PEOPLE’S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

Belhadj Bouchaib University – Ain- Temouchent

Institute of Letters, Social Sciences and English Language

Department of Letters and English Language

An Introduction to Corpus Linguistics
(Master Two Level)

Elaborated by: ***Dr. Chahrazed HAMZAOUI***

Docent at the Department of Letters and English Language

Academic Year: 2019-2020

Table of Contents

Introduction

Corpus Linguistics (CL), for many, is an end in itself. That is, it provides a means for the empirical analysis of language and in so doing, adds to its definition and description. This process has led to the refinement of descriptions of lexis, leading to immensely enhanced coverage in dictionaries. Now the application of CL is diverse in the extreme, as are the needs of its users. While a lexicographer is interested in how best to profile a word semantically, someone using CL in the study of second language acquisition may be interested in how aspects of language develop over time in one individual or a group of users. The course of corpus linguistics at the Master level is divided between lectures presented by the teacher and students' contributions in terms of research papers, assignments and oral presentations. It is mainly based on the lectures mentioned in this course-book. However, it will always remain open to the possibility of elaboration and adoption of other lectures as a response to the new advances in the area of corpus linguistics without, of course setting aside students' needs and expectations.

Objectives

In this course-book, we will try to bring together as diverse as possible miscellaneous definitions of CL and corpora as well as their advantages so as to capture the state-of-the-art in terms of how CL is being applied and might be applied in the future. Crucially for the use, development and vibrancy of CL, this process of application of CL to other areas has a wash-back effect for CL and in particular on how corpora and corpus software are designed. The course seeks to equip students with pertinent information about corpus linguistics' definitions, historical overview, scope, types, usage and advantages. More importantly, a focus is placed on the application of corpus linguistics to linguistic theory.

ICL/M2

Dr. HAMZAOUI

1. A Historical Perspective on Corpus Linguistics

Introduction

Corpus linguistics, nowadays, is perhaps most readily associated in the minds of linguists with searching through screen after screen of concordance lines and wordlists generated by computer software, in an attempt to make sense of phenomena in big texts or big collections of smaller texts. This method of exegesis (analysis) based on detailed searches for words and phrases in multiple contexts across large amounts of text can be traced back to the thirteenth century.

The etymology of concordantia is the Latin cum, meaning 'with', and core meaning 'heart', which ties in with the original ideological underpinning of this painstaking (careful) endeavour. The works of Shakespeare were also the subject of concordancing as a means of assisting scholars, for example Becket's (1787), a concordance to Shakespeare illustrates by way of extracts from Becket's concordance, the word and its linguistic context and location in the **Shakespeare canon is given**. For a literary scholar, this provides an immense resource. Though concordances from former times were laboriously compiled by hand, their spirit and intentions live on in the software programmes, we are now familiar with.

1.What drove the creation of modern corpora?

While the process of concordancing and indexing has its origins in the tidy work of literary scholars, the drive to create electronic corpora did not come from these quarters entirely. There was an influence from the work of Jesuit priest Roberto

Busa, who created an electronic lemmatized (converting a word to its base form) index of the complete works of St Thomas Aquinas, beginning in the 1950s and completing it in the late 1970s. At least two other forces are more significant, namely the work of lexicographers and that of pre-Chomskyan structural linguists. In both cases, collecting attested data was essential to their work. Dr Samuel Johnson's first comprehensive dictionary of English, published in 1755, was the result of many years of working with a paper corpus: that is, endless slips of paper logging samples of usage from the period 1560 to 1660.

And perhaps the most famous example of the 'corpus on slips of paper is the more than three million slips attesting word usage that the Oxford English Dictionary (OED) project had amassed by the 1880s, stored in what nowadays might serve as a garden shed. These millions of bits of paper were, quite literally, pigeon-holed in an attempt to organize them into a meaningful body of text from which the world-famous dictionary could be compiled. McEnery et al. (2006) note that the more specific term corpus linguistics did not come into common usage until the early 1980s; Aarts and Meijs (1984) is seen as the defining publication as regards coinage of the term.

Technology has been the major enabling factor in the growth of corpus linguistics but has both shaped and been shaped by it. The ability to store masses of data on relatively small computer drives and servers meant that corpora could be as big as one wanted. In this regard, lexicographers led the way. Their aim has always been to collect the maximum amount of data possible, so as to capture even the rare events in a language.

The early COBUILD corpora were measured in tens of millions of running words, other publishing projects soon competed and pushed the game up to hundreds of millions of words and, by the middle of the first decade of the twenty-first century, the Cambridge International Corpus (Cambridge University Press) had topped a billion running words of text.

ICL/ M2

Dr. HAMZAOUI

2. What is Corpus Linguistics?

Course contents: Introduction- Definition of corpus linguistics- What is considered corpus linguistics and what is not- The role of the corpus linguist- Conclusion

Introduction

Corpus Linguistics is a hugely popular area of linguistics which, since its initiation in the sixties, has revolutionized our understanding of language and how it works. *Corpus Linguistics* is a systematic guide to creating and analyzing linguistic corpora. It starts with a discussion of the role that corpus linguistics plays in linguistic theory, demonstrating that corpora have proven to be very useful resources for linguists who believe that their theories and descriptions should be based on real, rather than artificial data.

1. Definition of corpus linguistics

Corpus linguistics is the analysis of naturally occurring language on the basis of computerized corpora. Usually, the analysis is performed with the help of the computer, i.e. with specialized software, and takes into account the frequency of the phenomena investigated.

Corpus linguistics is, indeed, considered as a methodology to gain and analyze the language data either quantitatively or qualitatively. It can be applied in almost any area of language studies. It is also an object of a study in authentic, naturally occurring language use. *Corpus linguistics is not a separate branch of linguistics (like e.g. sociolinguistics, neurolinguistics, etc)) or a theory of language, but rather a methodology.*

Corpus linguistics is simply a tool for linguistic inquiry. That is “a methodological basis for pursuing linguistic research” as Leech (1992: 105) correctly says. Corpus linguistics dates back to the 1960’s, when the first computer corpus, the Brown Corpus, was created. This new trend was not welcomed by generativists who dominated at that time, and considered this corpus as “a useless and foolhardy enterprise” (Francis 1992: 28).

2. Corpus Linguistics as a Methodology

Researchers have debated heavily about the position of corpus linguistics in the field of linguistics as a whole. Some influential researchers, such as John Sinclair (2004) and Tognini-Bonelli (2001), have made strong claims that corpus linguistics should be considered to be a unique branch of linguistics that provides us with completely new ways to observe and understand language.

However, others in the field lean toward the view that corpus linguistics is essentially a methodology; a bag of resources, tools, and techniques that are used to help us understand how language works. From this perspective, although the insights gained from corpus linguistics might be profound, it is not a true sub-discipline of linguistics in the same way that phonology, pragmatics, syntax, and so on are usually described. (For an in-depth discussion on this topic from the perspectives of multiple corpus linguists, see Viana, Zyngier, & Barnbrook, 2011).

If corpus linguistics is considered to be essentially a methodology, research that contributes to the development of that methodology can be considered to be in some way “fundamental” to the field. This is in contrast to the more conventional use of the term “fundamental,” which refers to research conducted primarily to acquire new knowledge of the underlying foundations of a field (European Union, 2006). Following this terminology, “fundamental” corpus linguistics research would include the creation of new corpus data resources, analytical tools, statistical methods, and visualization techniques.

In contrast, research that utilizes these resources, tools, and techniques in other fields can be considered to be “applied” corpus linguistics research. Research in this category would include studies in the area of language understanding (e.g., receptive/productive studies related to phonology, pragmatics, syntax, morphology, discourse, vocabulary, and so on). It would also include studies on language teaching, learning, and testing, and the development of language engineering applications (e.g., web search engines, data-mining tools, query systems, flash-card learning programs, plagiarism detection systems, chat bots, and so on).

3. What is corpus linguistics and why is it useful?

Corpus linguistics is the study of language by means of naturally occurring language samples; analyses are usually carried out with specialised software programmes on a computer. Corpus linguistics is thus a method to obtain and analyse data quantitatively and qualitatively rather than a theory of language or even a separate branch of linguistics on a par with e.g. sociolinguistics or applied linguistics. The corpus-linguistic approach can be used to describe language features and to test hypotheses formulated in various linguistic frameworks.

To name but a few examples, corpora recording different stages of learner language (beginners, intermediate, and advanced learners) can provide information for foreign language acquisition research; by means of historical corpora it is possible to track the development of specific features in the history of English like the emergence of the modal verbs *gonna* and *wanna*; or sociolinguistic markers of specific age groups such as the use of *like* as a discourse marker can be investigated for purposes of sociolinguistic or discourse-analytical research.

In other words, corpus linguistics is a method of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years.

The great advantage of the corpus-linguistic method is that language researchers do not have to rely on their own or other native speakers' intuition or even on made-up examples. Rather, they can draw on a large amount of authentic, naturally occurring language data produced by a variety of speakers or writers in order to confirm or refute their own hypotheses about specific language features on the basis of an empirical foundation.

4. What is considered corpus linguistics and what is not

Corpus linguistics approaches the study of language in use through corpora (sing: corpus). Briefly speaking, corpus linguistics serves to answer two fundamental research questions

1- What appropriate patterns are associated with lexical or grammatical traits?

2- How do these patterns vary within varieties and registers?

Many leading figures have contributed into the evolvement of today corpus linguistics: Leech, Biber, Johansson, Hunston, Francis, McCarty and Conrad to name a few. These scholars have made significant contributions to corpus linguistics, both past and present. Many corpus linguists, however, consider John Sinclair to be one, if not the most influential scholar of modern-day corpus linguistics. Sinclair discovered that a word in and of itself does not carry meaning, but that meaning is often made through several words in a sequence (Sinclair, 1991). This is the idea that forms the strength and determination of corpus linguistics.

However, it is also of paramount importance to understand what corpus linguistics is not. This could be explained through the following points where corpus linguistics is not

- able to provide negative evidence
- able to explicate why
- able to provide all possible language in one time

Corpus linguistics is not able to provide negative evidence. This means a corpus cannot tell us what is possible or correct or not possible or incorrect in language; it can only tell us what is or is not present in the corpus. Many instructors mistakenly believe that if a corpus does not present all manners to express a certain idea, then the corpus is altogether faulty. Instead, instructors should believe that if a corpus does not present a particular manner to express a certain idea, then perhaps that manner is not very common in the register represented by the corpus.

Corpus linguistics is not able to explain why something is the way it is, it only tells us what it is. To discover why, we, as users of language, use our intuition. Corpus linguistics is not able to provide all possible language at one time. By definition, a corpus should be principled: "a large, principled collection of naturally occurring texts . . .," meaning that the language that enters a corpus is not random, but planned. However, no matter how planned, principled or large, a corpus is, it cannot be representative of all language. In other words, even in a corpus that contains one billion words, such as the Cambridge International Corpus (CIC), not all instances of use of a language may be present.

5. The role of the corpus linguist

Among the roles of the corpus linguist is to create a corpus; he is a corpus compiler and hence becomes a field linguist because he goes out collecting and recording speech in various locations: homes, offices, schools, universities, work place, and so forth. It is noteworthy that the corpus linguist creates a corpus not for studying it. Instead, he creates it for others to study.

Corpus linguists show a kind of dissatisfaction vis à vis the abstract and decontextualized linguistic data supported by generative grammarians, and recently "linguists of various persuasions use corpora in their research and are united in their belief that one's linguistic analysis will benefit from the analysis of real language" (Meyer 2002: 2). The contribution of corpus linguistics in linguistic research is overwhelming in many respects.

6. What data do linguists use to investigate linguistic phenomena? Roughly, four types of data for linguistic analysis can be distinguished:

1) data gained by intuition

a) the researcher's own intuition ("introspection")
b) other people's (informant's) intuition (accessed, for example, by elicitation tests)

2) naturally occurring language

a) randomly collected texts or occurrences ("anecdotal evidence")
b) systematic collections of texts ("corpora") (For further reading on corpora vs. intuition, see Fillmore 1992).

Conclusion

In conclusion, this course has introduced you to what is known as corpus linguistics which is above all a methodology and not a discipline of linguistics. It has also shed light on what corpus linguistics is and what it is not in order to avoid any bias about this special area of study.

Even though descriptive/theoretical linguists and computational linguists use corpora for very different purposes, they share a common belief: that it is important to base one's analysis of language on real data – actual instances of speech or writing – rather than on data that are contrived or "made-up". In this sense, then, corpus linguistics is not a separate paradigm of linguistic, but rather a methodology.

References/ Further readings

Francis, W. N. (1992). Language corpora B.C. In Svartvik (ed). 17-32.

Hardie, A. and McEnery, T. (2010) 'On two traditions in corpus linguistics, and what they have in common', *International Journal of Corpus Linguistics* 15(3): 384-94.

Leech, G. (2011) 'Principles and applications of corpus linguistics' in Viana, V., Zyngier, S. and Barnbrook, G. (eds.) *Perspectives on Corpus Linguistics*, pp155-70. Amsterdam: John Benjamins.

Meyer, C. F. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). Trust the text: Language, corpus and discourse. Abingdon, UK: Routledge. Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Philadelphia, PA: John Benjamins.

