# 10. CORPUS LINGUISTICS AND PROBABILISTIC GRAMMAR

## 10.1 Computational Linguistics and its Methods of Research

Findings gained from the analysed corpus material can be processed in an attempt at a more appropriate grammatical description of language use meeting the demands of language users. As Svartvik puts it "...it is maintained that corpus-studies will help to provide descriptively more adequate grammars" (1966.Preface).

The traditional approach to grammar rested on rules which were usually exemplified from written texts (fiction, newspaper reporting and scientific literature). In many cases, however, these examples were chosen at random. The lack of sources, especially with regard to spoken language, did not enable the linguist to apply a different, more satisfactory procedure.

With the rise of corpus linguistics, especially the existence of spoken language corpora such as *A Corpus of English Conversation* (Lund 1980), or the British National Corpus, possibilities of an extensive and more realistic grammar analysis and description of spoken and written language have opened up. The existence of large corpora has facilitated access to existing language use. The scope of the texts available in the corpora is sufficient for the delimitation of the register and genre variation in language use.

**Predictability** in the occurrence and use of grammatical structures typical of texts of specific types and different registers is an issue of great importance. Halliday (1991.31) expresses his concern in the following observation: "It had always seemed to me that the linguistic *system* was inherently probabilistic, and that frequency in text was the instantiation of probability in the grammar". Halliday suggests: "What is predicted is the general pattern" and, furthermore, "...its relevance (i.e. the relevance of predictability, the present author's note) lies not in predicting but in interpreting" (1991.32).

Halliday has coined the term **lexicogrammar** (1991.31) to denote the interface between the meaning and form of the message. The concept of lexicogrammar is compatible with my interpretation of the phenomenon of semantic indeterminacy and its manifestations. Indirectness, impersonality, attenuation and accentuation represent the interface between the grammatical structure and word meaning.

In order to characterize predictability of grammatical structure with regard to semantic indeterminacy, it is necessary to observe three basic aspects of grammatical phenomena occurring in the text, namely

(1) **the frequency of occurrence of indeterminacy phenomena**
(2) **the grammatical structure of indeterminacy**
(3) **grammaticalness and grammatical acceptability**

A close scrutiny of the analysed material reveals the presence and a high frequency of occurrence of indeterminacy phenomena. Quantitative analysis of the texts from the corpus is thus a precondition for further research. Up to the present, my investigation has excluded the use of exact methods of computational linguistics, because every single occurrence of semantic indeterminacy is heavily context-bound. At this stage, **semantic tagging**, i.e. the search for meaning in context, cannot be done by computer, because the identification of meaning can proceed with certainty only by close examination on the part of the researcher. It will, however, be possible in the future to carry out the search for key words, which can be computerized. Their function in the text would then be matched with the relevant context and individually classified.

It is also be possible to work with **concordance lists**, which again seem to have a limited scope of application. In my view, the crucial problem of concordance listing is that the language data can only include the **immediately relevant co-text**, i.e. the verbal context, not the broader context, i.e. the situational context, or the context of general experience. All types of context are relevant in the disambiguation of meaning.

## 10.2 Representativeness of Text Selection

In this section an attempt is made to match the findings related to semantic indeterminacy with methods used in the field of corpus linguistics. Linguists claim that **key word identification** and **concordance listing** should be based on a selection which is "representative". Let me examine the notion of **representativeness** in this particular field of computational linguistics. Within the framework of spoken discourse analysis the use of a subtle classification of text types is inevitable.

The classification of corpus texts used in this book is based on the criteria of **tenor, mode** and **domain.**

**The tenor** "has to do with the relationship between a speaker and the addressee(s) in a given situation, and is often characterised by greater or lesser formality" (Leech et al. 1982.9). In my research, I have attempted to analyse informal face-to-face conversation, formal and informal interviews, and telephone conversation representing small talk (administration).

**The level of formality** is the basic distinctive feature of any conversational behaviour. It is based on power relations and the distribution of knowledge which can be either symmetrical or asymmetrical. Attributes of discourse such as indirectness or politeness vary according to the level of formality present in discourse.

**The mode** "has to do with the effects of the medium in which the language is transmitted" (Leech et al. 1982.9). Face-to-face spoken language is accom-

panied by paralinguistic features which are usually substituted by language in telephone conversations or radio interviews. In my investigation I deal with three modes which differ in their linguistic realization.

**The domain** "has to do with how language varies according to the activity in which it plays a part" (Leech et al. 1982.9). In my investigation face-to-face conversation is part of everyday use while the interviews are political and the telephone conversation is business-like. As has been shown, the field of activity in which the particular conversation occurs strongly determines its lexico-grammar, i.e. the choice of vocabulary and the structural properties of the text.

## 10.3 Remarks on the Grammatical Structure of Conversational Language

The indeterminacy of utterance meaning does not exclusively depend on word meaning; it is also simultaneously produced by its grammatical structure.

Let me summarize the effect individual manifestations of semantic indeterminacy have on the grammatical structure of utterances.

**Indirectness** is frequently delivered by means of a declarative question. The results of my investigation show that in authentic English conversation the declarative sentence structure is preferred to the interrogative sentence structure. Degrees of indirectness can be achieved by specific combinations of various question markers in discourse. Clusters of indirect elicitations tend to appear in face-to-face conversation; they are less frequent in telephone conversation and radio interviews.

