*Jürgen Esser**

Medium-transferability and corpora: Remarks from the consumer-end of corpus linguistics

Abstract

A distinction is made between units and categories that are medium-independent (e.g. word class, noun phrase and clause) and those that are tied to the medium of realization. While the **orthographic sentence** is a typical, highly conventionalised unit that is tied to the written medium, the **tone unit** is a typical unit of the spoken medium. There are, however, some problems related to this unit of realisation. Not only is the tone unit and its organisation into higher-level units subject to theoretical dispute, it also has a different status in speaking and reading respectively, which so far has been largely ignored in corpus linguistics.

1. Consumers of corpus linguistics

To my mind the image of supply and demand from the area of economics can be well applied to corpus linguistics. On the one hand, there are the designers, compilers and analysts of corpora, and on the other hand there are the linguists who have no corpora or tagging programs of their own but who want to use corpora to assist their own research. It is the latter that I would like to call consumers of corpus linguistics, evidently a substantial target group invited to buy and use the many corpora and tools that are being made available.

2. Medium-independent units, categories and structures

There is an important point that the consumer of corpus linguistics must be aware of: The bytes of the ASCII-code which represent the corpora in electronic form do not all have the same status as linguistic data. There are units, categories and structures that are independent of the

* *Jürgen Esser*
Institut für Anglistik der RWTH
Karmanstraße 17/19
52062 Aachen (D)

medium of realisation and those that are dependent on it. This distinction was already made by Halliday/McIntosh/Strevens (1964: 51) and I think that it can be a useful consideration for corpus linguistics:

Table 1

Without discussing some disputable details of Table 1, it is fair to point out that grammatical word-forms (which separate homographs and homophones), word-class labels and the structures of phrases and clauses are medium-independent. These units are manifestations of de Saussure's (1916) and Halliday et al.'s (1964) abstract concept of 'form' (as opposed to 'substance') and they demonstrate Lyons' (1981) concept of 'medium-transferability'. Lyons uses this notion not only in its trivial sense, i.e. everything that is spoken can be written and everything that is written can be read aloud. Rather, it indicates for him (p. 60): "not only that a language-system has a structure, but that it is a structure". But, as Halliday's distinctions make clear, in linguistic description we must reckon not only with medium-independent units but also with units that depend on the medium of realisation.

3. Medium-dependent choice of medium-independent units

Before I come to medium-dependent language units I must mention the medium-dependent choice of medium-independent units. This choice

makes for the distinction between the styles of the spoken and the written language and it is usually related to the medium in which a language activity originates as the left-hand part of Table 2 makes clear.

Table 2

Basically, the stylistic choice between spoken and written English can be described in terms of *elements* and *configurations.* Elements are directly searchable in ASCII-code, separately or in combination such as first-person pronouns, past-tense forms, that-clauses or *by*-passives. Biber's (1988) feature study, for example, shows how medium-independent elements are correlated with situational variables of the communication situation. On the other hand, choices can be described in terms of configurations, notably in terms of complex sentences. So far, there are only few studies which deal with configurations of medium-independent elements in larger structures because the parsing of real complex sentences still offers some difficulties.

An interesting study in this direction is Altenberg's (1993) article on recurrent verb-complement constructions in the London-Lund corpus. He deals, for example, with SVC constructions that form the matrix clause of an extraposed subject and that function as "attitudinal prefaces", for example:

(1) it's (very/rather/a bit/so) difficult (to)

Further observations show, quite expectedly, that extraposition in the spoken London-Lund Corpus tends to occur in sentences that are less complex compared to sentences with extraposition in the written Learned-Scientific part J of the LOB-Corpus. Compare (2) and (3) as typical examples. [Notation convention: Each clause starts a new line with a new code. There is a number starting from 1 for each new clause, the main clause is underlined. Nominal clauses keep the number of their matrix code and receive a subscript for their respective type: s for subject clause, d for direct object clause. Postmodifying clauses are marked additionally "*" and ing-clauses (gerunds) "Δ ".]

(2)	*1*	*it really <u>is</u> a tremendous problem* \	
	2	*because [@:m] it's very difficult*	
	2s	*to* adapt *yourself* \	
		to another human being \	*(S-5-10)*

(3)	*1*	*After Dr. Aukrust's careful analysis of the Norwegian*	
		figures, and the extensive figures for other countries	
	*2**	*quoted above,*	
	1	*it is going to be very difficult for anyone*	
	1s	*seriously to contend*	
	(1s)d	*that increased investment is a sure way of*	
	3Δ	*increasing the rate of economic growth.*	*(J-42)*

The tree banks of parsed corpora, like the Lancaster Parsed Corpus, will help to make it easier to study such medium-dependent configurations of medium-independent units by way of comparison.