Viana, V., Zyngier, S., & Barnbrook, G. (Eds.). (2011). Perspectives on corpus linguistics. Philadelphia, PA: John Benjamins.

Self-assessment: Answer the following questions using your own words.

1. Is corpus linguistics a branch of linguistics? Why?
2. What is considered corpus linguistics and what is not?

3. Definition of a Corpus

***Course Contents**:* Introduction- What is a Corpus?- Conspicuous Features of a Corpus- Size of a Corpus- Use of a Corpus- Utility of a corpus- Conclusion

Introduction

The term *corpus* is derived from Latin *corpus* "body". As Leech (1992) points out, it was in the 1950s, in the era of American structuralists such as Harris, Fries and Hill, just to name a few, when the notion of collecting real data appeared.

1. What is a Corpus?

The term corpus is derived from the Latin word corpus that means “body”. The Latin term, however, displays two distinct descendants in modern English:

(a) corpse (it came via Old French cors) and

(b) corps (it came via modern French corps in the 18th century).

The first form (i.e. corpse) entered into English in the thirteenth century as cors and during the fourteenth century it had its original Latin ‘p’ reinserted. At first it meant simply ‘body’, but by the end of the fourteenth century, the sense ‘dead body’ became firmly established. However, on the other hand, the original Latin term corpus itself was acquired in English in the fourteenth century (Ayto 1990: 138).

Within the domain of modern corpus linguistics, the term ‘corpus’ refers to "a large collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language" (Crystal 1995). Thus, it refers to a large collection of written and spoken text samples, available in machine-readable form, accumulated in scientific manner to represent a particular variety or use of a language.

According to scholars, a corpus is a collection of linguistic items that are selected and ordered according to some explicit linguistic criteria defined by the users in order to be used as a sample of a language. It is methodically designed to contain millions of word compiled from diverse text types across many demographic variations to encompass the diversity a natural language exhibits through its multifaceted use.

McEnery and Wilson (1996: 215) have classified corpus in a finer scheme of classification characterised by its inherent features:

(a) Loosely, a corpus refers to anybody of text;

(b) Most commonly, it refers to a body of machine-readable text;

and (c) More strictly, it refers to a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of it.

In principle, a corpus is actually designed for accurate study of the linguistic properties, features, and phenomena observed in a language. Therefore, we have argued that a systematically compiled corpus, however small in size, should adhere to the following criteria (Dash 2005: 12):

• A corpus should faithfully represent both the common and special linguistic features of the language from which it is designed and developed. The idea of text representation in a corpus indirectly refers to the total sum of its components (i.e. words, phrases, clauses, sentences, etc.) included in it. However, in practice, the total number of words included in a corpus may determine its size but may fail to abide by the principle of proper text representation.

Therefore it is better to keep fields open for a corpus as well as keep number of words unlimited for the benefit of language and users.

• A corpus should be large and wide to encompass texts from various disciplines. In other words, directional varieties of language use manifested in various disciplines and domains should have proportional representation in it. For instance, text samples from the fields of natural sciences should carry equal weight as those from aesthetics, literature, mass media, engineering, and social sciences. Thus, a balanced representation of text samples obtained from all disciplines and domains of language use will ensure its reliability.

- A corpus should be a true replica of physical texts available in printed form. Thus, it should faithfully preserve various word forms, spelling variations, punctuation marks as well as various other orthographic symbols used in the source texts. Else, the actual image of a language or the language variety will be distorted and a corpus will lose its value and authenticity.

- A corpus should be available in the electronic form for easy access by the end users in order to enable common users as well as language researchers to use the database in multiple tasks related to language description and analysis, statistical analysis, language processing, translation, etc. As corpus designers our basic task is to gather large amount of representative text samples covering wide varieties of language used in various domains of our regular linguistic interaction. Since a corpus is capable of representing potentially unlimited selections of text, it may be defined acrostically from the letters used to compose the term in following way (Dash 2005: 4):

C : Compatible to both man and computer,

O : Operational in research and application,

R : Representative of a language or a variety,

P : Processable by both man and machine,

U : Unlimited in the amount of data and samples,

and S : Systematic both in formation and representation. Unless defined otherwise, let us assume that a corpus will possess all the properties mentioned above. Exception may be made for historical corpora, which have limited use due to their diachronic form and composition. Historical corpora are mostly used within specific areas of historical linguistics that attests indirect importance in the field of empirical language research. In essence, a well-defined and systematically developed corpus is an empirical standard, which acts as a valuable benchmark for validation of usage of all linguistic properties available in a natural language.

A corpus (plural *corpora*, German "das Korpus", not "der") is a collection of texts used for linguistic analyses, usually stored in an electronic database so that the data can be accessed easily by means of a computer. Corpus texts usually consist of thousands or millions of words and are not made up of the linguist's or a native speaker's invented examples but on authentic (naturally occurring) spoken and written language.

A corpus can be defined as a systematic collection of naturally occurring texts (of both written and spoken language). "Systematic" means that the structure and contents of the corpus follows certain extralinguistic principles ("sampling principles", i.e. principles on the basis of which the texts included were chosen). For example, a corpus is often restricted to certain text types, to one or several varieties of English, and to a certain time span.

If several subcategories (e.g. several text types, varieties etc.) are represented in a corpus, these are often represented by the same amount of text. "Systematic" also means that information on the exact composition of the corpus is available to the researcher (including the number of words in each category and in the whole corpus, how the texts included in the corpus were sampled etc).

Although "corpus" can refer to any systematic text collection, it is commonly used in a narrower sense today, and is often only used to refer to systematic text collections that have been computerized. The majority of present-day corpora are "balanced" or "systematic". This means that the texts are collected ("compiled") according to specific principles, such as different genres, registers or styles of English (e.g. written or spoken English, newspaper editorials or technical writing); these sampling principles do not follow language-internal but language-external criteria. For example, the texts for a corpus are not selected because of their high number of relative clauses but because they are instances of a predefined text type, say broadcast English in a hypothetical corpus of Australian British English. Examples of balanced corpora are the *International Corpus of English (ICE)*, the *British National Corpus (BNC)*, or the *Brown* and *Lancaster-Oslo/Bergen (LOB)* corpora and their Freiburg updates (*Frown* and *F-LOB*).

A corpus is thus a systematic, computerised collection of authentic language used for linguistic analysis.

2. Some of the definitions of the term 'corpus' are as follows:

A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language.

(David Crystal, A Dictionary of Linguistics and Phonetics, Blackwell, 3rd Edition, 1991)

A collection of naturally occurring language text, chosen to characterize a state or variety of a language.

(John Sinclair, Corpus Concordance, Collocation, OUP, 1991)

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Sinclair, J. 2005. "Corpus and Text - Basic Principles" in Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books.

Actually, corpus means representative collection of texts of a given language, dialect or other subset of a language to be used for linguistic analysis. In finer definition, it refers to

(a) (loosely) any body of text;

(b) (most commonly) a body of machine readable text;

and (c) (more strictly) a finite collection of machine-readable texts sampled to be representative of a language or variety (McEnery and Wilson 1996: 218).

Corpus contains a large collection of representative samples of texts covering different varieties of language used in various domains of linguistic interactions. Theoretically speaking, a corpus is able to represent potentially infinite selections of texts. It is compatible to computer, operational in research and application, representative of the source language, processable by man and machine, unlimited in data, and systematic in formation and representation (Dash 2005: 35).

3. Conspicuous Features of a Corpus

Quantity: It should be big in size containing large amounts of data either in spoken or written form. Size is virtually the sum of its components, which constitute its body.

· **Quality (=** authenticity). All texts should be obtained from actual instances of speech and writing. The role of a linguist is crucial here. He has to check whether language data is collected from ordinary communication, and not from experimental conditions or contrived circumstances.

· **Representation:** It should embrace samples from a wide range of texts. It should be balanced to all areas of language use to represent maximum linguistic diversities, as future analysis may need verification and authentication of information from the corpus representing a language.

· **Simplicity:** It should contain clear texts in simple format. This means that we expect an unbroken string of characters (or words) without any additional linguistic information marked-up within texts. A simple clear text is opposed to any kind of annotation (explanatory note) with different types of linguistic and non-linguistic information.

· **Equality:** Samples used in corpus should be of even size. However, this is a controversial issue and will not be adopted everywhere. Sampling model may change considerably to make a corpus more representative and multi-dimensional.

· **Retrievability:** Data, information, examples, and references should be easily retrievable from corpus by the end-users. This pays attention to preserving techniques of language data in electronic format in computer. The present technology makes it possible to generate corpus in PC and preserve it in such way that we can easily retrieve data as and when required.

· **Verifiability:** Corpus should be open to any kind of empirical verification. We can use data form corpus for any kind of verification. This puts corpus linguistics steps ahead of intuitive approach to language study.

· **Augmentation:** It should be increased regularly. This will put corpus in equilibrium to register linguistic changes occurring in a language in course of time. Over time, by addition of new linguistic data, a corpus achieves historical dimension for diachronic studies, and for displaying linguistic cues to arrest changes in life and society.

· **Documentation:** Full information of components should be kept independent from the text itself. It is always better to keep documentation information independent from the text, and include only a minimal header containing reference to documentation. In case of corpus management, this permits effective disconnection of clear texts from annotation with only a small amount of programming effort.

4. Size of a Corpus

How big a corpus will be? This implies that size is an important issue in corpus generation. It is concerned with total number of words (tokens) and different words (types) to be taken into a corpus. It also involves the decision of how many categories we like keep in the corpus, how many samples of texts we need put into each category, and how many words we shall keep in each sample. Although the question of size affects validity and reliability of a corpus, it is stressed that any corpus, however big, is nothing more than a minuscule sample of all speech and writing varieties produced by the users of a language.

In early days of corpus generation, when computer technology for procuring language data was not much advanced, it was considered that a corpus containing one million words is large enough to represent a language or variety. However, by the middle of 1980s, computer technology went through a vast change with unprecedented growth of its storage, processing, and accessing abilities that have been instrumental in changing the concept regarding the size of a corpus.

Now it is believed that the bigger the size of corpus the more it is faithful in representing the language under consideration. With advanced computer technology, we can generate corpus of very large size containing hundreds of millions of words. For instance*, the British National Corpus, the COBUILD Corpus,* the *Longman/Lancaster Corpus, the International Corpus of English*, the *American National Corpus, etc.* are indeed very large in size – each one containing more than hundred million words.

5. Use of a Corpus

There are a number of areas where language corpus is directly used as in *language description, study of syntax, phonetics and phonology, prosody, intonation, morphology, lexicology, semantics, lexicography, discourse, pragmatics, language teaching, language planning, sociolinguistics, psycholinguistics, semiotics, cognitive linguistics, computational linguistics* — to mention a few. In fact, there is hardly any area of linguistics where corpus has not found its utility. This has been possible due to great possibilities offered by computer in collecting, storing, and processing natural language databases. The availability of computers and machine-readable corpora has made it possible to get data quickly and easily and to have this data presented in a format suitable for analysis.

. **Corpus as knowledge resource:** A corpus is used for developing multilingual libraries, designing course books for language teaching, compiling monolingual dictionaries (printed and electronic), developing bilingual dictionaries (printed and electronic), multilingual dictionaries (printed and electronic), monolingual thesaurus (printed and electronic version), various reference materials (printed and electronic version), developing machine readable dictionaries (MRDs), developing multilingual lexical resources, electronic dictionary (easily portable, can be duplicated as many copies as needed, can be modified easily for newer versions, can be customised according to need of users, can be ready and accessed easily, more durable than printed dictionary, etc.).

· **Corpus in language technology:** corpus is used for designing tools and systems for word processing, spelling checking, text editing, morphological processing, sentence parsing, frequency counting, item-search, text summarisation, text annotation[1], information retrieval, concordance, word sense disambiguation, WordNet (synset), semantic web, Semantic Net, Parts-of-Speech Tagging, Local Word Grouping, etc.

· **Corpus for translation support systems:** corpus is used for language resource access systems, Machine translation systems, multilingual information access systems, and cross-language information retrieval systems, etc.

· **Corpus for human-machine interface systems:** corpus is used for voice recognition, texto-speech, E-learning[2], on-line teaching, e-text preparation, question-answering, computer-assisted language education, computer-aided instruction, etc.

· **Corpus in speech technology:** Speech corpus is used to develop general framework for speech technology, phonetic, lexical, and pronunciation variability in dialectal versions, automatic speech recognition, automatic speech synthesis,

[1] T**ext Annotation** is the practice and the result of adding a note or **gloss** to a text, which may include highlights or underlining, comments, footnotes, tags, and links.

[2] learning conducted via electronic media, typically on the Internet.

automatic speech processing, speaker identification, repairing speech disorders, and forensic linguistics, etc.

· **Corpus in mainstream linguistics:** corpus is used for language description, lexicography, lexicology, grammar writing, semantic study, language learning, dialect study, sociolinguistics, psycholinguistics, stylistics, bilingual dictionary, lexical selection restriction, dissolving lexical ambiguity, semiotics, pragmatic and discourse study, etc.

