ZAA 54.2 (2006): 163-176 ©

VOLKER GAST

The Distribution of *Also* and *Too*: A Preliminary Corpus Study

Abstract: This paper discusses distributional differences between the two additive particles *also* and *too* on the basis of corpus evidence. Three hypotheses are explored: (i) *also* and *too* are characteristic of different registers or styles, (ii) the use of *also* and *too* is sensitive to structural properties of the 'added constituent', and (iii) the use of the two particles depends on the distribution of 'added material' over the sentence. The discussion centres around the question to what extent corpus-based methods are appropriate to test these hypotheses. While hypothesis (ii) can straightforwardly be tested (and confirmed), the other two hypotheses pose methodological challenges of different types. Corpus-based test procedures are outlined for both hypotheses and preliminary results are given, but it is pointed out that specific types of questions could be answered more easily by applying common elicitation procedures.

1. Introduction*

The two additive particles *also* and *too* are commonly regarded as synonyms with different diaphasic associations. *Also* is generally associated with a more formal style and written language, while *too* is taken to be characteristic of an informal style and spoken language. Moreover, there is a clear structural difference between the two particles: while *also* either takes up a 'medial' position in the clause, in terms of the structural description given by Quirk et al. (1985, 490-8), or immediately precedes the focus or 'added constituent' (cf. Reis & Rosengren 1997 for this term), *too* is usually clause-final. Consequently, *also* may either precede or follow the focus whereas *too* always follows it (cf. also Quirk et al. 1985, 609-10; Huddleston & Pullum 2002, 592-5). This contrast is directly related to another difference between the two particles, which concerns the placement of stress. In English as well as in many other European languages, additive particles are generally unstressed when they precede the added constituent while they attract stress when

* This paper was written during a visit at the Center for Grammar, Cognition and Typology at the University of Antwerp. Financial support from the Alexander-von-Humboldt Foundation and the University of Antwerp is gratefully acknowledged. Thanks are also due to the audience of the Workshop on 'The Scope and Limits of Corpus Linguistics' in Berlin (June 11, 2005) for valuable comments and helpful suggestions. Moreover, I am indebted to Stefan Gries for commenting on the paper and pointing out some shortcomings to me. All remaining mistakes are my own.

they follow it. Accordingly, (additive) *too* is invariable stressed, whereas *also* may be either stressed or unstressed, depending on its position relative to the focus. These structural and phonological differences are illustrated in (1) and (2). The added constituent/focus is enclosed by brackets bearing a subscript AC ($[\ldots]_{AC}$), and the main accent of a sentence is indicated by small capitals.

(1) a. (*Bill has bought a car, and…*)
[*Sue*]$_{AC}$ *has ALSO bought a car.* (*also* follows AC)
b. (*Sue has bought a bicycle, and…*)
she has also bought [*a CAR*]$_{AC}$. (*also* precedes AC)

(2) a. (*Sue has bought a bicycle, and…*)
she has bought [*a car*]$_{AC}$, *TOO.* (*too* follows AC)
b. (*Bill has bought a car, and…*)
[*Sue*]$_{AC}$ *has bought a car, TOO.* (*too* follows AC)
c. (*Sue has bought a bicycle, and…*)
**she has too bought* [*a CAR*]$_{AC}$. (*too* may not precede AC)

In addition to the 'canonical' configurations illustrated above there are alternative, less widespread uses of the two particles under discussion as well. First, *also* may be used sentence-initially, as a near equivalent of *moreover*. In such occurrences of *also* the entire sentence can be assumed to represent added information (cf. [3a]). This use is often considered to be restricted to, or at least characteristic of, spoken language. Moreover, *also* is sometimes used in a clause-final position, more or less like (additive) *too* (cf. [3b]). Note that *too* may also be right-adjacent to the added constituent (cf. [4a]) or occupy some intermediate position, as illustrated in (4b) (the examples are taken from Huddleston and Pullum 2002, 593).

