. As expected, including all variables gives the best performance, but the performance is only marginally better than using only the CLAWS-5 tags. The results obtained using the RASP POS tags and grammatical relations are worse than expected: the CLAWS-5 tags give more information with respect to locating premodifying -ing participles. Note that each point on any of the lines corresponds to a set of rules.

When we want to pick a good set of rules, we are left with a trade-off between precision and recall. Low precision is bad because we get many false positives and have to do more manual post-processing. Low recall is bad because it produces more false negatives and that may cause bias in the final results. For our further analysis we require that the recall should be above 90 %. A reasonable trade-off seems to be the simple and straightforward rule to include all words categorized by the CLAWS-5 tagger as adjective and exclude all other -ing words. This model results in an estimated recall of ca. 92% and precision of ca. 67%, i.e. about one third of the final results are not premodifying -ing participles. An alternative that would allow automated processing of the results is based on using both tags and the grammatical information and gives 99% precision, but only 38% recall. However, the latter seems to be a poor choice for our purposes.

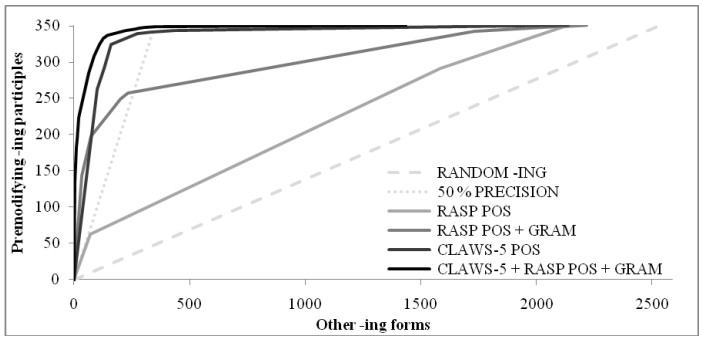

Figure 1: Precision vs. recall trade-off for different variables. The y-axis represents the number of returned premodifying -ing participles. The x-axis gives the number of other -ing forms recalled. Each line represents a set of variables.

3. Premodifying -ing participles in the BNC

As described in the section above, we chose to retrieve the -ing forms relevant for this study by using a script that had very high recall and a reasonably high precision. The script was applied to the four registers we were interested in, after which each participle was analysed manually. After completing the analysis, we were left with 3,434 NPs with a premodifying -ing participle. These NPs were divided across the four registers in the following way:

Table 3. Data included in the study.

Register	Files	Words	Nouns	-ing pcpls
Academic prose	15	595,380	152,802	1,490
Conversations	15	195,757	20,348	31
Fiction	15	603,199	106,971	1,519
Newspaper	10	187,632	46,889	394
Total	**55**	**1,581,968**	**327,010**	**3,434**

The files chosen to represent each register were selected randomly. For academic prose, we chose eight texts from the domain of social sciences and seven texts from natural and hard sciences. Although the selection process was indeed random in the sense that we did not study the data beforehand, we did try to choose texts of roughly equal length to ensure that the length of the text would have as little bearing on the overall results as possible. However, this could not be achieved for conversation data, as the conversation files in the BNC tend to be quite short. Therefore, the conversations in our data range from 2,161 words (file KPC) to 31,141 words (file KBG), the average number of words being 13,050. Moreover, the newspaper articles included in this study are actually collections of articles, which means that it is not sensible to compare lexical variation within the texts for this register (see Figure 4 below).

Before studying the use of adjectival and verbal -ing participles in more detail, let us take a look at the overall frequency of premodifying -ing participles in the four registers studied. Earlier research has indicated that conversations include more pronouns and fewer nouns than the other three registers, whereas academic prose and news contain more nouns and fewer pronouns (Biber et al. 1999: 92). As premodifying -ing participles by definition occur only before nouns, we have compared the number of -ing participles against the number of nouns instead of the number of words. Figure 2 illustrates the participle/noun ratio in the four registers:

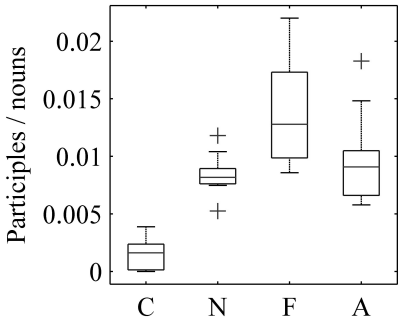

Figure 2. The frequency of the premodifying -ing participle in different registers.

We may first observe from Figure 2 that conversations are indeed markedly different from the other three registers: of the 20,347 nouns that have the potential of occurring with a premodifying -ing participle, only 31 actually do so (see Table 3). As previous studies (e.g. Biber et al. 1999) have indicated that prenominal modifiers are rare in conversations, this is an expected result.

Less expected may be the fact that premodifying -ing participles are used more frequently in novels than in academic prose or newspaper articles. This is contrary to earlier claims (e.g. Biber et al. 1999: 65), according to which -ing adjectives (a class into which other premodifying -ing participles are also

included) are much more common in newspaper texts and in academic prose than in fiction.