6. Utility of Corpus

In essence, a corpus is an empirical standard, which acts as a criterion for validation of usage of linguistic properties found in a language. If one analyses a corpus database, one can retrieve the following information about a language or variety.

· Information about all the properties and components used in a language, e.g., sounds, phonemes, intonation, letters, punctuations, morphemes, words, compounds, phrases, idioms, set phrases, proverbs, clauses, sentences, etc.

· Grammatical and functional information of letters, allographs, morphemes, words, phrases, sentences, idiomatic expressions, proverbs, etc.

· Usage-based information of letters, characters, phonemes, morphemes, words, compounds, phrases, sentences, etc., relating their descriptive, stylistic, metaphorical, allegorical, idiomatic, and figurative usages, etc.

· Extralinguistic information relating to time, place, situation, and agent of language events, sociocultural backgrounds of linguistic acts, life and living of target speech community, discourse and pragmatics, as well as of the world knowledge of the language users at large.

Conclusion

It is understandable that developing a corpus in accordance with these pre-conditions is really a hard task. However, we can simplify the task to some extent if we redefine the entire concept of corpus generation based on object-oriented and work-specific needs. Since it is known that all types of corpus should not follow the same set of designing and composition principles, we can have liberty to design a corpus keeping in mind the works we are planning to do with it (Dash 2008: 47). The basic proposition is that the general principles and conditions of corpus generation may vary depending on the purpose of a corpus developer or a user.

Corpus linguistics is, however, not the same thing as obtaining language databases through the use of computer. It is the processing and analysis of the data stored within a corpus. The main task of a corpus linguist is not to gather databases, but to analyse these. Computer is a useful, and sometimes indispensable, tool for carrying out these activities.

References/ Further Readings

Biber, D., Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Exploring Language Structure and Use. Cambridge: Cambridge University Press.

Leech, G. (1992) ‘Corpora and theories of linguistic performance’, in Svartvik, J. (ed.) Directions in corpus linguistics: proceedings of Nobel symposium 82. Berlin and New York: Mouton de Gruyter, pp. 125–148.

Sinclair, J. (1991a) Corpus, Concordance and Collocation. Oxford: Oxford University Press.

4. Corpus Analysis and Linguistic Theory

Course contents: Introduction- Corpus-based research in linguistics-Main fields of application of corpus linguistics- Conclusion

Introduction

When the first computer corpus, the Brown Corpus, was being created in the early 1960s, generative grammar dominated linguistics, and there was little tolerance for approaches to linguistic study that did not adhere to what generative grammarians deemed acceptable linguistic practice. However, even though generative grammarians and corpus linguists have different goals, it is wrong to assume that the analysis of corpora has nothing to contribute to linguistic theory: corpora can be invaluable resources for testing out linguistic hypotheses based on more functionally based theories of grammar, i.e. theories of language more interested in exploring language as a tool of communication.

1. Corpus-based research in linguistics

As has been noted, corpus linguistics is essentially a methodology or set of methodologies, rather than a theory of language description. Essentially, corpus linguistics means this:

. looking at naturally occurring language;

. looking at relatively large amounts of such language;

. observing relative frequencies, either in raw form or mediated through statistical operations;

. observing patterns of association, either between a feature and a text type or between groups of words.

Linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various research agendas. For instance, many lexicographers have found that they can more effectively create dictionaries by studying word usage in very large linguistic corpora.

1.1. Grammatical studies of specific linguistic constructions

Studies of this kind can test hypotheses arising from grammatical descriptions based on intuition or on limited data. Experiments have been designed specifically to do this (Nelson et al., 2002: 257–283). For example, Meyer (2002: 7–8) describes work on ellipsis from a typological and psycholinguistic point of view that predicts that of the three possible clause locations of ellipsis in American spoken English, one will be much more frequent than the others.

1.2. Reference grammars

More recent reference grammars have relied even more heavily on corpora. These grammars use corpora to provide information on the form and use of grammatical constructions, but additionally contain extensive numbers of examples from corpora to illustrate the grammatical constructions under discussion. As an illustration, Biber et al.'s *Longman Grammar of Spoken and Written English* (1999) is based on the Longman Spoken and Written English Corpus, a corpus that is approximately 40 million words in length and contains samples of spoken and written British and American English. This grammar provides extensive information not just on the form of various English structures, but also on their frequency and usage in various genres of spoken and written English.

1.3. Lexicography

To understand why dictionaries are increasingly being based on corpora, it is instructive to review precisely how corpora and the software designed to analyze them, can not only automate the process of creating a dictionary but also improve the information contained in the dictionary. A typical dictionary, as Landau (1984: 76f.) observes, provides its users with various kinds of information about words: their meaning, pronunciation, etymology, part of speech, and status (e.g. whether the word is considered "colloquial" or "non-standard"). In addition to making the process of creating a dictionary easier, corpora can improve the kinds of information about words contained in dictionaries and address some of the deficiencies inherent in many dictionaries.

Dictionaries have also been criticized for the unscientific manner in which they define words, a shortcoming that is obviously a consequence of the fact that many of the more traditional dictionaries were created during times when well-defined theories of lexical meaning did not exist. For instance, Fillmore's (1992) analysis of the various meanings of the word *risk* in a corpus effectively illustrates the value of basing a dictionary on actual uses of a particular word. As Fillmore (1992: 39) correctly observes, "the citation slips the lexicographers observed were largely limited to examples that somebody happened to notice . . ." But by consulting a corpus, the lexicographer can be more confident that the results obtained more accurately reflect the actual meaning of a particular word.

1.4. Language variation

In sociolinguistics, the primary focus is how various sociolinguistic variables, such as age, gender, and social class, affect the way that individuals use language. One reason that there are not more corpora for studying this kind of variation is that it is tremendously difficult to collect samples of speech, for instance, that are balanced for gender, age, and ethnicity.

However, despite the complications that studying linguistic variables poses, designers of some recent corpora have made more concerted efforts to create corpora that are balanced for such variables and that are set up in a way that information on these variables can be extracted by various kinds of software programmes. For example, prior to the collection of spontaneous dialogues in the British National Corpus, a sub corpus known as the Corpus of London Teenage

English (COLT) contains a valid sampling of the English spoken by teenagers from various socioeconomic classes living in different boroughs of London.

1.5. Historical linguistics

There exist a number of historical corpora – corpora containing samples of writing representing earlier dialects and periods of English – that can be used to study not only language variation in earlier periods of English but changes in the language from the past to the present.

Much of the interest in studying historical corpora stems from the creation of the Helsinki Corpus, a 1.5-million-word corpus of English containing texts from the Old English period (beginning in the eighth century) through the early Modern English period (the first part of the eighteenth century).

Historical corpora have greatly enhanced our ability to study the linguistic development of English: such corpora allow corpus linguists not only to study systematically the development of particular grammatical categories in English, but also to gain insights into how genres in earlier periods differed linguistically and how sociolinguistic variables such as gender affected language usage.

1.6. Language acquisition

Although studies of language acquisition have always had an empirical basis, researchers in the areas of first- and second-language acquisition have tended not to make publicly available the data upon which their studies were based. However, this situation is changing and there now exist corpora suitable for studying both first- and second-language acquisition.

To facilitate the study of both first- and second-language acquisition, the CHILDES (Child Language Data Exchange System) was developed. This system contains a corpus of transcriptions of children and adults learning first and second languages that are annotated in a specific format called “CHAT” and that can be analyzed with a series of software programmes called “CLAN” (MacWhinney 1996: 2).

1.7. Language pedagogy

One consequence of the development of learner corpora is that researchers are taking information from them to develop teaching strategies for individuals learning English as a second or foreign language. In addition to using information from learner corpora to develop teaching strategies for learners of English, many have advocated that students themselves study corpora to help them learn about English, a methodology known as “data-driven learning” (Johns 1994 and Hadley 1997). This method of teaching has students investigate a corpus of native-speaker speech or writing with a concordance programme to give them real examples of language usage rather than the contrived examples often found in grammar books.

2. Main fields of application of corpus linguistics

Applied linguistics has been described as the use of knowledge about language to solve real-world problems. In recent years, the benefits of looking at large amounts of naturally occurring language, in the form of corpora, have been welcomed by applied linguists. Although corpora have many applications – notably Lexicographic and lexical studies ,Grammatical studies, Register variation and genre analysis, Dialect distinction and language variety, Contrastive and translation studies, Diachronic study and language change, Language learning and teaching, Semantics, Pragmatics, Sociolinguistics, Discourse analysis, Stylistics and literary studies, Forensic linguistics -this section describes briefly just two of the more frequently encountered applications of corpus linguistics: language teaching and translation.

3. Language Teaching

3. Corpora have influenced language teaching in three distinct ways. Firstly, the findings from corpus research have been used extensively to improve reference materials for learners, such as dictionaries and grammars. Secondly, learners are increasingly being encouraged to explore corpora for themselves. Finally, corpus techniques have been applied to study of learners’ language. Since the publication in the mid-1980s of the first

learners' dictionary based on corpus research (Sinclair et al., 1987), corpora have become an indispensable resource for lexicographers and grammarians.

4. Modern learners' dictionaries typically pay more attention to phraseology, and in particular to collocation, than previous ones did. Similarly, grammar books for learners pay more attention to register variation, to spoken usage, and to the role of lexis in grammar (Sinclair et al., 1990; Biber et al., 1999). To a lesser extent, course books have also changed, now placing more emphasis on collocation and phraseology than previously. Corpora have influenced the method, as well as the content of language teaching. Advanced learners are frequently invited to access corpora themselves and to engage in ''data-driven learning'' (Johns, 1991; Bernadini, 2000), in which they use a corpus to make their own generalizations about language use.

5. One of the consequences of this is that learners are exposed to all the complexity of a language, and the task of teaching explicitly every aspect of that language looks less viable than it did before. As a result, data-driven learning coincides happily with the view of language learning that stresses guided observation on the part of the learner rather than exposition on the part of the teacher (Willis, 2003; Bernardini, 2004).

6. Finally, the language of learners themselves has been studied extensively through the development of learner corpora (Granger, 1998), that is, corpora consisting of collections of written or spoken texts produced by learners of a language. These allow the learners' output to be compared with that of native speakers and for persistent errors in learner language to be identified.

7. A common methodology is to identify features of language that occur significantly more or less frequently in the learner corpus than in a comparable corpus of native-speaker texts and to use such disparities as the starting point for more qualitative research. The features investigated

include groups of words such as adverbials (Altenberg and Tapper, 1998) and modal auxiliaries (Aijmer, 2002), as well as more abstract categories, such as word class (Granger and Rayson, 1998) and sequences of part-of-speech tags (Aarts and Granger, 1998).

4. Translation

8. Corpora can be used to train translators, used as a resource for practicing translators, and used as a means of studying the process of translation and the kinds of choices that translators make. Parallel corpora are often used in these applications, and software exists that will ‘align’ two corpora such that the translation of each sentence in the original text is immediately identifiable.

9. This allows one to observe how a given word has been translated in different contexts (see, for example, Teubert’s work on travail and work/labor mentioned in the section ‘Languages and Varieties’). One interesting finding is that apparently equivalent words – such as English go and Swedish ga˚, or English with and German mit (Viberg, 1996; Schmied and Fink, 2000) – occur as translations of each other in only a minority of instances. This suggests differences in the ways those languages use the items concerned.

10. More generally, examination of parallel corpora emphasizes that what translators translate is not the word but a larger unit (Teubert and Cˇ erma´kova´, 2004). Although a single word may have many equivalents when translated, a word in context may well have only one such equivalent. For example, although travail as an individual word is sometimes translated as work and sometimes as labor, the phrase travaux pre´paratoires is translated only as preparatory work.

11. Thus, Teubert and Cˇ erma´kova´ argue, travaux pre´paratoires and preparatory work may be considered to be equivalent translation units, whereas no such claim can be made for travaux and work. As well as

giving information about languages, corpus studies have also indicated that translated language is not the same as nontranslated language. Studies of corpora of translated texts have shown that they tend to have higher incidences of very frequent words and that they tend to be more explicit in terms of grammar (Baker, 1993).

12. They may also be influenced by the structure of the source language, as was indicated in the discussion of wh- clefts in English and Swedish in the section ‘Languages and Varieties.’ In communities where people read a large number of translated texts, the foreign language, via its translations, may even influence the home language. Gellerstam (1996) notes that some words in Swedish have taken on the meanings of English that look similar and argues that this is because translators tend to translate the English word with the similarlooking Swedish word, thereby using the Swedish word with a new meaning, which then enters the language. One example is the Swedish word dramatisk, which used to indicate something relating to drama but which now, like the English word dramatic, also means ‘substantial and surprising.

Conclusion

Corpora have numerous uses, ranging from the theoretical to the practical, making them valuable resources for descriptive, theoretical, and applied discussions of language. Because corpus linguistics is a methodology, all linguists – even generativists – could in principle use corpora in their studies of language. In many other disciplines of linguistics, corpora have proven to be valuable resources: they are used for creating dictionaries, studying language change and variation, understanding the process of language acquisition, and improving foreign- and second-language instruction.