(3) a. *Also,* [*it was pouring with rain*]$_{AC}$.
b. *We plan to visit* [*Paris*]$_{AC}$ *also.*

(4) a. [*I*]$_{AC}$ *too think that the proposal has a merit.*
b. [*I*]$_{AC}$ *realised too that he was in great pain.*

The central question to be addressed in this paper is: What are the major distributional differences between *also* and *too*? The following three (families of) hypotheses will be explored:

1. DIAPHASIC hypotheses: (i) *also* is characteristic of written language while *too* is typical of spoken language, and (ii) *also* is characteristic of a formal style while *too* is typical of an informal style.
2. STRUCTURAL hypotheses: the distribution of the two particles depends on structural properties of the added constituent, in particular its grammatical function and length.
3. An INFORMATION-STRUCTURAL hypothesis: the use of the two particles is sensitive to the distribution of added material over the sentence; added material must always be located to one side of the particle, which excludes the use of 'medial' *also* in specific cases.

These descriptive objectives are important in their own right, since in my view the distribution of *also* and *too* has not been accurately accounted for so far. However, there is a second, more general, objective pursued in this paper, namely to determine the scope and limits of corpus analyses. In particular, we will consider to what extent corpus methods are suited to an analysis of the hypotheses made above and what implications there are for methodological choices in this area.

We will start with a consideration of the diaphasic hypotheses in Section 2, then turn to the structural hypotheses in Section 3, and finally consider the information-structural hypothesis in Section 4. As will be seen, the structural hypotheses can straightforwardly be tested (and confirmed), whereas the diaphasic and the information-structural hypotheses pose methodological challenges of different types. As for the diaphasic hypotheses, a distinction between spoken and written language does not give any clear results, which is attributed to the heterogeneous nature of spoken language. The hypothesis that the contrast between formal and informal style is relevant seems to be more promising but is difficult to test using corpus evidence since it presupposes a well motivated classification of sub-corpora according to different degrees of 'formality', which is a non-trivial task of its own. Finally, the information-structural hypothesis is likewise shown to be difficult to test, since it predicts the complete absence of specific types of configurations. If those configurations are not found in a given corpus, this may be due either to a very low discourse frequency – a fact concerning the USE of language – or to a syntactic rule disallowing the relevant constructions, i.e. a fact about the language SYSTEM. The problem that comes to light at this point is that corpora do not give us any negative evidence. The paper concludes with a brief summary in Section 5.

2. The diaphasic hypotheses

The assumption that *also* and *too* are register-specific is a robust intuition shared by many native speakers of English. It has, however, rarely been tested empirically. Biber et al. (1999, 800), who distinguish between the four registers CONVERSATION, FICTION, NEWS and ACADEMIC PROSE, describe the situation as follows:

> In both expository registers [news and academic prose], the common additive adverbial *also* serves to mark information being added to previous information ... In fiction, the meaning of addition is spread more evenly over two adverbs, *also* and *too*, with *also* carrying a more formal tone: ... *Too* is used more informally, often in dialog or reports of dialog: ... Interestingly, this use of *too* is actually more common in fictional dialog than in conversation. (Biber et al. 1999, 800)

As this passage shows, the 'diaphasic hypotheses' have a number of ramifications that go beyond dichotomies like spoken vs. written language or formal vs.

informal style. In particular, Biber et al. (1999) mention a difference between fictional dialog and conversation. Let us nevertheless start with the higher-level distinctions between spoken and written language, and between formal and informal style, considering to what extent they can account for the distribution of *also* and *too*. Relatively strong versions of the relevant hypotheses are given in (5):

(5) a. M_1: In written language, *also* is more frequent than *too*.
M_2: In spoken language, *too* is more frequent than *also*.
b. S_1: In formal style, *also* is more frequent than *too*.
S_2: In informal style, *too* is more frequent than *also*.

Let us first consider hypotheses M_1 and M_2 ('M' stands for 'medium, 'S' for 'style'). The figures given by Biber et al. (1999, 796) appear to confirm both hypotheses. For the sake of the argument, we will assume that the four registers CONVERSATION, FICTION, NEWS and ACADEMIC PROSE systematically differ in the portion of spoken language, with a decrease from left to right (note that this assumption is of course itself in need of empirical justification, and that it is not made by Biber et al. 1999 themselves). In order to compare the relative frequencies of *also* and *too*, we can simply determine the ratio of the relative frequency of *also* to that of *too*. Let us refer to this ratio as the 'A/T-index' ('*also*-to-*too* index'). An A/T-index of '1' indicates that the relative frequencies of *also* and *too* are identical. If the A/T-index is above zero, *also* is relatively more frequent than *too*. Since Biber et al. (1999) provide only approximate numbers, the figures in (6) are generously rounded up. They represent (approximately) the distribution of *also* and *too* in the 'Longman Spoken and Written English Corpus' (~40m, 1990's). The subscripts on '<' and '>' indicate factors. For instance, '$<_3$' reads 'is three times smaller than'.

