Lecture4: Corpora

1. Special features of corpus linguistics

English has been analysed from a corpus linguistic perspective since the late 1970s. Corpus linguistics gives priority to descriptive adequacy. A diversity of text types in corpora makes it possible to test out linguistic hypotheses and describe the use of language as a communicative tool. The use of corpora provides language researches with controlled access to large amounts of usage data. Corpora reveal the range and frequencies of patterns of a language that learners assimilate. Various sophisticated tools have been designed for doing both quantitative and qualitative research these days. However, Aarts (2000: 7-8) points out that modern linguists should focus more on meaningful questions about the language being studied and go beyond the bare statistics. Chafe (1992: 96) states that: "A corpus linguist is a linguist who tries to understand language by carefully observing extensive natural samples of it and then constructing plausible understandings that encompass and explain those observations."

Corpora are valuable resources for descriptive, theoretical and applied discussions of language (Meyer 2002: 28). Corpora have been introduced into different linguistic disciplines and are used to study language change and variation, to understand the process of language acquisition, to improve foreign- and second-language instruction. Moreover, corpora are used for creating dictionaries. Corpora open up new areas of research and bring new insights to traditional research questions.

2. Corpora typology

Granger (1998, 2002) and Meyer (2002) give a full account of learner corpus design and analysis. They speak about a collection of texts or parts of texts that are used to carry out some linguistic research. According to whether English is learnt in an English-speaking country or not, "the learning context distinguishes between English as a Second Language (ESL) and English as a Foreign Language (EFL)" (Granger 1998: 9).

Corpora have numerous uses, ranging from the theoretical to the practical ones. "What one discovers in a corpus can be used as the basis for whatever theoretical issue one is exploring" (Meyer 2002: 4). For the current research the use of corpora is relevant in terms of studying of learner grammar and discourse.

Corpora vary in terms of the overall length of the corpus, the types of genres included, the number and age of texts, the length of individual text samples (see Meyer 2002: 30-45).

Historical corpora, such as the Helsinki and ARCHER provide resources for studies of the linguistic development of English. They contain samples of writing that represent earlier dialects and periods of English and allow for the study of changes in the language from the past to the present. These corpora are also useful for studying grammar and vocabulary.

Corpora of Modern English are often used for the study of language variation. For example, FLOB and FROWN consist of texts published in 1991. As synchronic corpora, on the one hand, they permit the study of varieties in British and American English. On the other hand, FLOB and FROWN replicate the LOB and Brown corpora (with texts published in 1961), and allow for studies of linguistic change in BE and AmE over a period of thirty years (Meyer 2002: 21).

Meyer (2002) notes that for the study of language varieties or for conducting a contrastive analysis, as well as for synchronic or diachronic comparison, it is better to use corpora of the same size. In this respect, the corpora of Brown family are suitable. They are divided into 2,000-word samples in varying genres (Meyer 2002: 145). The only limitation is that they exclude spoken material. Chafe (1992: 88) suggests that spoken corpora have a more favored place since "speaking is natural to the human organism in ways that writing can never be".

Multi-purpose corpora, such as BNC and the ICE Corpus29 consist of both written and spoken texts of different types (see Meyer 2002: 31, 35). These corpora represent similar genres and are used for studies of vocabulary, grammatical features, differences between various national varieties and genres of English.

There have been created other corpora for special purposes. Those that facilitate contrastive analyses of English and other languages are known as parallel corpora. For example, the English-Norwegian Parallel Corpus contains English and Norwegian fiction and non-fiction of similar types. This material can be used to study genre variation between two languages and to conduct a contrastive translation analysis. Parallel corpora are also valuable in terms of enhancing foreign language teaching.

3. Contribution of Learner Corpora to SLA Research

The so called learner corpora have been developed to facilitate the study of second-language acquisition. Current learner corpora are big in size and are used for particular SLA and LT purposes. Learner corpora give access to learners' total interlanguage and make it possible to conduct a contrastive interlanguage analysis (see Granger 1998: 12). In this respect learner corpora are used to study and compare the structure of various interlanguages that individuals from different first-language backgrounds develop. Moreover, researches can use learner corpora to test what non-native and native speakers of a language do in comparable situations.

A learner corpus has important implications for language teaching since it allows for a quantitative investigation of distinctive features of interlanguage: the frequency of use of certain words, phrases and structures, whether they are overused or underused. Descriptions of learner language can help to develop new pedagogical methods and approaches which target more accurately learners' needs.

Granger (1998: 4) describes SLA as a mental process and notes that learner performance data is necessary to uncover the principles that govern the process of learning a foreign or a second language. Three main data types are distinguished[1]: language use, metalingual judgements and self-report data.