Corpus linguistics is a relatively new discipline, and a fast-changing one. As computer resources, particularly web-based ones, develop, sophisticated corpus investigations come within the reach of the ordinary translator, language learner, or linguist. Our understanding of the ways that types of language might vary from one another, and our appreciation of the ways that words pattern in language, have been immeasurably improved by corpus studies. Even more significant, perhaps, is the development of new theories of language that take corpus research as their starting point.

References/ Further reading

Aarts J & Granger S (1998). 'Tag sequences in learners corpora: a key to interlanguage grammar and discourse.' In Granger (ed.). 132–142.

Altenberg, Bengt and Marie Tapper (1998) The Use of Adverbial Connectors in Advanced Swedish Learners' Written English. In Granger (1998). 80−93.

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

Fillmore, Charles (1992) Corpus Linguistics or Computer-Aided Armchair Linguistics. In Svartvik(1992). 35−60.

G & Hunston S (eds.) System and corpus: exploring connections. London: Equinox.

MacWhinney, Brian (1996) The CHILDES System. *American Journal of Speech-Language Pathology* 5. 5−14.

(2000) *The CHILDES Project: Tools for Analyzing Talk*. 3rd edn., vol. 1: *Transcription Format and Programs*, vol 2: *The Database*. Mahwah, NJ: Erlbaum.

Meyer C (2002). English corpus linguistics: an introduction.Cambridge: Cambridge University Press.

Nelson G, Wallis S & Aarts B (2002). Exploring natural language: working with the British component of the International Corpus of English. Amsterdam: Benjamins.

ICL/M2

Dr. HAMZAOUI

5. The Corpus Approach

Course contents: Introduction- The different corpus approaches- The corpus-driven approach- The corpus-based approach- Characteristics of The corpus approach – Target features Conclusion.

Introduction

This lecture is designed to introduce the students to the different corpus approaches, the main characteristics of the corpus approach and the target features such as phraseology, lexicogrammar, registers, English for Specific Purposes (ESP) and appropriate syllabus design.

In order to broaden their knowledge as regards corpus linguistics. *Elexiko*, the first German hypertext dictionary compiled exclusively on the basis of an electronic corpus, offers a new way of presenting sense relations, using a variety of approaches to extract the necessary data. In this paper, I will show how *elexiko* presents a differentiated system of paradigmatic relations including synonymy, various subtypes of incompatibility (such as antonymy, complementarity, converseness, reversiveness, etc.), and vertical structures (such as hyponymy and meronymy).

Primary attention, however, will focus on the question of how data for a paradigmatic description is retrieved from the corpus. Whereas a corpus-driven approach is mainly used for various semantic information and a corpus-based method plays an important part in obtaining data for the grammatical description in *elexiko*, it will be argued that both the corpus-driven and the corpus-based approach can be complementary methods in gaining insights into sense relations.

1. Preliminaries

The study of contextual relations, such as sense relations, is significant when investigating the structures of the lexicon of a language.

> Natural vocabularies are not random assemblages of points in semantic space: there are quite strong regularizing and structuring tendencies, and one type of these manifests itself through sense relations. (Cruse 2004: 143)

Sense relations offer insights into the meaning and use of a word, and they reveal the interrelatedness of the vocabulary. As Cruse (1986: 16) points out "the meaning of a word is fully reflected in its contextual relations". However, contextual relations not only possess a fascination for semanticists, but they also attract the interest of lexicographers. Contextual relations contribute to the semantic identity of a word, and they have therefore always played an important role in disambiguating word senses in lexicography (cf. Reichmann 1989: 111-114). The lexicographic treatment of paradigmatic structures, as one major type of sense relations, will be the focus of this paper.

Judging by the relatively large number of dictionaries that cover paradigmatic items (pairs, triplets, or more complex word sets), dictionary users have a strong interest in this type of information. Such dictionaries are consulted in specific situations of text production when a user searches for alternative expressions in order to specify, to generalize or simply to vary in style or register (cf. Wiegand 2004: 36). However, in many monolingual German dictionaries the description of paradigmatic relations is often problematic and limited to a few types, such as synonymy and antonymy, and their presentation is inadequate.

Paradigmatic patterns can illustrate specific semantic choices of a lexical item within a context, and their investigation can help to detect particularities of word meanings. A dictionary that aims at describing the meaning and the use of a lexical item should also include a semantic description of paradigmatic contextual partners, not only to illustrate the semantic identity of a lexical item but also to demonstrate the interdependency of words. As Hanks (1990: 35) argues:

> […] there is a tendency for human lexicographers to focus on the way words are used to describe the world rather then on the way words interrelate with one another.

With the availability of large computer corpora, paradigmatic contextual choices can be studied empirically, revealing selectional preferences and contextual constraints and conditions. Although corpora offer fundamental methodological advantages, corpus-assisted approaches have, thus far, not played a central part in extracting and describing paradigmatic relations in German lexicography.

Elexiko is a relatively new lexicographic project based at the *Institut für Deutsche Sprache* in Mannheim (IDS) which aims to explain and document German and its present-day usage (cf. Haß-Zumkehr 2004, Storjohann 2005, and http://www.elexiko.de) including a detailed paradigmatic description of each lexical item. This electronic dictionary offers a differentiated presentation of sense relations and uses various corpus approaches to retrieve the necessary data.

First, I will briefly outline the types of sense relations that are of interest to *elexiko*. Attention is then turned to the principal objective of this paper. I will explore how the required data for the paradigmatic description of a word is elicited from the corpus using a variety of methods. Finally, I will demonstrate how sense relations are presented lexicographically in *elexiko*.

2. The System of Paradigmatic Relations

The specificity of a lexeme's meaning in context can vary enormously. Following a contextual approach this meaning reveals itself through contextual relations. In order to account for a detailed description of the meaning and use of a word, lexical patterns, such as manifested paradigmatic sense relations, need to be examined. In *elexiko*, the illustration of paradigmatic patterns is part of the semantic description of a lexeme comprising the comprehensive demonstration of the horizontal and vertical relations which exist between the senses of lexical items (cf. lexical units in Cruse 1986: 84).

These concern relations of inclusion and identity, as well as relations of exclusion and opposition. *Elexiko* has primarily adopted a classification following that offered by Cruse (1986) and by Lutzeier (1981), and this can be summarized as follows:

Table 1: Horizontal structures vs. vertical structures (adapted from Cruse (1986).

horizontal structures		**vertical structures**
incompatibility	antonymy	hyperonymy
	complementarity	hyponymy
	converseness	holonymy
	reversiveness	meronymy
synonymy		

The major differences between this classification and paradigmatic categories in other existing German dictionaries (e.g. DUDEN 8, DUDEN WUG, WSA, WGDS, DORNSEIFF) concern the detailed distinction of terms of exclusion. The relations of contrast and opposition, of which incompatibility is the most general sense relation, are divided into four categories. Whereas in other dictionaries the main relation of opposites is defined as antonymy, in *elexiko* (following Cruse 1986) this relation is a special case of incompatibility that is restricted to semantically gradable adjectives. Complementarity, converseness, and reversiveness are also specific sense relations of opposition and subtypes of incompatibility.

Within vertical patterns, lexical relations are separated into hyponymy/hyperonymy and meronymy/holonymy. More precise definitions of individual relations, including specific types and subgroups, can be found in Cruse (1986 and 2004). Synonymy in particular is not further subclassified in *elexiko*, but is used to refer to all types of semantic identity, ranging from absolute sameness and propositional identity to more vague categories such as near-synonymy.

3. Corpus Retrieval of Sense Relations

As far as the lexicographic process of describing lexemes and their uses is concerned, the corpus is primarily being used exploratorily. Instances of natural language are studied in order to identify rules and patterns, and linguistic proto-typicalities are then interpreted and classified. Finding copious illustrative text samples is only a by-product of corpus-aided analysis. Besides an extensive and maximally representative corpus serving as an empirical basis, the lexicographic process of obtaining paradigmatic sense relations requires a good corpus query tool assisting the search of the corpus and processing data.

> Computers do not get bored; they notice only what they are told to notice; and they notice every occurrence of the word or usage pattern in the corpus that they have been told to notice, no matter how many there may be. Only a large corpus of natural language enables us to identify recurring patterns in the language and to observe collocational and lexical restrictions accurately. (Hanks 1990: 36)

However balanced the underlying corpus might be and however well the necessary software to search and analyse language data might work, another crucial prerequisite of good lexicographic work is the linguistic competency of data interpreting. Language data used for our lexicographic interpretation is retrieved exclusively from the *elexiko*-corpus, a monitor corpus currently comprising about

1,300 million words. For the extraction of paradigmatic partners, both the corpus-driven and the corpus-based approaches are applied (cf. Sinclair 1996 and Tognini-Bonelli 2001), as in practice, it was observed that an interplay of both methodologies can have substantial benefits for the retrieval of this type of sense relation.

3.1. The Corpus-driven approach

The corpus-driven approach (CDA) is a methodology whereby the corpus serves as an empirical basis from which lexicographers extract their data and detect linguistic phenomena without prior assumptions and expectations (Tognini-Bonelli 2001). Any conclusions or claims are made exclusively on the basis of corpus observations. CDA proves indispensable since it provides information on significant and typical sense relations.

There might be a large number of potentially meaningful patterns that escape the attention of the traditional linguist; these will not be recorded in traditional reference works and may not even be recognised until they are forced upon the corpus analyst by the sheer visual presence of the emerging patterns in a concordance page (Tognini-Bonelli 2001: 86).

Although one can derive valuable results from CDA, in a number of cases, it cannot provide a comprehensive description of paradigmatic structures. Here, the corpus-based approach is used complementarily.

3.2. The Corpus-based approach

The corpus-based approach (CBA) is a method that uses an underlying corpus as an inventory of language data. From this repository, appropriate material is extracted to support intuitive knowledge, to verify expectations, to allow linguistic phenomena to be quantified, and to find proof for existing theories or to retrieve illustrative samples. It is a method where the corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as additional supporting material.

In this case, however, corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out

> according to pre-existing categories. Although it is used to refine such categories, it is never really in a position to challenge them as there is no claim made that they arise directly from the data (Tognini-Bonelli 2001: 66).

CBA offers an additional, complementary method of tracing paradigmatic pairs. The corpus-based approach implies a specific corpus inspection, where the lexicographer has a specific paradigmatic word in mind and searches the corpus for samples to either invalidate or verify and quantify the assumption. With the help of introspective expectation, through the collation of existing dictionaries and the use of specific search options, valuable evidence can be elicited from the corpus and incorporated into the paradigmatic description.

To understand the distinction between CBA and CDA, consider the following points:

- Corpus-based approach: theories are conceived and then proofed against corpora.
- Corpus-based linguists tend to use annotated corpora.
- Corpus-driven approach: theories are drawn to explain the existing data from corpora.
- Corpus-driven linguists tend to use raw corpora.

4. Characteristics of the corpus approach

The Corpus approach contains four major characteristics

a- It is empirical, analyzing the actual patterns of language use in natural texts.

b- It utilizes a large and principled collection of natural texts as the basis for analysis.

c- It makes extensive use of computers for analysis.

d- It depends on both quantitative and qualitative analytical techniques.

a- The corpus approach is empirical, analyzing the actual patterns of language use in natural texts.

The core of this pattern of the corpus approach is authentic language. The idea that corpora are principled has been mentioned, but not what language a corpus is comprised of. Corpora are comprised of textbooks, fiction, nonfiction, magazines, academic papers, world literature, newspapers, telephone conversations, business meetings, class lectures, radio broadcasts, and TV shows, among other communication acts. In short, any real-life situation in which any linguistic communication takes place can constitute a corpus.

b- The corpus approach utilizes a large and principled collection of natural texts as the basis for analysis

This characteristic of the corpus approach involves the corpus itself. You may work with a written corpus, a spoken corpus, an academic spoken corpus, etc.

c- The corpus approach makes extensive use of computers for analysis

Computers not only hold corpora, but they also help analyze the language in a corpus. A corpus is accessed and analyzed by a concordance programme. In short, you cannot effectively use corpora, or employ the corpus approach, without a computer.

d- The corpus approach depends on both quantitative and qualitative analytical techniques

This feature of the corpus approach highlights the significance of our intuition as expert users of a language. We take the quantitative results generated from the corpus and then analyze them qualitatively to find significance.

1. Target Features

Despite intuition may not always be reliable for drawing conclusions about language in general, it often gives an answer to the question ‘why’. Intuition is often useful for aiding us from queries for a corpus. Many of the questions that corpora reply fall into certain areas of language teaching, such as phraseology, lexicogrammar, registers, English for Specific Purposes (ESP) and appropriate syllabus design.

a- **Phraseology:** is the study of phrases and is considered as a major element of corpus linguistics. Sinclair (1991) noted that a meaning of a word is found via various lexical items in a sequence, through phrases. Phraseology encompasses the study of collocations, lexical bundles, and language occurring in preferred sequences.

Collocation: is the statistical tendency of words to co-occur. This means that when a word is used, there is a high statistical probability that a given word or words will take place alongside of it. For instance, when looking at the noun form of the word *deal* , the words *big, good, and great* are collocations of *deal* as a noun, we often refer to a big deal, a good deal, and/ or a great deal. A big deal is usually an event or situation that has significant meaning; a good deal generally refers to a bargain; a good deal often refers to a quantity. Studying collocations gives a more profound comprehension of the meaning and use of a word, such as deal, than plainly studying a word alone.