In order to obtain a clearer picture of the use of premodifying -ing participles in the four registers, we annotated each participle as a verbal participle or an adjectival participle according to the distributional properties of the participle (see e.g. Vartiainen, forthcoming, for these distributional tests). Figure 3 shows the frequency of verbal participles in the four registers:

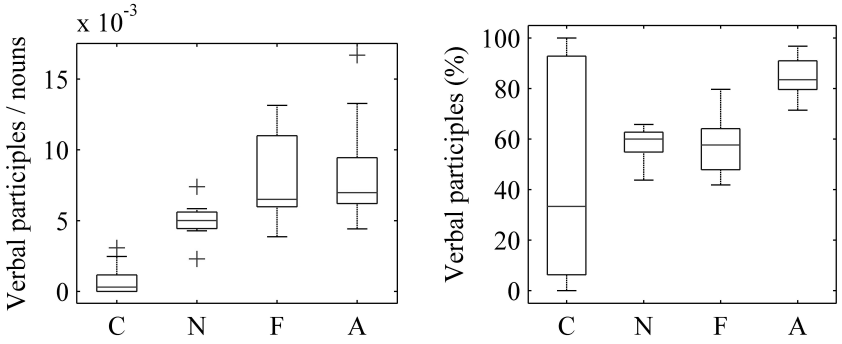

Figure 3. The frequency of verbal participles in different registers (left) and the proportion of verbal -ing participles of all premodifying -ing participles (right).

We can see from Figure 3 that, on average, the number of verbal participles is slightly larger in academic prose than in fiction or newspaper texts. However, the most significant difference across the four registers can be seen when the proportion of verbal and adjectival participles are compared: over 80 percent of all premodifying -ing participles are verbal participles in academic prose, while the corresponding proportions for newspaper texts and fiction are at around 60 percent. Clearly, the more fine-grained division of -ing participles into adjectives and verbs reveals a difference in the use of premodifying -ing participles.

The large proportion of verbal participles in academic prose can be explained in part by the use of certain frequently occurring participles, such as *following*, *preceding* and *succeeding*. Compared to fiction, where words like *following* almost always have a temporal function (e.g. *the following morning*), the function of these participles in academic prose is textual (e.g. *the following examples*). We suggest that in academic prose these participles offer the author a way to guide the reader's attention to a particularly significant or illustrative passage in a text (e.g. *the following examples*), or to re-introduce a previously mentioned referent into the reader's mind (e.g. *the preceding sections*). In other words, these participles are used as *foregrounding elements* in academic discourse. Moreover, their use is very specific to the academic register: 11.4 percent of all premodifying -ing participles in academic prose are foregrounding participles, whereas in newspaper articles only one percent of the participles have

similar function. We found only one single foregrounding participle in our fiction data, while there were no occurrences in the conversation data.

However, the use of foregrounding participles cannot by itself explain why the proportion of verbal participles is so large in academic prose. Another reason for the frequent use of verbal participles is that topical phrases including a premodifying -ing participle are often repeated in academic prose. For example, the file CS3, a text about social classes and different kinds of government, includes 160 premodifying -ing participles. Of these, the participle *ruling* occurs 54 times, while *governing* occurs 23 times. The use of foregrounding participles and the repetition of topical phrases also mean that there is much less lexical variation in academic prose than in fiction, for example. This difference is depicted in Figure 4. Quite strikingly, the two registers can be distinguished with very high precision by this single parameter (each data point represents a text file in the BNC).

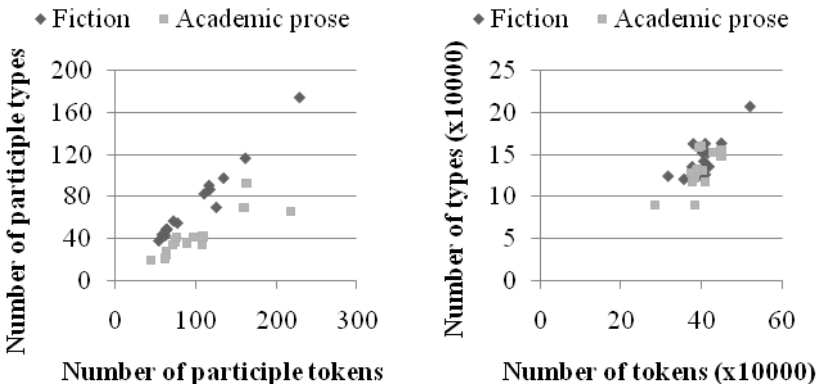

Figure 4. The number of unique participles in academic prose and fiction (left) and the type/token ratio in academic prose and fiction (right).

Importantly, the variance between the two registers described in Figure 4 is independent of the type/token ratio of the texts, which indicates that the difference in the lexical variation of -ing participles is not motivated by a more general lexical tendency.

4. Discussion and conclusion

In this article, we have concentrated on a single group of nominal premodifiers: -ing participles. We studied the frequency and use of the -ing participle in four registers, and our results can be seen as largely corroborating earlier research, while also presenting novel ideas. First, contrary to earlier research (Biber et al.