(6)

	CONV		FICT		NEWS		ACAD
A/T-index (LSWE)	~0.33	$<_3$	~1	$<_6$	~6	$<_{>1.5}$	>9

As can be seen quite clearly, the A/T-index increases monotonically from left to right, thus correlating negatively with the (assumed) portion of oral discourse in the relevant registers. However, the picture becomes much less clear if we compare the data of Biber et al. (1999) to those presented in an earlier corpus study on *also* and *too*, which was carried out by Fjelkestam-Nilsson (1983). Fjelkestam-Nilsson (1983) uses three corpora: the *London-Lund Corpus* (LLC) for spoken British English (~0.5m, 1960s-70s), the *London-Oslo-Bergen Corpus* (LOB) for written British English (~1m, 1961), and the *Brown University Corpus* (Brown) for written American English (~1m, 1961). According to her data, *also* is considerably more frequent than *too* in the LLC, i.e., in spoken British English. We can determine an A/T-index of 1.9 on the basis of her data – 1.47 for spontaneous conversation (texts I – III), and 2.19 for other types of oral discourse (texts IV – XII; cf. Fjelkestam-Nilsson 1983, 26). *Too* on the other hand is more frequent in the imaginative part of the LOB-corpus (texts K-R; A/T-index 0.68). In informative texts (A-J), *also* outnumbers *too* by far (A/T-

index 4.5). In sharp contrast to the results arrived at by Biber et al. (1999), the A/T-indices determined on the basis of Fjelkestam-Nilsson's (1983) data do not correlate with the (assumed) portion of oral discourse in the relevant registers. Moreover, unlike in Biber et al.'s (1999) data, the A/T-index of (spontaneous) conversation (LLC A-C) is higher than 1 (1.47), which means that *also* is more frequent than *too*. The A/T-indices based on Fjelkestam-Nilsson (1983, 26) are summarized in (7).

(7)

	LLC (CONV)		LOB (IMAG)		LOB (INFO)
A/T-index	1.47	$>_{1.7}$	0.68	$<_{6.4}$	4.5

The question arises as to how the marked discrepancies between the data shown in (6) and (7) can be explained. In order to obtain an additional set of reference values, I carried out a corpus study myself, using the BNC as a data source. The relative frequency of *also* in four sub-corpora of the BNC (corresponding to the registers of Biber et al. 1999) could easily be determined. For *too*, the relative frequency had to be estimated, since *too* is homophonous with an adverb of degree (e.g. *too small*). This estimation was carried out on the basis of a sample of 100 randomly chosen occurrences of *too* for each of the four registers under discussion (i.e., 400 tokens altogether). The results arrived at in this investigation are compared to the data presented above in (8). The most noteworthy fact is the extremely high A/T-index for conversation (4.23). The registers NEWS and ACADEMIC PROSE are here represented by their mean value under INFO, which is done for the sake of comparison.

(8)

A/T-indices	CONV		FICT/IMAG		INFO
Biber et al. (1999)	~0.33	$<_{3.0}$	~1	$<_{>7.5}$	>7.5
Fjelkestam-Nilsson (1983)	1.47	$>_{2.2}$	0.68	$<_{6.6}$	4.5
present study (BNC)	4.23	$>_{5.3}$	0.80	$<_{12.3}$	9.87

While the scores for FICT/IMAG and INFO can be considered more or less homogeneous (the variation coefficient *V* of FICT/IMAG is 20%, that of INFO ~37%; $V = \mu/\sigma$), the A/T-indices for conversation diverge much more strongly ($V_{CONV} \cong 100\%$, which means that the mean value and the standard deviation are approximately identical). Why should that be so?