Lexical bundles: Phraseology also looks at variation in somewhat fixed phrases, which are often referred to as lexical bundles. Biber et al (1999: 990) define a lexical bundle as a recurring sequence of three or more words. In conversation, 'Do you want me to' and 'I don't know what' are among the most common lexical bundles (ibid: 994). It is important to understand that lexical bundles differ from idioms. Unlike lexical bundles, idioms have a meaning not derivable from their parts.

Preferred sequences: Phraseology also embraces the study of preferred sequences of words. Hunston (2002: 9-11) explains that students often confuse the words *interesting* and *interested*, and explanations of their distinct meanings do not usually help learners use the words accurately. Looking at the phrases *someone is interested in something, an interesting thing, what is interesting and it is* *interesting to see,* can provide learners the capacity to use the individual words effectively by providing an established pattern of use for each word.

b- **Lexicogrammar:** Lexicogrammar is Sinclair's (1991) idea that there is no disparity between lexis and grammar. An example of this idea includes certain words (lexicon) tied up with certain tenses (grammar): *know, matter and suppose* occur more than 80 percent of the time in the present tense while *smile, reply and pause* occur more than 80 percent in the past tense (Biber et al, 1999: 459).

c- **Register:** Register is defined as situation of use. We use distinct language with distinct audiences, our parents, children or colleagues at distinct times and distinct reasons. Corpus linguistics addresses language teaching through the study of register by examplifying the several phraseology and lexicogrammar used from register to register. For instance, 90 percent of lexical bundles in conversation are declarative or interrogative clauses (Biber et al, 1999: 999); pronouns are used slightly more in conversation than nouns, but nouns are used significantly more than pronouns in fiction, news and academic writing (Biber et al, 1999: 235).

d- **English for Specific Purposes (ESP):** ESP is probably one of the most evident and pointed applications of corpus linguistics. The areas of register, lexicogrammar and phraseology can all be applied to specific purposes. The Academic World List (AWL) is a well-known example of using corpus linguistics to address ESP. Corpora are available for nurses and health care professionals, air traffic controllers, just to name a few.

e- **Syllabus design:** The last area of language teaching that corpus linguistics tackles is syllabus design. Phraseology, lexicogrammar, register, ESP. These areas can be used to more effectively and correctly design syllabi by helping us see what learners really need to know about language, frequency and collocation for vocabulary, grammar patterns for different registers, and specific knowledge for specific purposes. As a teacher, you can supplement course materials with information that is relevant for students.

References/ Further readings

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

Sinclair, J. (1991a) Corpus, Concordance and Collocation. Oxford: Oxford University Press.

ICL/ M2

Dr. HAMZAOUI

6. Key Considerations for Building and Designing a Corpus

Course contents: Introduction- What are the basics for building a corpus?- How do I collect texts?- How much mark-up do I need?- Issues in designing a spoken corpus- Conclusion

Introduction

If you have decided that using a corpus will help you with your research but that no corpus already exists which is suitable for your purposes, you will need to design your own corpus. In order to build a corpus there are a number of factors which need to be taken into consideration.

1. What are the basics for building a corpus?

A corpus is essential when exploring issues or questions related to language use. Each year, the number of corpora that are available for researchers to use is increasing. Therefore, before tackling the task of building a corpus, one has to be sure that there is not an existing corpus that meets his/her needs. Each day, more and more corpora of different languages are becoming available on the web. Having a clearly articulated question is an essential first step in corpus construction since this will guide the design of the corpus. Before designing a corpus, one has to consider several factors. These incorporate representativeness, size, sampling and balance which will be discussed below.

a- ***Representativeness:*** The corpus must be representative of the language being investigated. If the goal is to describe the language of newspaper editorials, collecting personal letters would not be representative of the language of newspaper editorials. There must be a match between the language being examined and the type of material being collected (Biber 1993).

A corpus can be said to be representative if the findings from that corpus are generalisable to language or a particular aspect of language as a whole. Obviously, it is not possible to collect an entire language to test the representativeness of a corpus. Ultimately, building corpora is about collecting texts. Thinking carefully about the components of a corpus helps to decide what sorts of texts constitute a representation of the particular language variety under investigation. This, in turn, helps with the overall probable representativeness of the corpus.

b- ***Size:*** Here, we have to take into consideration the following relevant question: What kind of data do I use and how much? The question of corpus size is a difficult one. There is not a specific number of words that answers this question. Corpus size is certainly not a case of one size fits all. For explorations that are designed to capture all the senses of a particular word or set of words, as in building a dictionary, then the corpus needs to be large, very large – tens or hundreds of millions of words. However, for most questions that are pursued by corpus researchers, the question of size is resolved by two factors: representativeness (have I collected enough texts (words) to accurately represent the type of language under investigation?) and practicality (time constraints). For example, it is possible to capture all the works of a particular author, or historical texts from a certain period, or texts from a particular event (e.g. a radio or TV series, political speeches). In these cases, complete representation of the language can be achieved. An example of this is the 604,767-word corpus of nine seasons of the popular television sitcom Friends (Quaglio 2008).

However, it is possible to get much useful data from a small corpus, particularly when investigating high frequency items. In fact, this may be desirable to do this rather than being overwhelmed by too much data from a big corpus.

You may also be constrained by more practical considerations. If you need to transcribe spoken data with a high degree of detail, then it may only be feasible to work with thousands rather than millions of words. With written texts, you may be limited by what you can obtain permission for from the copyright holder.

c- Sampling: Since it is practically impossible to investigate entire language varieties, corpus building will inevitably involve gathering together enough samples of that language variety (i.e. texts) to adequately represent it. Sampling, then, is crucially important in corpus building. Narrowing down which texts you need to collect in order to represent the language variety you are investigating could start with some simple questions, such as:

□ What language or language variety is being studied? (e.g. English, Australian English, Yorkshire English.)

□ What is the location of the texts you need to collect? (e.g. UK, Australia, Yorkshire.)

□ What is the production/publishing date of the texts you want? (This might be one day, one year, or a span of years)

These questions are quite broad, but need careful consideration. Further questions you might ask to help you specify the sorts of texts you want to collect could include:

□ What is the mode of the texts? (spoken or written, or both)

□ What sorts of texts (text-types) need to be included? (e.g. newspaper articles, short stories, letters, text messages)

□ What domain will the texts come from ? (e.g. academic, legal, business)

d- Balance: If you are collecting data for a spoken corpus, it may be necessary to carefully consider the types of people you use as informants. This will allow you to decide if your data balanced in terms of the gender, class, age, ethnic background, etc. of the participants and thus how representative any claims you might make will be of the wider population. Getting this balance right is not an exact science and there are no reliable ways of determining whether a corpus is truly balanced. One approach to achieving balance is to use an existing corpus as a model.

2. How do I collect texts?

Once a research question is articulated, corpus construction can begin. The next task is identifying the texts and developing a plan for text collection. In all cases, before collecting texts, it is important to have permission to collect them. When collecting texts from people or institutions, it is essential to get consent from the parties involved. The rules that apply vary by country, institution and setting, so be sure to check before beginning collection. There are texts that are considered public domain. These texts are available for research and permission is not needed.

When creating a corpus there are certain procedures that are followed, regardless of whether the corpus is representing spoken or written language. Some issues that are best addressed prior to corpus construction include: What constitutes a text? How will the files be named? What information will be included in each file? How will the texts be stored (file format)?

3. How much mark-up do I need?

The term ‘mark-up’ refers to adding information to a corpus file. Not all corpora contain mark-up; however, certain types of mark-up can facilitate corpus analysis. Mark-up can be divided into two types: document mark-up and annotations. Document mark-up refers to markings such as paragraphs, fonts, sentences, including sentence numbers, speaker identification, and marking the end of the text.

Annotation refers to all the extra information that is added to the texts in order to aid the researcher to retrieve as much relevant information as possible. The most common form of corpus annotation involves including parts of speech (POS) tags which label each word in a corpus as to its grammatical category (e.g. noun, adjective, adverb, etc.). These tags can be very useful for addressing a number of questions and help to resolve many of the issues related to simply searching on a particular word.

In corpus linguistics, part-of-speech tagging also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context, i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

4. Issues in designing a spoken corpus

Unlike writing, the nature of spoken discourse means that it is subject to the observer's paradox. If the goal of your research is to collect what might be described as 'natural' or 'real' data, there are a number of issues to consider. Obviously, the best way to get this kind of data is for the participants in your study to be unaware that they are being recorded. Surreptitious recording has been used by corpus linguists in the past but is now regarded as unethical at best and, in some circumstances, may well be illegal.

Instead, the compilers of many recent corpora have asked contributors to wear lapel microphones and carry recorders around with them for a part of their day. The data captured using this method is often characterised by questions regarding the microphone from other interlocutors at the outset of conversations, but these do not last long and conversations tend to proceed as normal.

5. Points to consider when conducting corpus-linguistic research

- Make sure you have enough time to conduct your corpus-linguistic research! Don't start two or three days before your actual presentation – you should be finished by then! Depending on your topic/research question, you'll need two or three weeks to analyse your features.

- Choose your corpus/corpora carefully: a large corpus is usually suitable for any kind of linguistic research (1,000,000 words or more), while a small corpus (200,000 to 500,000 words) may only be sufficient for frequent syntactic structures such as the present perfect or the analysis of the more common modal verbs.

- Get to know your corpus/corpora: text types, size, language variety, etc.

- If you compare two or more different corpora, e.g. the German and Swedish *ICLE* sub-corpora, be aware that each sub-corpus may consist of a different number of words (e.g. 265,341 words in the German *ICLE*, 248,578 words in the Swedish *ICLE*). When you present your corpus-linguistic results, you have to make sure that your figures are comparable. It is no use saying that feature X occurred 5 times in the German *ICLE* and 5 times in the Swedish *ICLE*, if the total numbers of words differ in the two corpora – you have to have a common basis. You can solve this problem by extrapolating your figures to a common denominator.

  A frequently used common denominator in corpus-linguistic research is 1 million words, but you can also use other figures, e.g. 250,000 words. This is how you calculate the extrapolation: Multiply the feature you counted in corpus A by 1,000,000, then divide this figure by the actual size of corpus A. E.g. feature X occurred 5 times in 265,341 words in the German *ICLE* à 5 multiplied by 1,000,000 = 5,000,000 divided by 265,341 = 18.84 occurrences of feature X in 1 million words. You then do the same calculation with the results from the Swedish *ICLE*: 5 multiplied by 1,000,000 divided by 258,978 = 20.11 occurrences of feature X in 1 million words. Although the differences between the two corpora used here are only minimal you still have to do the extrapolation. Otherwise you would be comparing apples and oranges!

- Be careful to find all occurrences of your feature! If, for example, you search for the collocation make a decision, your search strategy has to be such that you find all inflectional variants of MAKE (mak* would give you *make, makes, making* but not *made*. However, it also gives you *maker*. *Ma** would give you also *made* but then you are faced with any word starting in *ma-*, such as *man, mankind, mad, Mary*, etc.) Also, DECISION might be pre-modified by an adjective such as *useful* or *personal* which you might want to include in your analysis as well, so make sure you don't forget these examples during your search (e.g. by using the search string *a * decision*).

- Not all concordance lines need to be relevant for your research. If you search for the phrasal verb make up, you will find a number of nominal or adjectival uses of this phrasal verb, such as "She put on her make up" or "Her beautifully made up face". In WordSmith, you can discard such unwanted concordance lines by highlighting them, then pressing delete. When you have marked all unwanted examples in your concordance in this way, you use the "zap" function so that the unwanted examples are discarded and you are left only with those occurrences you actually need.

- A high frequency of your researched feature does not necessarily mean that your feature is distributed evenly across the entire corpus you used. Check the corpus' file names in order to exclude that maybe only one or two authors or speakers produced all the examples you have found.

- Make sure you don't over-generalise your results. If, for example, you used a very small corpus of written academic American English, you mustn't claim that your results are valid for American English as a whole or even for English in general. Qualify your research results by saying that your results hold only as far as written academic American English is concerned and that further research into other types of English needs to be conducted for more general conclusions about the features you researched.

Conclusion

As can be seen from this course, a corpus can serve as a useful tool for discovering many aspects of language use that otherwise may go unnoticed provided that we consider the number of factors before compiling it. Unlike straightforward grammaticality judgments, when we are asked to reflect on language use, our recall and intuitions about language often are not accurate. Therefore, a corpus is essential when exploring issues or questions related to language use. The wide range of questions related to language use that can be addressed through a corpus is strength of this approach. Questions that range from the level of words and intonation to how constellations of linguistic features work together in discourse can all be explored through the lens of corpus linguistics.

References/ Further reading

Biber, D. (1990) 'Methodological Issues Regarding Corpus-Based Analysis of Linguistic Variation', Literary and Linguistic Computing 5(4): 257–69.

Biber, D., Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Exploring Language Structure and Use. Cambridge: Cambridge University Press.