2002, Biber 2007), we found that in our data, premodifying -ing participles are most frequently used in fiction. This suggests that there is more to the use of NP-internal participial modifiers than just economy of expression or information packaging (see e.g. Biber and Clark 2002, Biber and Gray 2010). Second, we found that the division of -ing participles into verbal and adjectival participles is not motivated only on morphosyntactic grounds; rather, the differences in the syntactic behaviour of the participle classes have consequences for language use as well: while both adjectival and verbal participles are common in novels and newspaper articles, academic prose strongly favours verbal participles. This tendency can be explained by the repetition of topical phrases (e.g. *the ruling elite*), on the one hand, and the frequent use of foregrounding participles, such as *following*, on the other. The foregrounding function of this subclass of -ing participles can also be seen as a means of (re-)introducing a discourse referent. Thompson (1988) already noted that premodifying (attributive) adjectives often introduce a new discourse referent, while predicative adjectives typically have a characterising function (see also Englebretson 1997; Ford, Fox and Thompson 2003). Our data complements these findings, indicating that not only adjectival but also verbal premodifiers may have such referent-introducing function. We also suspect that adjectival participles can be used to express the author's attitude or stance more readily than verbal participles. This may also contribute to their relative infrequency in academic prose, which tends to be written in a more formal and more objective tone than novels and newspaper texts.

To conclude, we find that the differences in the morphosyntactic behaviour of adjectival and verbal -ing participles imply that the debate over the categorisation of the participle is not just a theoretical concern but that it is actually relevant for language users and is part of linguistic reality. This observation may present difficulties for the automatic (and often overly simplified) annotation of corpora, and it may also be inconvenient for the linguist who is trying to retrieve the relevant -ing forms from a corpus. On the other hand, mislabelled and miscategorised modifiers may create a false picture of the way language is used in different registers both synchronically and diachronically.

5. References

Andersen, Ø., J. Nioche, E.J. Briscoe and J. Carroll (2008), 'The BNC parsed with RASP4UIMA', in: *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC08)*, Marrakesh, Morocco.

Biber, D. (1998), *Variation across speech and writing*, Cambridge: Cambridge University Press.

Biber, D. (2007), 'Compressed noun-phrase structures in newspaper discourse', in: W. Teubert and R. Krishnamurthy (eds.), *Corpus linguistics: Critical concepts in linguistics*, Vol. 5, London: Routledge, 130–41.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *The Longman grammar of spoken and written English*. London: Longman.

Biber, D. and V. Clark. (2002), 'Historical shifts in modification patterns with complex noun phrase structures: how long can you go without a verb?', in: T. Fanego, M.J. López-Couso and J. Pérez-Guerra (eds.), *English historical syntax and morphology*, Amsterdam and Philadelphia: John Benjamins, 43–66.

Biber, D. and B. Gray (2010), 'Challenging stereotypes about academic writing: Complexity, elaboration, explicitness', *Journal of English for Academic Purposes*, 9, 2–20.

Conrad, S., D. Biber and G. Leech (2002), *The Longman student grammar of spoken and written English: Workbook*. London: Longman.

Briscoe, T., Carroll, J. and Watson, R. (2006), 'The second release of the RASP system', in: *Proceedings of the COLING/ACL on interactive presentation sessions*, Sydney: Australia.

Burnard, L. (2007), *Reference guide for the British National Corpus XML edition*. http://www.natcorp.ox.ac.uk/XMLedition/URG/

Huddleston, R. (1984), *Introduction to the grammar of English*. Cambridge: Cambridge University Press.

Denison, D. (2007), 'Playing tag with category boundaries', in: A. Meurman-Solin and A. Nurmi (eds.), *Annotating variation and change, VARIENG e-Series 1*, Helsinki: Research Unit for Variation, Contacts, and Change in English (VARIENG). http://www.helsinki.fi/varieng/journal/volumes/01/denison/

Laczkó, Tibor (2001), 'Another Look at Participles and Adjectives in the English DP', in: M. Butt and T. Holloway King (eds.), *Proceedings of the LFG01 Conference*. CSLI Publications.

Englebretson, R. (1997), 'Genre and grammar: Predicative and attributive adjectives in spoken English', *Berkeley Linguistics Society*, 23, 411–21.

Ford, C.E., Fox, B.A. and Thompson, S.A. (2003), 'Social interaction and grammar', in: M. Tomasello (ed.) *The new psychology of language: Cognitive and functional approaches to language structure*, Vol. 2, London: Erlbaum, 119–44.

Pullum, G. (2009), 'Lexical categorization in English dictionaries and traditional grammars', *Zeitschrift für Anglistik und Amerikanistik*, 57(3), 255–73.

Vartiainen, Turo (forthcoming), 'Telicity and the Premodifying ing-participle in English'. *Proceedings of the 30th ICAME Conference*.

Thompson, S.A. (1988), 'A discourse approach to the cross-linguistic category "adjective"', in: J. Hawkins (ed.), Explaining linguistic universals, Oxford: Basil Blackwell, 167–85.