**Declarative sentence structure** is enhanced by **question markers** such as intonation and/or lexical markers, e.g. a question phrase of the type *I think, I believe, I suppose, I hope* etc., afterthoughts of the type *as far as I know, if possible*, prompters like *you know, you see* etc. The declarative sentence is a common way of expressing a **query** in authentic English conversation. The **polite request** is frequently worded as a declarative sentence, besides the generally accepted interrogative sentence structure of the type *Can you help me, please.*

**If-clauses** represent grammatical structures which are **hypothetical.** Due to their dubitative nature, they are frequently used for interrogative purposes as expressions of uncertainty and doubt, such as **the conversational formulae** *if necessary, if possible, if need be* etc. At the same time, if clauses can carry the meaning of an invitation in the sentence *listen if you feel like a film tomorrow night Mike* (S.1.7.1207-1208)

**Impersonality** is rendered by means of personal pronouns *we* and *they*, indefinite pronouns *one* and *people*, passive voice, nominal expressions of the type *the problem is* and the *there is* construction. Various **degrees of impersonality** can be achieved through a combination of these means.

Impersonality is linked with the pragmatic category of **detachment.** It can be used in face-to-face conversation as a depersonalized way of expression,

which does not occur very frequently. Additional motives for the use of impersonal structures are lack of knowledge, or the speaker's self-defence. Some conversation genres feature impersonality to a great extent, e.g. radio interviews. Impersonality is inevitably linked with **formality** and **matter-of-factness**.

The weakening of the illocutionary force affects both the grammatical structure and the choice of vocabulary. In the attenuation function lexical means prevail, e.g. *probably, perhaps, possibly, in a way* etc. The grammatical expression of attenuation is connected with the occurrence of question phrases *I think, I don't think, I suppose, I mean* etc. followed by an object clause.

It can be summarized that certain pragmatic means are **double-edged**, being capable of producing either indirectness or attenuation. The two indeterminacy phenomena are interrelated, since the meaning expressed indirectly is equivalent to the meaning accompanied by hedging devices in being **implicit**. The line of demarcation between the two, however, can be easily drawn. Indirectness is related to the **speech act type**: a declarative question is an indirect speech act (compared with a direct question representing a direct speech act), whilst attenuation and accentuation modify the **illocutionary force of the speech act**; the speech act itself remains the same. Attenuation is generally understood as an expression of involvement producing the final effect of mitigation.

**Example 104: attenuation (apology)**

*I don't know I don't know no I don't think they will I hope not and I'll gloss it a bit of course the bit there is in that* (S.1.4.216-221)

**Example 105: indirectness (inquiry)**

**elicitation:** *this is very tricky I should have thought there were* (S.1.5.528-529)

**response:** *yes well quite they do that sort of thing you see* (S.1.5.530-532)

**Accentuation** reinforces the illocutionary force of the message. The language means utilized to achieve this effect are lexical rather than grammatical. The use of prosodically marked question phrases such as *I am absolutely convinced* is part of the grammatical repertoire, the same as exclamations expressed by means of full syntactic structures such as *that's a devil.*

## 10.4 Grammaticalness and Grammatical Acceptability

Aarts (1991) develops the notion of **language use** by trying to specify its characteristic features. From the grammatical point of view, he sees the intersection between **intuition-based** and **observation-based** grammar, i.e. between **grammatical** and **acceptable** sentences, on the one hand. On the other

hand, the existence of the two above-mentioned sentence categories has to be confronted with **corpus sentences**. Corpus sentences, i.e. all the sentences that are part of the corpus, are split into **grammatical** and **acceptable** (i.e. normative), and **ungrammatical** (i.e. non-normative).

In conclusion, it is necessary to point out that the pragmatic principles differ considerably from the rigid system of grammatical rules. A communicative grammar, such as Leech and Svartvik's *A Communicative Grammar of English* (1994.3), does not exclude grammatical competence but tries to integrate it within socio-cultural competence. Thomas (1995.105) advocates creativity rather than prescription with regard to language use: "...it is often the case in pragmatics that the most interesting effects are achieved when categories overlap or are blurred (such as one interactant can exploit the uncertainty) or are unclear to one of the participants. This applies not just in the case of speech acts, but to many other linguistic phenomena (such as discourse roles and activity types...) and it is a mistake to sacrifice the potential to exploit all the potential richness of meaning of speech acts for the sake of (the appearance of) a tidy system of rules".

Aarts (1991.58) speculates about the notion of **acceptability** with regard to corpus sentences and claims that: "if we write a grammar that accounts for every single sentence in a corpus, that grammar loses its (potential) generalizability... ." In conclusion he states that: "Whether or not a particular construction found in the corpus should be accounted for in the grammar is determined by the currency of that construction. The currency of a construction is compounded of its frequency of occurrence and its 'normalcy'" (1991.59). Frequency of occurrence cannot be regarded as a decisive factor, because even frequently occurring sentences need not be acceptable. **Normalcy** can be explained as acceptability "by a large number of language users" (1991.58).

In the research material I have analysed the problem of acceptability has not arisen, since the vast majority of utterances can be considered **acceptable**. Syntactic anacoluthon is rare, and the majority of "loose" structures are prevailingly incomplete, but not unacceptable.