As the comparative (British and American) data given by Fjelkestam-Nilsson (1983) show, the marked contrasts between the three data sets cannot be explained in terms of diatopic differences. The great discrepancies in the data are probably due to the fact that the categories 'conversation' and, more generally, 'spoken language', subsume an extremely heterogeneous set of discourse types, which are represented differently in the various corpora. The heterogeneous nature of spoken language, and hence, its limited utility as a criterion of register classification, has repeatedly been noticed in the relevant literature. For instance, Kim & Biber (1994), referring to data from Koran, Tuvaluan, Somali and English, observe that "no dimension defines an absolute dichotomy between spoken and written registers; rather, there is some overlap on each dimension, as

certain spoken registers have relatively literate character, and certain written registers have relatively oral characteristics" (Kim & Biber 1994, 179). It seems, thus, that cover terms like 'conversation' or 'spoken language' are too general to allow for any reliable distributional generalizations.

What needs to be done, thus, is to use a sub-classification of oral register types. Since a comprehensive register study is beyond the scope of this paper, we will only consider three registers (of the BNC) to illustrate what a relevant investigation could look like, and to demonstrate the marked asymmetries within the category of 'spoken language': (i) 'courtroom conversation' (~0.13m), (ii) 'broadcast discussion' (~0.76m), and (iii) 'medical consult' (~0.14m). The following A/T-indices were obtained using the VIEW-interface for the BNC:

(9)

	COURTR CONV		BROADC DISC		MEDIC CONS
A/T-index	34	$>_{7.2}$	4.7	$>_{3.0}$	1.59

As the data in (9) show, there are in fact marked contrasts between the A/T-indices of different types of spoken discourse. The highest score given in (9) – the one for 'courtroom conversation' – is way above the average value for INFO (~7). The 'medical consult' register is the one with the lowest A/T-index, and the score for 'broadcast discussion' corresponds more or less to the average in spoken language as determined in the study carried out on the BNC (cf. [8] above; note that the other studies lead to much lower A/T-indices).

The data given in (9) seem to suggest that the distributional differences between *also* and *too* may be related to the distinction between formal and informal style: from 'courtroom conversation' to 'medical consult' there is a felt cline in the degree of 'formality'. This brings us to the second pair of hypotheses made above (S_1 and S_2), which present us with a new challenge: How can these hypotheses be tested using corpus data?

The most obvious problem that immediately becomes apparent when approaching hypotheses S_1 and S_2 is: What does it mean for a register to be 'formal'? How can we determine degrees of formality, or establish a binary distinction between 'formal' and 'informal' language? These questions can be approached from two different directions. First, we can interpret the term 'formal' in a 'functional' manner, i.e. as a term referring to extra-linguistic discourse features like: the social setting or circumstances under which the discourse takes place, the social distance between the speaker and the addressee, the communicative purpose of the conversation, etc. Such an approach would require an elaborated sociolinguistic classification of discourse types, for instance along the lines of the work done by M.A.K. Halliday ('field', 'tenor' and 'mode' of discourse; cf. Halliday 1985, 12). Carried out in a methodologically sound way, we may hope for a relevant analysis to corroborate the intuition that the results given in (9) above reflect a difference in 'formality'. Since I am not aware of any detailed and empirically well-motivated sociolinguistic typology of natural language (sub-)corpora in terms of formality, however, and given that

establishing such a typology would be a research programme of its own, I will not pursue this line of reasoning any further.

The second way of approaching the distinction between formal and informal language could be based on linguistic features, in a multi-dimensional approach along the lines of Biber (1988, 1995). We could try to isolate a set of linguistic features that is, in some way, characteristic of either formal or informal language, and we could determine the extent to which these features tend to co-occur with *also* and *too*. Obviously, the classification of linguistic features as either 'formal' or 'informal' would, again, have to be motivated extra-linguistically, but such a classification seems to be more straightforward for single linguistic items than for entire register types. Moreover, the very fact that specific linguistic features tend to co-occur would be an interesting finding in itself, even without recursion to the sociolinguistic or pragmatic notion of 'formality'.