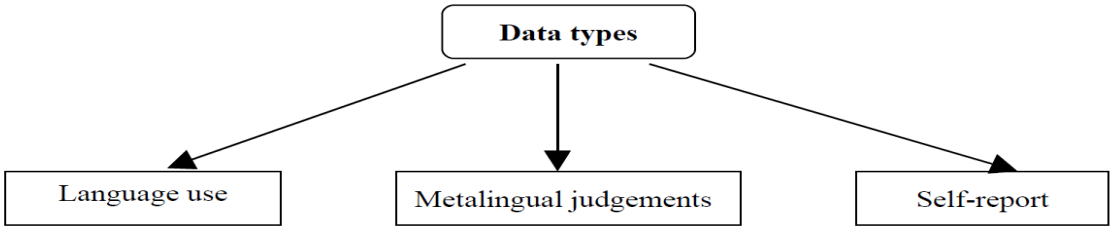

Figure 1. Learner performance data types

[1] This classification is taken from Granger (1998: 4, 2002: 5), but the data types are distinguished by Ellis (1994: 670).

The first data type reflects how learners use a second language in either comprehension or production. If no control is exerted on the language performance, the data will be natural. Language use data is elicited if it is based on the results of a controlled experiment. Metalingual judgements type concerns learners' intuition when they judge some instances of a language. The third data type is based on questionnaires or think-aloud tasks used to explore the ways learners acquire a second language.

The development of learner corpora contributes to the development of teaching strategies for individuals learning English as a second or foreign language (Meyer 2002: 27). The use of corpora helps to depict how learners are actually using the language. Various kinds of grammatical distinctions in English can be investigated by students themselves. Students of English as a foreign language can examine and figure out to what extent the speech or writing of native speakers of English is different from their English. Real examples of language usage taken from corpora differ obviously from those found in a majority of text- and grammarbooks. Vast amounts of data provided by learner corpora allow for exploration of real language. The only challenge concerns interpretation of data discovered. Coming back to Aarts (2000), corpus linguistics should focus more on qualitative research.

4. The International Corpus of Learner English

One of the larger learner corpora is called the International Corpus of Learner English (ICLE). The current size of this corpus is more than two million words. It is comprised of written English that represents one type of genre – essay writing. ICLE is divided into 500-word essays written by students from fourteen different linguistic backgrounds learning English as a foreign language (Granger 1998: 10).

What distinguishes a learner corpus from other corpora are design criteria for a specific purpose. ICLE shares some features with its subcorpora and has some variable ones. **Figure 2 illustrates ICLE design criteria.**

Table2: ICLE design criteria

Shared features	Variable features
Age	Sex
Learning context	Mother tongue
Level	Region
Medium	Other foreign languages
Genre	Practical experience
Technicality	Topic
	Task setting

ICLE includes mostly argumentative essay writing and a small proportion of literature exam papers. ICLE's medium distinguishes this corpus from spoken corpora, and within this medium the argumentative genre is

distinguished from narrative writing. This corpus contains writing by young male and female learners at an advanced level (university undergraduates) who study English as a foreign language in a non-English-speaking environment. This kind of environment refers to language context and is a crucial distinction between ESL and EFL. Learners' mother tongue background and their knowledge of other foreign languages are recorded in the corpus. It is an important factor that makes it necessary and useful to be aware of how learners' English may be influenced by other foreign languages.

The content of the essays included in ICLE is similar, but these written productions cover a variety of topics. It is a relevant factor since topics can affect the choice of lexical items and such a language feature as technicality. The degree of technicality can affect both the lexis and the complexity, as well as the frequency of grammatical items.

Each corpus has its **limitations**, and ICLE is not an exception. On the one hand, it is a lengthy corpus and allows for the study of lexis and grammar within the context of a complete text. On the other hand, only one genre and non-professional writing make up the corpus. As Biber (1993: 252) notes, diversity across text types contributes more to the achievement of broader linguistic representation. It is important to be aware of limitations when one chooses a corpus for a particular type of investigation to be carried out.

ICLE as a learner corpus is valuable in terms of providing researchers with information about English learnt by students of different mother tongues. Accurate descriptions of learner language can help to develop new classroom practices, especially those that concern developing writing skills. A qualitative account of research findings can help teachers of English to figure out what targets more accurately the needs of their learners.

Source:
https://www.duo.uio.no/bitstream/handle/10852/26174/ENG4190xThesisx.pdf?isAllowed=y&sequence=1

(pp 36-42)