4. Medium-dependent language units, categories and structures

I now come to the four language activities mentioned in the right-hand part of Table 2. For writing there is a high degree of conventionalisation for units, categories and structures that are medium-dependent. They include the orthographic word, the orthographic sentence (with a capital letter at the beginning and a special punctuation mark at the end) and the paragraph. These conventions were also used in the past for the transcription of spoken material, as for example by Gregory and Carroll (1978: 39):

(4) *A: Going to buy one?*

B: Don't know. Perhaps.

Today, one tends to use tone units in the written representation of spoken material instead of orthographic conventions.

On the other hand, there is less standardisation for the spoken medium, both for speaking and reading. For the spoken medium the units, categories and structures are represented by quite a number of different models. There seems to be agreement about the central role of the tone unit and that it has a nucleus. But opinions differ about the number and types of tones, about the status of prominent syllables other than the focus, and about pitch levels or key. In any case, the medium-dependent intonation elements in the spoken corpora are much more subject to theoretical dispute than medium-independent categories like noun, article, past-tense form of the verb etc.

Medium-dependent language units of a given theoretical model can again be studied as elements and in configurations. There exist, for example, statistical studies of prosodic elements in the London-Lund Corpus by Altenberg (1987) and Nevalainen (1992). One result of Nevalainen's study is the following (p. 419 f.):

> *"The falling type [of tone] predominates in personal face-to-face conversations between equals and intimates [...]. As the social or physical distance increases, as in telephone conversations and broadcasts, the rising type will gain ground."*

This study of elements is comparable to Zettersten's (1969: 2) finding that the letter h is more frequent in the Fiction genres K-P of the Brown Corpus than in the other genres, due to the frequent occurrence of the pronouns *he, his, him, she* and *her.* But there is some limitation in the exploitation of elements. So, one of the conclusions that can be drawn from Nevalainen's investigation is that the study of configurations of intonation elements should be further developed.

5. Studying larger intonation structures in corpora

Studying larger intonation structures in corpora is like studying complex sentences. In both cases we are dealing with pragmatic configurations of higher-level structures and not only with elements. Just as there can be no list of all possible complex sentences in English, there can be no list of all possible larger intonation structures. Nevertheless we are trying to establish some recurring patterns with the help of suitable intonation models. But here, as a consumer of corpus linguistics, I find

myself in difficulties: I do not want to be restricted to the prosodic model of Crystal 1969, which concentrates on elements, but I want to explore the corpora in the light of new or alternative theories that accommodate configurations of intonation elements.

Basically, there are two directions in the study of larger intonation structures: the declining tonal envelope and the relation of adjacent tone units.

The declining tonal envelope is responsible for intonational paragraphs, called for example, "paratone" by Couper-Kuhlen (1986: 1989) or "pitch sequence" by Brazil et al. (1980: 61). The pitch sequence usually begins with high key. Formally it

> *"begins immediately following a tone unit with low termination and includes all succeeding tone units until the next one with low termination."*

While the booster signals in the London-Lund Corpus can be readily interpreted as high key there is unfortunately no indication for low key in Crystal's system and hence in the corpus.

Therefore, the consumer of spoken corpora should perhaps first turn in the other direction in the study of larger intonation structures: the relation of adjacent tone units. This is the old idea, already to be found in Palmer (1922: 88), that successive tone units with identical intonation elements express coordination or communicative equivalence, whereas successive tone units with different intonation elements express subordination or superordination, cf. Fox (1984) and House (1990). It is assumed, for example, that falls signal more relevance than rises and that high key signals more relevance than normal key, and normal key more than low key. A cline of relevance, as explored in Esser (1988: 66), could look in part like (5). [The tone unit is represented in abstract form by an underline (for the nucleus) and by pitch direction: \ falling, / rising. Nuclear high key is marked by subscript H. The angled brackets point to subordinated material.]