In what follows, I will only present the outlines of a co-occurrence analysis for *also* and *too*. We will consider three linguistic features, which are taken to be indicative of an either formal or informal style: (i) (the presence or absence of) contracted negation (*-n't*), which is interpreted as an indicator of little social distance between the interlocutors; (ii) nominalizations in *-ment*, which are considered to be indicators of abstract topics; and (iii) the use of the conjunction *however*, which can be regarded as an indicator of an explicit and dialectic discourse organization. In order to capture the amount of contracted negation, the ratio of non-contracted (*not*) to contracted negation (*n't*) has been determined for each of the text types mentioned in (9) ('N/T-index'). For nominalizations in *-ment* and the occurrence of *however*, the relative frequencies in the relevant sub-corpora of the BNC were obtained, using the VIEW-interface. The results are given in (10):

(10)

	COURTR CONV		BROADC DISC		MED CONSULT
N/T-index	1.05	$<_{1.5}$	1.62	$<_{1.4}$	2.30
-ment	3.36	$>_{1.3}$	2.62	$>_{1.7}$	1.58
however	2.12	$>_{2.1}$	1.03	$>_{6.9}$	0.15

Each of the three features shows monotonic curves from 'courtroom conversation' to 'medical consult', which seems to confirm the impression that the three registers can be ordered on a scale of 'formality'. Given that the A/T-index also correlates with the three registers shown in (10) (cf. [9] above), we can in fact – still tentatively – conclude that there is a (positive) correlation between the 'degree of formality' of a given register, and the A/T-index of that register.

In order to obtain more reliable results, we would of course have to consider more linguistic features and, ideally, carry out a large-scale multi-factorial analysis à la Biber, which I leave as a suggestion for future research. Moreover, we should be careful not to overgeneralize from the data in (10). For example, it is sometimes assumed that discourse connectors are *per se* indicative of a formal style; but if we consider the distribution of the subordinator *because* relative to the three registers mentioned above, it turns out that *because* is more frequent in broadcast discussion

(24.44) than in the two other registers (consult 20.72 and courtroom 20.71). This shows, once again, that establishing a linguistically well-motivated classification of sub-corpora is a non-trivial task of its own.

3. The structural hypotheses

We can now turn to the second family of hypotheses under consideration, i.e. the ones concerning structural properties of the added constituent. The following two hypotheses were explored by Fjelkestam-Nilsson (1983):

1. *Also* and *too* are sensitive to the grammatical function of the added constituent.
2. *Also* and *too* are sensitive to the length of the added constituent.

Fjelkestam-Nilsson (1983) distinguishes between three types of grammatical functions: (i) subjects, (ii) predicates, and (iii) 'other constituents', i.e. (prepositional or non-prepositional) objects and adjuncts. All three types of constituents combine more frequently with *also* than with *too*. However, if we determine the A/T-index for the different types of added constituents, a marked contrast can be observed. As is shown in (11), the A/T-index is highest with predicative AC's and lowest with subject AC's (cf. Fjelkestam-Nilsson 1983, 52; the figures are based on the LOB):

(11)

	SUBJ		OTH		PRED
A/T-index	1.56	$<_{2,3}$	3.59	$<_{1,2}$	4.46

Why should *too* be more prone to combine with subjects than with other constituents, in comparison to *also*? There are several possible explanations. One such explanation could be based on a tendency to maintain a certain distance between the added constituent and the additive particle if the latter is stressed. Added constituent often receive a secondary stress when the primary stress falls on the additive particle. If the additive particle immediately follows the added constituent, two stress positions come to stand side by side, which is sub-optimal from the perspective of prosodic sentence organization. This fact is illustrated in (12), where primary and secondary stress are indicated by the common phonetic symbols ['] and [ˌ], respectively.

(12) a. [ˌ*Mary*]$_{AC}$ *kissed John*, ˈ*too*. (ok)
b. Mary kissed [ˌ*John*]$_{AC}$, ˈ*too*. (sub-optimal)

The correlation illustrated in (11) may also be related to the second factor investigated by Fjelkestam-Nilsson (1983), viz. the length of the added constituent. Fjelkestam-Nilsson (1983) distinguishes between added constituents of one, two and three or more words (3+). Her study gives particularly clear results showing that the A/T-index correlates positively with the length of the added constituent. (13) summarizes the results for subject AC's and for other AC's:

(13)

	1 word		2 words		3+ words
subj. AC's	0.47	$<_{3.0}$	1.42	$<_{7.3}$	10.38
other AC'S	0.64	$<_{3.1}$	2.00	$<_{7.2}$	14.38

There may be a direct relationship between the data given in (11) and the data in (13) insofar as the grammatical function of a constituent correlates with its length. As is well known, subjects tend to represent given information and are often pronominal, whereas objects and adjuncts are more often than not used to introduce new information (a point already made and tested by DuBois 1987). It is therefore conceivable that the distributional contrast between added constituents with different grammatical functions is at least partly a consequence of the length of the relevant constituents. But why should additive particles be sensitive to the length of a constituent? Consider the contrast between the example in (14), which has been taken from the BNC, and the one in (15), where *too* is used instead of *also*:

(14) *It is almost certain that targets will be set, including efforts to reduce by a third the number of smokers by the year 2000. There will* also *be* [*targets aimed at reducing the incidence of strokes, heart disease and preventable cancers*]$_{AC}$.