McEnery, T, Xiao, R. and Tono, Y. (2006) Corpus Based Language Studies: An Advanced Resource Book. London: Routledge.

Leech, G. (2005) 'Adding Linguistic Annotation', in M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbrow Books, pp. 17–29; also at http://ahds.ac.uk/linguistic-corpora/

www.helsinki.fi/varieng/CoRD/corpora/index.html

www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite.../Skript_02.pdf

ICL/M2

Dr. HAMZAOUI

7. A step-by-step guide to building a corpus

The following represents a possible set of steps for building a corpus:

1. The first step when building a corpus should be the corpus design. Thinking carefully about sampling and representativeness issues will help to build a good corpus, as well as save valuable time and effort.

2. Decide on what your corpus is attempting to represent and, therefore, what will be in it.

3. Use your answer to the above to create a hierarchical model of the components of the corpus.

4. For each of the components at the bottom level of the hierarchy, list the possible text-types or texts (depending on the detail of your hierarchy) that you expect to find.

5. Consider whether each text-type should have equal standing in the corpus.

6. Think about the size of each component, taking into consideration the number of text-types available, their real-world importance, and the practical issues in gathering them.

7. Decide whether whole texts will be collected, or extracts, or both. If you are collecting extracts, where possible use random sampling. This can be achieved by using a random number generator (there are many available on the internet) and employing those numbers to select, for example, page numbers. Whatever

strategy you decide upon, keep in mind that lexical choice can be influenced by textual position.

8. Choose, locate and gather the texts that will populate the corpus. This can be a very time consuming step, depending on the nature of the texts being collected. For instance, gathering spoken texts will require a number of further steps including making audio recordings of the spoken data, and then transcribing them.

9. Make and keep a security copy of the text in its original form. This copy might be electronic, paper or audio/visual, depending on the nature of the text(s) being collected.

10. Make and keep an electronic, plain text format copy of the text. This format is the simplest electronic format and, at the moment, a requirement of most corpus-tools. It is also the most portable, flexible and future-proof format, so a good bet for archive-copies of the corpus. If required, the plain text copies can be (usually very easily) converted into other formats.

11. Add some information about the text at the beginning of the text file. Sinclair (2005) suggests that the easiest way to do this is to simply add an identification code or serial number, which can then be cross-referenced with a database or spread-sheet that holds useful information about the text (i.e. metadata). This might include details such as the author, date of publication, genre, page-numbers sampled, and so on. Another option is to include this sort of information actually in the text file in what is known as a header.

12. Add further annotation or mark-up required by the investigation (for example, part-of-speech annotation). Keep copies of the corpus before the annotation is applied, and make copies afterwards.

13. Once you have a working copy of the corpus, MAKE A COPY OF IT! Ideally, you should make copies along the way as you build the corpus. This means that you can always go back if something goes wrong at some stage.

14. Keep notes of what you do and why you are doing it. Record decisions about corpus design, the data you use, where you got the data from, any annotation you apply to it, and so on. These notes will act as an *aide memoire*, and help others see and understand what you did.

15. Do some analysis!

16. If you need to, repeat any of the above steps in light of your analysis.

ICL/M2

Dr. HAMZAOUI

8. Advantages of a Corpus

Course contents: Introduction- The kinds of things that a corpus can aid us with- Translation- Stylistics- Language and ideology-Advantages of a corpus- Conclusion

Introduction

"It is no exaggeration to say that corpora, and the study of corpora, have revolutionised the study of language, over the last few decades." (Hunston 2002: 1). Even if you have never used a corpus before, it is increasingly likely that you have used dictionaries and grammar books which were written using information derived from corpora as their bases, especially if English is not your first language. The following course looks at how corpora have been used to enhance understanding in three areas: translation, stylistics, and language and ideology.

1. The kinds of things that a corpus can aid us with

They are listed as follows:

a- Translation

On a practical level, a parallel corpus can be used by a translator to look at a number of alternatives for a particular term and aid in the solution of a translation problem. A parallel corpus is a richer resource than a bilingual dictionary as it allows the user to see the search term with more of the co-text and with a broader range of contexts and collocates. This in turn shows the translator a wide range of possible renderings: from the ‘zero’ option, where something has been missed out by the translator, possibly for pragmatic reasons, to a phrase which differs a great deal in terms of lexical equivalence but retains the semantic content of the original.

On a more theoretical level it is possible to compare a corpus of texts translated into a language with those originally written in that language. Studies of this nature have shown how original and translated texts differ in particular ways. For example, Laviosa (1997: 315 see Hunston 2002: 127) has shown how translations are often less lexically varied than their ‘original’ equivalents and McEnery et al (2006: 93) demonstrate that “… the frequency of aspect markers in Chinese translations is significantly lower than that in the comparable L1 Chinese data.” This information may be useful for those who study how translators work, or who are involved in the training of translators to help their students to avoid ‘translationese’ creeping into their work.

b- Stylistics

There are a number of ways in which corpus-based approaches can contribute to the study of not just literary works but ‘literariness’ in general. The statistical analysis of literary texts, known as Stylometrics, has been used to establish authorship of contested texts. As has been mentioned before, a smaller corpus of literary texts can be compared with a reference corpus to investigate literary ‘devices’ to see how they vary from more ‘everyday’ varieties of English.

Louw (1997: 245) demonstrates how students can confirm their intuitions about literary texts using corpus data. In one example, students investigated the term *wielding a* in order to confirm that the use of this term in the line ‘And crawling sideburns, wielding a guitar’ from the poem *Elvis Presley* by Thom Gunn was being used ironically. As expected they found that *wielding a* is most frequently used with some kind of weapon. What was unexpected was the very high frequency with which the term is used ironically, prompting one student to comment that it may soon lose its power as a writer’s device.

c- Language and ideology

There is an increasing interest in using corpora to investigate the ideological stance of writers and speakers in texts. Frequently occurring patterns allow the observer to make deductions about what a group or society sees as valuable or important. Information about collocation means that new concepts and the range of associations of a word can be monitored. Stubbs (1996: 195) argues that if a collocation becomes more common in the language then it is more likely to become fixed in the minds of speakers and therefore, more difficult to challenge. As we saw with stylistics, semantic prosody, the semantic associations of a word or phrase, can be used to carry covert messages.

Studies in this area have covered a wide variety of areas such as sexism and racism in media discourse, Euroscepticism, political correctness and the difference in rhetorical styles of Bush and Blair in relation to the war in Iraq (see McEnery et al 2006: 108-113). Hunston (2002: 121) points out some of the assumptions that such studies can be based on. O'Halloran and Coffin (2004) argue that using corpora can actually help the researcher to avoid over- and under-interpretation when working with texts. While caution should be exercised regarding the verifiability of claims about ideology found in corpora, they remain valuable resource in such studies.

2. Benefits of a corpus

First, the corpus gives the linguist an empirical data which permits him to form objective rather than subjective linguistic statements. Second, corpus linguistics helps the researcher to get rid of any linguistic generalizations that may be based upon his internalized cognitive perception of language. Third, qualitative and quantitative linguistic research can be conducted in few seconds owing to the powerful computers and software that are able to perform complex calculations without errors, thus saving both effort and time. Lastly, studying the language empirically can aid linguists not only to conduct new linguistic research adequately, but also to revise and test the existing theories.

Conclusion

A corpus is used for various reasons, like: objective verification of results, Corpora show how people really use the language. They do not provide imaginary or idealised examples. Quantitative data show what occurs frequently and what occurs rarely in the language.Thank to IT-technology, we can conduct fast, complex studies and process more material than by hand.

References/ Further reading

Louw, B. (1997) The Role of Corpora in Critical Literary Appreciation in Wichmann, A., Fligelston, S., McEnery, T. & Knowles, G. (eds) *Teaching and Language Corpora* Harlow: Longman

McEnery, T, & Wilson, A. (1996) *Corpus Linguistics* Edinburgh: Edinburgh University Press

McEnery, T., Xiao, R. & Tono, Y. (2006) *Corpus-Based Language Studies* Abingdon: Routledge

Meyer, C. (2002) *English Corpus Linguistics* Cambridge: Cambridge University Press

Stubbs, M. (1996) *Text and Corpus Analysis* Oxford: Blackwell

ICL/M2

Dr. HAMZAOUI

9. Types and Aims of Corpora

Course Contents: Introduction- General corpora- Specialized corpora-Comparable corpora-Parallel corpora- Historical corpora- Monitor corpora

Introduction

A corpus is principled because texts are selected for inclusion according to pre-defined research purposes. Usually texts are included on external rather than internal criteria. For example, a researcher who wants to investigate metaphors used in university lectures will attempt to collect a representative sample of lectures across a number of disciplines, rather than attempting to collect lectures that include a lot of figurative language. Most commercially available corpora are made up of samples of a particular language variety which aim to be representative of that variety.

1. Types of corpora

Here are some examples of some of the different types of corpora and how they represent a particular variety:

1.1. General corpora

An example of a general corpus is the British National Corpus which "… aims to represent the universe of contemporary British English [and] to capture the full range of varieties of language use." (Aston & Burnard 1998: 5). As a result of this aim the corpus is very large (containing some 100 million words) and contains a balance of texts from a wide variety of different domains of spoken and written language. Large general corpora are sometimes referred to as reference corpora because they are often used as a baseline against which judgements about the language varieties held in more specialised corpora can be made.

1.2. Specialized corpora

Specialised corpora contain texts from a particular genre or register or a specific time or context. They may contain a sample of this type of text or, if the dataset is finite and of a manageable size, for example all of Shakespeare's plays, be complete. There are numerous examples of specialised corpora; these include The Michigan Corpus of Spoken English (approximately 1.7 million words of spoken data collected from a variety of different encounters at the University of Michigan), the International Corpus of Learner English (20,000 words taken from essays of students learning English as a foreign language) and the Nottingham Health Communication Corpus (see section 5.3 for more details)

1.3. Comparable corpora

Two or more corpora constructed along similar parameters but each containing a different language or a different variety of the same language can be regarded as comparable corpora. An example of this type is the CorTec Corpus which contains examples of technical language in texts from five areas in both English and Portuguese.

1.4 Parallel corpora

These are similar to comparable corpora in that they hold two or more collections of texts in different languages. The main difference lies in the fact that they have been aligned so that the user can view all the examples of a particular search term in one language and all the translation equivalents in a second language. The Arabic English Parallel News Corpus contains 2 million words of news stories in Arabic and their English translation collected between 2001 and 2004, and is aligned at sentence level.

1.5 Historical (or diachronic) corpora

In order to study how language changes over time texts from different time periods can be assembled as a historical corpus. Two examples of this type are the Helsinki Diachronic Corpus of English Texts (containing 1.5 million words written between 700 and 1700) and the ARCHER (A Representative Corpus of Historical English Registers) corpus (1.7 million words covering the years 1650 to 1990).

1.6 Monitor corpora

A monitor corpus is one that is 'topped up' with new texts on a regular basis. This is done in such a way that "… the proportion of text types remains constant …" which means that each new version of the corpus is comparable with all previous versions. (Hunston 2002: 16). The best example of this type is the Bank of English, held at the University of Birmingham.

2. What other corpora are there?

There are many types of corpora, which can be used for different kinds of analyses (cf. Kennedy 1998). Some (not necessarily mutually exclusive) examples of corpus types are (for a description of the individual corpora see below):

- general/reference corpora (vs. specialized corpora) (e.g. BNC = British National Corpus, or Bank of English): aim at representing a language or variety as a whole (contain both spoken and written language, different text types etc.)

- historical corpora (vs. corpora of present-day language) (e.g. Helsinki Corpus, ARCHER) aim at representing an earlier stage or earlier stages of a language

- regional corpora (vs. corpora containing more than one variety) (e.g. WCNZE = Wellington Corpus of Written New Zealand English) aim at representing one regional variety of a language - learner corpora (vs. native speaker corpora) (e.g. ICLE = International Corpus of Learner English) aim at representing the language as produced by learners of this language

- multilingual corpora (vs. one-language corpora) aim at representing several, at least two, different languages, often with the same text types (for contrastive analyses) - spoken (vs. written vs. mixed corpora) (e.g. LLC = London-Lund Corpus of Spoken English) aim at representing spoken language A further distinction of corpus types refers not to the texts that have been included in the corpus, but to the way in which these texts have been treated:

- annotated corpora (vs. orthographic copora) in annotated corpora, some kind of linguistic analysis has already been performed on the texts, such as sentence analysis, or, more commonly, word class classification

3. The sorts of things that a corpus can help you with

"It is no exaggeration to say that corpora, and the study of corpora, have revolutionised the study of language, over the last few decades." (Hunston 2002: 1) The following section outlines some of the areas where corpora have had an impact. The intention is to help you to see whether corpus analysis techniques may be useful to you in your research. Even if you have never used a corpus before, it is increasingly likely that you have used dictionaries and grammar books which were written using information derived from corpora as their bases, especially if English is not your first language.

The following section looks at how corpora have been used to enhance understanding in three areas: translation, stylistics and language and ideology. The parallel and comparable corpora that were mentioned in section 1 can be used for both practical and theoretical translation studies. On a practical level, a parallel corpus can be used by a translator to look at a number of alternatives for a particular term and aid in the solution of a translation problem. A parallel corpus is a richer resource than a bilingual dictionary as it allows the user to see the search term with more of the co-text and with a broader range of contexts and collocates.