(5) ___$_H$ \ > ___ \ > ___ /

Adapting an example from the London-Lund Corpus we get:

(6) *<u>immediately</u> before / [<] I was teaching in a <u>school</u>*$_H$ \
[>] in <u>Egypt</u> \ [>] but before <u>that</u> / [<] I was <u>in India</u>$_H$ \
(S-1-6)

According to the scale of relevance in (5), we can identify the second and the last tone unit in (6) as presentational peaks, marked by asterisks in abstract form:

(6+) ___ / < *___H^* \ > __ \ > ___ / < *___H^* \

Note that the scale of relevance makes it posible to recognise synonymous intonation patterns. In the following examples adapted from Altenberg (1987: 181) it is always the word difference that is presented as a peak:

(7) *this made no difference to this girl *
this made no difference \ to this girl /
*this made no difference$_H$ \ to this girl *

6. Phonic presentation structure of encoder (speaking)

With orally originating texts, the medium-dependent presentation structure is created by the speaker in the act of encoding. In this respect it differs radically from the decoding-encoding process of reading. As has been frequently observed, speaking intonation differs from reading intonation. One point is the predominance of falling tones. This does not mean that there are more neutral statements or commands (functions often associated with falls). Rather, the falls have to be seen as elements in larger intonation structures. They function perfectly well in sentence-medial position as we have seen in examples (2) and (6) where the afterthought-like presentation of *to another human being* in (2) and *in Egypt* in (6) are part of larger presentation structures that are typical of orally originating texts.

7. Phonic presentation structure of decoder-encoder (reading)

Reading, on the other hand, is a decoding-encoding process. The reader has to produce a medium-dependent presentation structure on the basis of a configuration of medium-independent units. Not only are there many possible readers for one text, even one reader can produce several configurations of intonation elements. Therefore, the status of the intonation symbols in reading corpora is different from that in spoken corpora.

The concepts of intonational synonyms and abstract presentation structure can help to find recurring patterns in this infinity of possibili-

ties. Here are two parallel versions from my own reading corpus which show that the same abstract presentation structure can be expressed by different intonation means, namely high key in (8) and falling tone after several rises in (9):

(8) *He was confident \ assured \ in a sports shirt \ and light cotton slacks \ and open-toed sandals \ like a tourist$_H$ *

___ \ = ___ \ = ___ \ = ___ \ = ___ \ < *___H* \

(9) *He was confident / assured / in a sports shirt / and light cotton slacks / and open-toed sandals / like a tourist *

___ / = ___ / = ___ / = ___ / = ___ / < ___ \

The presentation strucure in (9) exemplifies the principle of resolution which is believed to be a property of reading intonation. By contrast, the presentation structure in example (2) from the spoken London-Lund Corpus does not make use of the principle of resolution, nor does the presentation structure in (6) which is also from the London-Lund Corpus.

The corpus study of intonation must therefore reckon with different presentation structures for speaking and reading. The intonation of reading is not a unique property of the realisation in the phonic medium like the phonological structure of words. It is something that must be worked out as a pragmatic achievement. The linguistic description of this decoding-encoding process relies on the analysis of corpora into medium-independent complex sentences and medium-dependent intonational presentation structures.

References

Altenberg, Bengt (1987): *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*. Lund: Lund University Press.

Altenberg, Bengt (1993): Recurrent verb-complement constructions in the London-Lund Corpus. In: Jan Aarts, Pieter de Haan & Nelleke Oostdijk (eds): *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi. 227-245.

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

Brazil, David, Malcom Coulthard & Catherine Johns (1980): *Discourse Intonation and Language Teaching*. London: Longman.

Couper-Kuhlen, Elizabeth (1986): An Introduction to English Prosody. Tübingen: Niemeyer.

Crystal, David (1969): *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Esser, Jürgen (1988): *Comparing Reading and Speaking Intonation*. Amsterdam: Rodopi.

Fox, A. (1984): Subordinating and co-ordinating intonation structures in the articulation of discourse. In: D. Gibbon & H. Richter (eds): *Intonation, Accent and Rhythm*. Berlin: de Gruyter. 120-133.

Gregory, Michael & Susanne Carroll (1978): *Language and Situation: Language Varieties and their Social Contexts*. London: Routledge & Kegan Paul.

Halliday, M.A.K., Angus McIntosh & Peter Strevens (1964): *The Linguistic Sciences and Language Teaching*. London: Longmans.

House, Jill (1990): Intonation structures and pragmatic interpretation. In: Susan Ramsaran (ed): *Studies in the Pronunciation of English: A Commemorative Volume in Honour of A.C. Gimson*. London: Routledge. 38-57.

Lyons, John (1981): *Language and Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Nevalainen, Terttu (1992): Intonation and discourse type. *Text* 12, 397-437.

Palmer, Harold E. (1922): *English Intonation*. Cambridge: Heffer.

Saussure, Ferdinand de ([1916] 1974): *Cours de linguistique générale*. Critical edition by T. de Mauro. Paris: Payot.

Zettersten, Arne (1969): *A Statistical Study of the Graphic System of Present-Day American English*. Lund: Studentlitt.