(15) *It is almost certain that targets will be set, including efforts to reduce by a third the number of smokers by the year 2000. There will be targets* [*aimed at reducing the incidence of strokes, heart disease and preventable cancers*]$_{AC}$, too.

Intuitively, the additive particle in (15) 'comes too late' (there also seems to be a possible ambiguity since *also* may be interpreted as being associated only with *preventable cancers*, which it would then add to *strokes* and *heart diseases)*. What (15) illustrates is that additive particles should not be too remote from the 'onset' of the added constituent. This effect could probably be explained in psycholinguistic terms, for example, with reference to a theory of parsing like the one proposed by Hawkins (1994). Since space prevents us from considering this theory in any detail, I will only summarize some central points (cf. also Kirby 1997 for discussion). In a few words, Hawkins' theory says that language tends to be arranged in such a way that the higher level constituent structure is recognized as early as possible. Assuming that *too* forms a constituent with the added constituent, the length of the 'constituent recognition domain' – the part of a sentence that needs to be processed in order to identify the major constituents – increases in direct proportion to the complexity of the added constituent; and the longer the CRD is, the more difficult is a sentence to process, according to Hawkins' (1994) theory. For instance, in a structure like [...[[ω_1 ω_2 ω_3]$_{AC}$ *too*]$_{C1}$], the 'constituent recognition domain' for C_1 comprises [[ω_1 ω_2 ω_3]$_{AC}$ *too*]. *Also* on the other hand has a weaker impact on the structural complexity of a sentence (if it precedes the added constituent) because the higher-level constituent structure can be determined after *also* and the head of the added constituent have been processed. For instance, in a structure of the form [... *also* [[ω_1 ω_2 ω_3]$_{AC}$], the 'constituent recognition domain' only comprises *also* and ω_1.

4. The information-structural hypothesis

Unlike the first two (families of) hypotheses considered above, the one to be tested in this section has, to my knowledge, not been explored anywhere else, or is at least not generally taken for granted. I will aim to show that *also* and *too* behave differently with respect to the information structure of their host sentences. This difference will be related to the structural contrast between *also* and *too* pointed out in Section 1 above. Consider the conversations in (16) and (17):

(16) Jane: *I love you*
Tarzan: a. *I love you, too.*
b. # *I also love you.*

(17) Jane: *Why are you so unhappy?*
Tarzan: a. *My house has burnt down, and my wife has left me, too.*
b. # *My house has burnt down, and my wife has also left me.*

The b-sentences in both (16) and (17) are semantically deviant when pronounced with an unmarked intonation contour and only one stress position. In (16b), using *also* would only be appropriate if Jane had said *Jack loves me*. (17b) would only be acceptable if the house's burning down were conceived of as a 'leave', i.e., if the house were personified (*My house has left me and my wife has left me, too*). I would like to make the following hypothesis to account for the data in (16) and (17):

(18) The information-structural hypothesis:
Additive particles can occur only to one side of the added material.

If we consider the information structure of the sentences at issue, the hypothesis in (18) can explain why (clause-final) *too*, but not (medial) *also*, is possible in all cases. Let us start with the contrast illustrated in (16). This sentence can be regarded as instantiating a 'contrastive topic' structure, i.e. a structure in which not only the focal, but also the topical, part of a sentence is contrasted with alternative values from the context (cf. Büring 1997). There is a contrast between two subjects – *Tarzan* and *Jane* – and, at the same time, between two objects – *Jane* and *Tarzan*. This is illustrated in (19). Accordingly, both the subject and the object represent added information (cf. [20]):

(19) Jane: I_{Jane} *love* you_{Tarzan}.
Tarzan: (and) I_{Tarzan} *love* you_{Jane} (*too*).

(20) a. $[I]_{AC}$ *love* $[you]_{AC}$ *TOO.*
b. # $[I]_{AC}$ *ALSO love* $[you]_{AC}$.