This in turn shows the translator a wide range of possible renderings: from the ‘zero’ option, where something has been missed out by the translator, possibly for pragmatic reasons, to a phrase which differs a great deal in terms of lexical equivalence but retains the semantic content of the original. On a more theoretical level it is possible to compare a corpus of texts translated into a language with those originally written in that language. Studies of this nature have shown how original and translated texts differ in particular ways.

For example, Laviosa (1997: 315 see Hunston 2002: 127) has shown how translations are often less lexically varied than their ‘original’ equivalents and McEnery et al (2006: 93) demonstrate that “… the frequency of aspect markers in Chinese translations is significantly lower than that in the comparable L1 Chinese data.” This information may be useful for those who study how translators work, or who are involved in the training of translators to help their students to avoid ‘translationese’ creeping into their work.

There are a number of ways in which corpus-based approaches can contribute to the study of not just literary works but ‘literariness’ in general. The statistical analysis of literary texts, known as Stylometrics, has been used to establish authorship of contested texts. As has been mentioned before, a smaller corpus of literary texts can be compared with a reference corpus to investigate literary ‘devices’ to see how they vary from more ‘everyday’ varieties of English. Louw (1997: 245) demonstrates how students can confirm their intuitions about literary texts using corpus data. In one example, students investigated the term wielding a in order to confirm that the use of this term in the line ‘And crawling sideburns, wielding a guitar' from the poem Elvis Presley by Thom Gunn was being used ironically.

As expected they found that wielding is most frequently used with some kind of weapon. What was unexpected was the very high frequency with which the term is used ironically, prompting one student to comment that it may soon lose its power as a writer's device. In language and ideology, there is an increasing interest in using corpora to investigate the ideological stance of writers and speakers in texts. Frequently occurring patterns allow the observer to make deductions about what a group or society sees as valuable or important. Information about collocation means that new concepts and the range of associations of a word can be monitored.

Stubbs (1996: 195) argues that if a collocation becomes more common in the language then it is more likely to become fixed in the minds of speakers and therefore, more difficult to challenge. As we saw with stylistics, semantic prosody, the semantic associations of a word or phrase, can be used to carry covert messages. Studies in this area have covered a wide variety of areas such as sexism and racism in media discourse, Euroscepticism, political correctness and the difference in rhetorical styles of Bush and Blair in relation to the war in Iraq (see McEnery et al 2006: 108-113).

Hunston (2002: 121) points out some of the assumptions that such studies can be based on. O'Halloran and Coffin (2004) argue that using corpora can actually help the researcher to avoid over- and under-interpretation when working with texts. While caution should be exercised regarding the verifiability of claims about ideology found in corpora, they remain valuable resource in such studies.

3. What you need to do corpus work

You can actually get started on some corpus work straight away, if you have internet access. There are corpora that you can browse (although not always in full) online. Examples include: MICASE [http://www.lsa.umich.edu/eli/micase/index.htm]BNC[http://www.natcorp.ox.ac.uk/] Business Letter Corpus [http://ysomeya.hp.infoseek.co.jp/] Or you can make concordances from the World Wide Web using the tools that can be found at these sites: WebCorp [http://www.webcorp.org.uk/] WebCONC [http://www.niederlandistik.fu-berlin.de/cgi-bin/webconc.cgi?art=google&sprache=en] If you intend to install some corpus investigation software onto a computer then the more RAM and the faster the processor, the easier the computer will be able to handle the tasks you might ask of it. Much of the software that has been developed thus far has been written for use with Windows operating systems.

Conclusion

This lecture has presented the different types of corpora and their major aims. However, the list remains long and open since there are always new types of corpora that are created to facilitate the task of analyzing data to researchers.

References/further reading

Aston, G. & Burnard, L. (1998) *The BNC Handbook* Edinburgh: Edinburgh University Press

Coffin, C. & O'Halloran, K. (2004) Checking Overinterpretation and Underinterpretation: Help from Corpora in Critical Linguistics in Coffin, C. Hewings, A. & O'Halloran, K. (eds.) *Applying English Grammar* London: Arnold

Hunston, S. (2002) *Corpora in Applied Linguistics* Cambridge: Cambridge University Press

Kennedy, G. (1998) *An Introduction to Corpus Linguistics* Harlow: Longman

Louw, B. (1997) The Role of Corpora in Critical Literary Appreciation in Wichmann, A., Fligelston, S., McEnery, T. & Knowles, G. (eds) *Teaching and Language Corpora* Harlow: Longman

McEnery, T, & Wilson, A. (1996) *Corpus Linguistics* Edinburgh: Edinburgh University Press

McEnery, T., Xiao, R. & Tono, Y. (2006) *Corpus-Based Language Studies* Abingdon: Routledge

Meyer, C. (2002) *English Corpus Linguistics* Cambridge: Cambridge University Press.

ICL/M2

Dr. CH.HAMZAOUI

10. Conceptual Classification of Corpora and Limitations of a Corpus

Introduction

Since electronic corpus is a new thing, we are yet to reach to a common consensus to what counts as a corpus, and how it should be classified. The classification scheme I propose here goes as far as it is prudent at the present moment. It offers a reasonable way to classify corpora, with clearly delimited categories wherever possible. Different criteria for classification are applied to corpora, sub-corpora, and their related components. Linguistic criteria may be external and internal. External criteria are largely mapped onto corpora from text typology concerned with participants, occasion, social setting, communicative function of language, etc. Internal criteria are concerned with recurrence of language patterns within the pieces of language. Taking all these issues under consideration I classify corpora in a broad scheme in the following manner: Genre of text, Nature of data, Type of text, Purpose of design, and Nature of application.

A- Conceptual Classification of Corpora

1 Genre of Text

- **Written Corpus:** A written corpus (e.g., TDIL Corpus) by virtue of its genre contains only language data collected from various written, printed, published and electronic sources.

- **Speech Corpus:** A speech corpus (e.g., Wellington Corpus of Spoken New Zealand English) contains all formal and informal discussions, debates, previously made talks, impromptu analysis, casual and normal talks, dialogues, monologues, various types of conversation, on line dictations, instant public addressing, etc. There is no scope of media involvement in such texts.

- **Spoken Corpus:** Spoken corpus (e.g., London-Lund Corpus of Spoken English), a technical extension of speech corpus, contains texts of spoken language. In such corpus, speech is represented in written form without change except transcription. It is annotated using a form of phonetic transcription.

2. Nature of Data

- **General Corpus:** General corpus (e.g., British National Corpus) comprises general texts belonging to different disciplines, genres, subject fields, and registers. Considering the nature of its form and utility, it is finite in number of text collection. That means, number of text types and number of words and sentences in it are limited. It has an opportunity to grow over time, and to append new data with availability of new texts. It is very large in size, rich in variety, wide and representation, and vast in utilisation scope.

- **Special Corpus:** Special corpus (e.g., CHILDES Database) is designed from texts sampled in general corpus for specific variety of language, dialect and subject with emphasis on certain properties of the topic under investigation. It varies in size and composition according to purpose. It does not contribute to the description of a language because it contains a high proportion of unusual features. Its origin is not reliable as it records the data from people not behaving normally. Special corpus is not balanced (except within the scope of its given purpose) and, if used for other purposes, gives distorted and 'skewed' view of language segments. It is different in principle, since it features one or other variety of normal, authentic language. Corpus of language of children, non-native speakers, users of dialects, and special areas of communication (e.g., auction, medical talks, gambling, court proceeding, etc.) are designated as special corpus because of their non-representative nature of the language involved. Its main advantage is that texts are selected in such a way that the phenomena one is looking for occur more frequently in it than in balanced corpus. A corpus that is enriched in such a way is smaller than a balanced corpus providing same type of data (Sinclair 1996b).

- **Sublanguage corpus:** It consists of only one text variety of a particular language. It is at the other end of the linguistic spectrum of a Reference corpus. The homogeneity of its structure and specialised lexicon allows the quantity of data to be small to demonstrate typically good and closure properties.

- **Sample corpus:** Sample corpus (e.g., Zurich Corpus of English Newspapers) is one of the categories of special corpus, which is made with samples containing finite collection of texts chosen with great care and studied in detail. Once a sample corpus is developed it is not added to or changed in any way (Sinclair 1991: 24) because any kind of change will imbalance its constitution and distort research requirement. Samples are small in number in relation to texts, and of constant size. Therefore, they do not qualify as texts.

- **Literary corpus:** A special category of sample corpus is literary corpus, of which there are many kinds. Classification criteria considered for generation of such corpus include author, genre (e.g., odes, short stories, fictions, etc.), period (e.g., 15th century, 18th century, etc.), group (e.g., Romantic poets, Augustan prose writers, Victorian novelists, etc.), theme (e.g., revolutionary writings, family narration, industrialisation, etc.) and other issues as valued parameters.

- **Monitor corpus:** Monitor corpus (e.g., Bank of English) is a growing, non-finite collection of texts with scope for constant augmentation of data reflecting changes in language. Constant growth of corpus reflects change in language, leaving untouched the relative weight of its components as defined by parameters. The same composition schema is followed year by year. The basis of monitor corpus is of reference to texts spoken or written in one single year (Sinclair 1991: 21). From monitor corpus we find new words, track variation in usage, observe change in meaning, establish long-term norm of frequency distribution, and derive wide range of lexical information. Over time the balance of components of a monitor corpus changes because new sources of data become available and some new procedures enable scarce material to become plentiful. The rate of flow is adjusted from time to time.

3. Type of Text

- **Monolingual corpus:** It (e.g., ISI Bengali Corpus) contains representative texts of a single language representing its use in a particular period or in multiple periods. It contains both written and spoken text samples so long their cohabitation and relational interface does not hamper proposed work of the investigators.

- **Bilingual corpus:** Bilingual corpus (e.g., TDIL Bengali-Oriya Corpus) is formed when corpora of two related or non-related languages are put into one frame. If these languages are genetically or typologically related they become parallel corpus (discussed below) where texts are aligned following some predefined parameters. Size, content, and field may vary from corpus to corpus, which is not permitted in case of parallel corpus.
- **Multilingual corpus:** Multilingual corpus (e.g., Crater Corpus) contains representative collections from more than two languages. Generally, here as well as in bilingual corpus, similar text categories and identical sampling procedures are followed although texts belong to different languages.

4. Purpose of Design

- **Un-annotated corpus:** It (e.g., TDIL Corpus) represents a simple raw state of plain texts without additional linguistic or non-linguistic information. It is of considerable use in language study, but utility of corpus is considerably increased by annotation.

- **Annotated corpus:** It (e.g., British National Corpus) contains tags and codes inserted from outside by designers to record some extra information (analytical marks, parts-of-speech marks, grammatical category information, etc.) into texts. In contrast to un-annotated corpus, annotated corpus is more suitable for providing relevant information useful in various tasks for language technology including morphological processing, sentence parsing, information retrieval, word sense disambiguation, machine translation, etc.

4. Nature of Application

- **Translation Corpora:** Translation corpora generally consist of original texts of source language and their translations taken from target language. These corpora usually keep meaning and function of words and phrases constant across languages, and as a consequence, offer an ideal basis for comparing realisation of particular meanings in two different languages under identical condition. Moreover, they make it possible to discover all cross-linguistic variants, i.e. alternative renderings of particular meanings and concepts. Thus, translation corpora provide more fruitful resources both for cross-linguistic data analysis and rule formulation necessary for translation (Altenberg and Aijmer 2000: 17).

- **Aligned corpus:** It is (e.g., The Canadian Hansard Corpus) a kind of bilingual corpus where texts in one language and their translations into other language are aligned, sentence by sentence, phrase by phrase, or even word by word.

- **Parallel corpus:** Parallel corpus (e.g., Chemnitz German-English Corpus) contains texts as and translations in each of the languages involved allowing double-checking translation equivalents. Texts in one language and their translations into another are aligned: sentence by sentence, phrase by phrase, or even word by word. Sometimes reciprocate parallel corpora are designed where corpora containing authentic texts and translations in each of the languages are involved.

- **Reference corpus:** It (e.g., Bank of English) is made to supply comprehensive information about a language. It is large enough to represent all relevant varieties of language and characteristic vocabulary, so that it can be used for writing grammars, dictionaries, thesauruses and other materials. It is composed on the basis of relevant parameters agreed upon by linguistic community. It includes spoken and written, formal and informal language representing various social and situational registers. It is used as a 'benchmark' for lexicons, for performance of generic tools, and language technology applications. With growing influence of internal criteria, reference corpus is used to measure deviance of special corpus.

- **Comparable corpus:** It is (e.g., Corpus of European Union) a collection of 'similar' texts in more than one language or variety. It contains texts in different languages where texts are not same in content, genre, or register. These are used for comparison of different languages. It follows same composition pattern but there is no agreement on the nature of similarity, as there are few examples of comparable corpora. It is indispensable for comparison in different languages and in generation of bilingual and multilingual lexicons and dictionaries.