In contrastive topic constructions only *too* (or final *also*) is possible, since there are added constituents to both sides of the particle if (medial) *also* is used (cf. [20b]). The same point can be made with regard to the example in (17) above. Since the

whole clause qualifies as added material, the additive particle has to be clause-final. The information structure of (17) is given in (21a), and the corresponding (deviant) sentence with (medial) *also* is given in (21b):

(21) a. *My house has burnt down, and* [*my wife has left me*]$_{AC}$, *TOO*.
b. *#My house has burnt down, and* [*my wife has*]$_{AC}$ *ALSO* [*left me*]$_{AC}$.

While the two examples given above seem to corroborate the hypothesis made in (18), it would of course be better to have more solid empirical evidence lending support to that hypothesis. How could such a hypothesis be tested using corpus data?

Given that the information-structural hypothesis makes reference to properties of sentences that can only be determined in relation to contextual information, some of which is rather specific, it is hard to imagine how a computerized procedure could be applied to test it. Even using a corpus with an excellent phonological annotation would not facilitate matters, since the mapping from information structure to sentence prosody is not one-to-one, especially in sentences with additive particles. The only option to test this hypothesis, it seems to me, is to do it manually: we have to try and find examples of medial *also* with added information to both its left and its right, in a sample of sufficient size.

However, even then a non-trivial problem remains which concerns a very general disadvantage of natural language corpora: they do not provide any negative evidence. Let us assume that we will not find a single occurrence of *also* contradicting the information-structural hypothesis. What could we actually conclude from this? Would this entitle us to claim that there is a *grammatical rule* disallowing the relevant configuration? Or are such configurations simply too rare to be attested in our sample? And here is a second, equally obvious, problem: what does it mean for a sample to be 'of sufficient size' to be indicative of a syntactic rule, rather than extreme rarity in discourse?[1]

---

[1] Stefan Gries has pointed out to me that there are statistical methods to deal with these questions, and that the no-negative-evidence problem "is actually not a big issue anymore". For instance, he suggests comparing the observed frequencies to the expected ones on the basis of a Poisson distribution. As far as I can see, however, statistical methods cannot give us any information about whether a given 'event' is rare or absent because it is at variance with the lexico-grammatical system of a language, or because speakers simply do not make use of an option which is, however, available. To illustrate this with an example from a different area of grammar, in earlier work (Gast forthcoming) I investigated distributional properties of 'intensifiers', i.e. *self*-forms in an adjunct function (*the president himself*). Intensifiers may occur in different positions of the clause, sometimes at a distance from the NP they are associated with (e.g. *He did it himself, He snores himself*). Among such 'head-distant' uses of intensifiers, two semantic types can be distinguished, namely an 'exclusive' one (~'alone', *He did it himself*) and an inclusive one (~'too', *He snores himself*). According to their semantics, the two types of intensifiers should be able to co-occur. For instance, the following context is conceivable: – *John has complained that Bill hasn't done his homework himself. – Well, John has himself not done his homework himself*. Such sentences with two occurrences of intensifiers are unattested in the major corpora of English. Is this due to a grammatical rule or to an extremely low discourse frequency?

On the other hand, let us assume that we *will* find one example that clearly contradicts our hypothesis. Would that be a good piece of evidence showing that the hypothesis is false? Or could it be possible that the corpus contains material which is not in accordance with the common rules of English grammar, i.e., so-called 'performance errors' – constructions which the speakers themselves would agree are ungrammatical?

Let us finally turn to what I actually found. I examined each of the 332 occurrences of *also* in the (augmented) LLC and assigned information-structural interpretations to the relevant sentences, which was quite straightforward in most but not all cases. The hypothesis held true for all but one occurrence of *also*. The example in question is given in (22) with some contextual information:

(22) [*Eight musicians are talking about the organisation of their rehearsals; two of them – speaker* a, *identified as Gill/ian*, *and speaker* b *– had a conversation about the topic the previous day;* b *is speaking*]
… yes, if it – if it's a successful compromise, that's fine. But I mean – compromises by their nature normally aren't – [*əm*] *– but I think it's probably –* [*əː*] *– I still think practically – it's probably the best thing to do.*
Whât – [ə] – whât – ˌGill and ˌI were ˈâlso disˌcussing yèsterday
– I think it's quite – important that if we do a sort of bunch of rehearsals based on Tuesday, say – I think there are – the schedule for the next one – is too far strung out for my liking […]

The information structure of (the critical part of) (22) can be represented as in (23) below. The subject *Gill and I* contrasts with the (contextually given) entire set of (eight) musicians; moreover, the adverb *yesterday* represents new information as well insofar as it contrasts with the (implied) adverb *today*:

(23) [*Gill and speaker*]$_{AC}$ *were* (*also*) *discussing* [*yesterday*]$_{AC}$.
↕ ↕
[*All eight musicians*]$_{AC}$ *are discussing* [*today*]$_{AC}$.

Given the existence of (22) in the LLC, should we discard our information-structural hypothesis? I believe that this would be premature. First, the assignment of information-structural specifications to the syntactic structure of (22) as given in (23) is not the only possible one. In particular, one may argue that *Gill and the speaker* do not properly contrast with *all eight musicians* since they are contained in that set of musicians. Second, we may interpret the scope relations holding between *also* and *yesterday* differently, insofar as *yesterday* could be regarded as being outside the scope of *also*. In that case, (22) would have more or less the following meaning (the scope of *also* is indicated by parentheses bearing a subscript SC):

(24) *Yesterday,* ([*Gill and I*]$_{AC}$ *were also discussing that*)$_{SC}$: …

Given such alternative information-structural interpretations, it seems to me that discarding our hypothesis would be premature. What is more, I doubt that corpus methods are adequate at all to test such a hypothesis. As has been seen, the information-structural interpretation of a syntactico-phonological input is not always straightforward. Moreover, we would have to find a non-arbitrary way of distinguishing 'ungrammaticality' from 'discourse rarity'. Given such difficulties, we seem to have a problem here indicating that there are limits to the application of corpus-based methods in certain areas of grammar. It is my impression that an empirical investigation using questionnaires, interviews and elicitation procedures – for instance, along the lines of Greenbaum and Quirk (1970) – would be a more promising way of finding answers to some of the questions addressed in this paper, since such procedures allows us to determine not only what is possible, but also what is impossible.

5. Summary

In this paper a number of hypotheses concerning the distribution of *also* and *too* were tested by using corpus data, with varying degrees of success. While strong support could be given to the hypothesis that structural properties of the added constituent play a role, the diaphasic and information-structural hypotheses proved more difficult to test. Certain indications were found showing that the distinction between a formal and an informal style are relevant, but no conclusive evidence could be offered. The hypothesis that *also* is characteristic of written language while *too* is typical of spoken language was not supported by the data. Finally, the distribution of added material over the sentence was hypothesized to be an important factor, which proved to be true to the extent that no clear counter-example could be found in the LLC. However, it was pointed out that corpus methods may not be fully adequate to test such a hypothesis.

Works Cited

Biber, Douglas (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison.* New York: Cambridge University Press.

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

Büring, Daniel (1997). *The Meaning of Topic and Focus – The 59th Street Bridge Accent*. London & New York: Routledge.

DuBois, John (1987). "The discourse basis of ergativity." *Language* 63, 805-55.

Fjelkestam-Nilsson, Brita (1983). *"Also" and "Too" – A Corpus-Based Study of their Frequency and Use in Modern English.* Stockholm: Almqvist & Wiksell.

Gast, V. (forthcoming). *The Grammar of Identity – Intensifiers and Reflexives in Germanic Languages*. London: Routledge.
Greenbaum, Sidney and Randolph Quirk (1970). *Elicitation Experiments in English. Linguistic Studies in Use and Attitude.* London: Longman.
Halliday Michael A.K. (1985). *Spoken and Written Language*. Oxford: Oxford University Press.
Hawkins, John (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
Huddleston, Rodney & Geoffrey Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Kim, Jong-Jin and Douglas Biber (1994). "A corpus-based analysis of register variation in Korean." D. Biber & E. Finegan, ed. *Sociolinguistic Perspectives on Register*, 157-81. New York: Oxford University Press.
Kirby, S. (1997). *Function, Selection and Innateness – The Emergence of Language Universals*. Oxford: Oxford University Press.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
Reis, Marga & Inger Rosengren (1997). "A modular approach to the grammar of additive particles: the case of German *auch*." *Journal of Semantics* 14, 237-309.