- **Opportunistic corpus:** An opportunistic corpus stands for inexpensive collection of electronic texts that can be obtained, converted, and used free or at a very modest price; but is often unfinished and incomplete. Therefore, users are left to fill in blank spots for themselves. Their place is in situations where size and corpus access do not pose a problem. The opportunistic corpus is a virtual corpus in the sense that selection of an actual corpus (from opportunistic corpus) is up to the needs of a particular project. Monitor corpus generally considered as opportunistic corpus.

There can be some other types of specification such as closed corpus, synchronic corpus, historical corpus, dialect corpus, idiolect corpus, and sociolect corpus, etc. Therefore, the scheme of classification presented here is not absolute and final. It is open for re-categorisation as well as for sub-classification according to different parameters.

B- Identification of Target Users

There are no fixed target users for a general corpus. Anybody and everybody can use it for any kind of linguistic or non-linguistic purpose. For a specialised corpus, however, the question of target user is important. Since, each investigator or researcher has specific requirement, a corpus has to be designed accordingly. For instance, a person who is working on developing tools for machine translation will require a parallel corpus rather than a general corpus. Similarly, a person who is working on the comparative studies between two or more languages will require a comparable corpus rather than a monitor corpus. In the following list (Table 1) I have summed up the type of corpus users and their needs with regard to the type of corpus.

Table 2: Type of corpus users and their needs with regard to the type of corpus

Target users	**Corpus**
Descriptive linguists	General, written, and speech corpus
Speech technology people	Speech corpus and spoken corpus
Lexicographers and terminologists	General, monitor, specialised, reference, opportunistic corpus
Dialogue researchers	Speech, spoken, annotated, specialised corpus
Sociolinguistics	General, written, speech, monitor corpus
Psycholinguistics	Specialised, speech, written corpus
Historians	Literary, diachronic corpus
Social scientists	General, speech, written and special corpus
Comparative linguists	Bilingual, multilingual, parallel, comparable corpus
Information retrieval specialists	General, monitor, and annotated corpus
Tagging, processing and parsing specialists	Annotated, monitor, written, spoken, general corpus
Core-grammar designer	Comparable, bilingual, and general corpus
Teachers and students	Learner, monitor, and general corpus
Linguists	All types of corpus

C- Limitations of Corpus

(a) **Lack of linguistic generativity:** Chomsky and his supporters have strongly criticised the value of corpus in linguistic research. At the University of Texas in 1958, he argued, “any natural corpus will be skewed. Some sentences won't occur because they are obvious; others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list". Generativists argue that corpus cannot provide evidence for linguistic innateness.

By virtue of its structure and content, it only can represent the linguistic 'performances' but does not reflect on the linguistic 'competence' and 'generitivity' of the users. A corpus, which records only the examples of performance, cannot be useful to linguists, who seek to understand the tacit, internalised knowledge of language rather than the external evidences of language use on various contexts.

(b) Technical difficulties: Corpus building is a large scale, multidirectional, enterprising work. It is a complex, time-consuming, error-prone, and expensive task. The whole enterprise requires an efficient data processing system, which may not available to all, particularly in a country like India. Linguists need to be trained in computer use and data handing. It is a troublesome task. Unlike linguists of other countries, Indian linguists are not eager to take up computer into their stride. Computer scientists, on the other hand, are also not enthusiastic to work with the linguists in tandem. The gap is wide apart. Let us hope for a mutual co-operational interface to develop between the two groups in near future.

(c) Lack of texts from dialogues: Present day corpus fails to consider the impromptu, non-prepared dialogues taking place spontaneously in daily linguistic exercises. Absence of texts from dialogic interactions makes a corpus cripple lacking in the aspect of spontaneity, a valuable trait of human language. Corpus, either in spoken or written form, is actually a database detached from the actual context of language use. Detachment from the contexts makes a corpus (corpse + carcass) a dead database, which is devoid of many properties of living dialogic interactions, discourse, and pragmatics. It fails to reveal the real purpose underlying a linguistic negotiation (a difficult action game), identify the language-in-use, determine the verbal actions involved within the dialogues, describe the background where from the interlocutors derive cognitive and perceptual means of communication.

(d) Lack of information from visual elements: Corpus does not contain graphs, tables, pictures, diagrams, figures, images, formulae and similar other visual elements, which are often used in a piece of text for proper cognition and understanding. A corpus devoid of such visual elements is bound to lose much of its information.

(e) Other limitations: Corpus creation and research works are unreasonably tilted towards written texts, which reduce importance of speech. In reality, however, speech represents our language in a more reliable fashion than writing. The complexities of speech corpus generation make it a rare commodity. Thus, easy availability of text corpus and the lack of speech corpus inspire people to turn towards the text corpus. However, this does not imply that speech corpus has lost is prime position in corpus linguistics research.

Moreover, language stored in corpus fails to highlight the social, evocative, and historical aspects of language. Corpus cannot define why a particular dialect is used as the standard one, how dialectal differences play decisive roles to establish and maintain group identity, how idiolect determines one's power, position and status in society, how language differs depending on domains, registers, etc. Corpus also fails to ventilate how certain emotions are evoked by certain poetic texts, songs and literature; how world knowledge and context play important roles to determine intended meaning of an utterance; how language evolve, divide, and merge with the change of time and society, etc.

Suggested Readings

Aarts, J. and Meijs, W. (Eds.) 1984. Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research. Amsterdam-Atlanta, GA.: Rodopi

Bouillon, P. and Busa, F. (Eds.) 2001. The Language of Word Meaning. Cambridge: Cambridge University Press.

Dash, N.S. (2007) Language Corpora and Applied Linguistics. Kolkata: Sahitya Samsad.

Halliday, M.A.K. 1989. Spoken and Written Language. Oxford: Oxford University Press.

Kennedy, G. 1998. An Introduction to Corpus Linguistics. New York: Addison-Wesley Longman Inc.

Souter, C. and Atwell, E. (Eds.) 1993. Corpus Based Computational Linguistics. Amsterdam: Rodopi.

Young, S. and G. Bloothooft (Eds.) 1997. Corpus-Based Methods in Language and Speech Processing. Vol-II. Dordrecht: Kluwer Academic Press.

ICL/M2

Dr. HAMZAOUI

11. The Absence of Arabic Corpus Linguistics

Introduction

In a world of a revolutionary computer technology in the field of linguistics, it seems that the common practice among the Arab linguists in the Arab world is very much frustrating. The only thing that an Arab linguist who is conducting a linguistic research can do is painstakingly sitting in his own office either contriving his linguistic data or extracting his own corpus – a tedious process that involves reading through printed texts and manually recording his data. The linguistic results of this huge effort are not highly accurate because these data are far removed from the real language use, not empirical and lack representation.

It is well-known that Arabic linguistic research in the Arabic countries is not based on corpora because Arab countries do not have Arabic corpus linguistics as compared to the existing English corpus linguistics. Corpora are very important in the advancement of different Arabic linguistics such as sociolinguistics, psycholinguistics, historical linguistics, geolinguistics, contrastive linguistics, grammar, lexicography, stylistics, language pedagogy, and translation.

1. English Corpus Linguistics versus Arabic Corpus Linguistics

There are several English corpora that have been created for the purpose of the empirical study of English linguistics since the 1960s. Undoubtedly, these efforts in the field of corpus linguistics led to the advance of different fields of English linguistics. Some of these corpora are presented here:

-The British National Corpus (BNC) contains 100 million words of British English. Ninety percent of the corpus consists of various genres of written English, and ten percent comprises different types of spoken British English (Meyer 2002).

-Michigan Corpus of Academic Spoken English (MICASE) is a speech corpus. It was created for the purpose of studying the type of speech used by individuals conversing in an academic setting, such as classroom discussions, students presentations, tutoring sessions, class lectures, and dissertation defenses (Powell and Simpson 2001).

-A Representative Corpus of English Historical Registers (ARCHER) is a historical corpus. It covers the period (1650 – 1990) which is divided into fifty-year subgroup texts. Also it is a “multi-purpose general” corpus because it contains many different texts that cover different periods of English (Rissanen 2000).

On the other extreme, although corpora are widely available for English, there is very little available for the Arabic language. In fact, we still have a long way to go before we catch up with English corpora. Throughout the Arab world we do not have one single corpus that we created ourselves and the existed handful Arabic corpora have been created by others who are not Arabs, as shown below:

-The European Language Resources Association (ELRA) provides two Arabic corpora. The first, 140 million words in length, is a corpus of six years work of Al-Nahar newspaper from Lebanon. The second is a corpus of Al-Hayat newspaper and contains 18 million words.

-The Linguistic Data Consortium (LDC), University of Pennsylvania, produced three corpora: a corpus of Arabic newspaper texts containing 76 million words, a corpus of Egyptian Arabic Speech, and a lexicon of Egyptian Arabic. The first corpus is composed of articles from the Agency France Press (AFP) Arabic Newswire. The second corpus consists of 60 unscripted telephone conversations. For each conversation, both the caller and the callee are native speakers of the Egyptian dialect of Arabic who are making calls from inside the USA and Canada. The third one (Gadalla et al. 1998) is a CALLHOME English Arabic corpus of telephone speech and consists of 120 unscripted telephone conversations between native speakers of Egyptian Colloquial Arabic (ECA).

2. Corpus Linguistics and Its Implications in Arabic Linguistic Research

Corpus linguistics serves most areas of linguistics as being the raw material on which the linguistic researcher is working. Corpora are used in linguistic research for the purpose of linguistic description and analysis. The following subsections will explain the role that corpus linguistics plays in different areas of linguistic research with reference to Arabic linguistics.

2.1. Historical Linguistics

There are a number of English historical corpora that contain samples of writing representing earlier periods. These corpora are used to study both language variation in the earlier periods of English as well as language changes and development. For example, the Helsinki Corpus, a 1.5-million-word corpus, contains texts from the Old English through the early Modern English. This corpus has been used by historical linguists to study the evolution of English (Rissanen 1992).

Moreover, Skaffari (2009) did significant studies in the middle English words that were borrowed from France. Most of Arabic historical linguistics studies are not corpus based. For example, Wafi (2000), depending on the linguistic data collected manually from different books, studied the history of the Semitic languages: its origin, life and development. His book covered the phonology, grammar, lexicon of these languages, the factors that led to the appearance of different dialects, phonological change and the collapse of some of these languages.

Creating an Arabic historical corpus that represents different periods of Arabic language history will allow historical linguists to investigate systematically the development of, for example, particular grammatical and phonological aspects in the earlier Arabic periods. Moreover, such a corpus can help linguists to study the sociolinguistic variables that affected language usage, such as gender. Various dialect regions can also be studied throughout different historical periods.

2.2. Psycholinguistics

Corpus linguistics serves psycholinguistics. An important corpus that serves the field of psycholinguistics is the CHILDES (Child Language Data Exchange). This corpus contains transcriptions of children learning first and second languages and it has been studied by psycholinguists who are interested in child language acquisition (MacWhinney 2000). It is very important for the development of the study of Arabic psycholinguistics to create an Arabic psycholinguistics corpus. Such a corpus can include, for instance, speeches from Arab normal children who are developing their normal linguistic skills and those who have language disorders such as aphasia and autism.

Studying this corpus psycholinguistically may give us a real picture of the normal and abnormal data of the Arabic language of normal as well as linguistically impaired children. Most importantly, forming Arabic psycholinguistic corpus will provide linguists with the chance to conduct contrastive psycholinguistic studies by comparing the linguistic behaviour of Arab children with that of the English ones.

2.3. Sociolinguistics

Corpora can be used to study some sociolinguistic variables such as gender, dialect region, social status and age. For example, in the spoken part of the British National Corpus, Aston and Burnard (1998) used the software program Sara to count the number of instances of the adjective lovely spoken by males and females. They found that this word is used more frequently by females than males. The intended Arabic National Corpus can include a sociolinguistic section as an Arabic sociolinguistic corpus.

This section can include, for instance, the language of Arab teenagers that can be similar to the COLT corpus (the Bergen Corpus of London Teenage English) that contains the speech of London teenagers (Stenström and Andersen 1996) or the language of educated people. Such a section can help sociolinguists to conduct comparative studies to compare different sociolinguistic variables. It also helps in holding contrastive sociolinguistic studies based on Arabic sociolinguistic corpora and English sociolinguistic ones.

2.4. Lexicography

It is customary that a dictionary provides the users with different kinds of information about words including their meaning, pronunciation, part of speech and examples that give the contextual meaning of the word. Before using the linguistic corpora in lexicography, all this information had to be collected manually and it was time consuming. For example, the Oxford English Dictionary took fifty years to complete. The dictionary included five million citations which were "painstakingly collected … subsorted … analyzed" (Landau 1984: 69).Recently, the advancement in computer corpora and software programs changed the way we look at the dictionaries.

The use of a software program called the concordancing program that can count the frequency of words in a corpus, detecting affixes and sorting the words by lemmas. As for the parts of speech, if the corpus is tagged, the parts of speech of each word can be automatically determined.

Not only can the corpus be used to create new dictionaries, but also to revise the existed ones. In this case, the corpus can either supplement or refute the lexicographer's intuitions. To illustrate this point, Atkins and Levin (1995) studied verbs in the semantic category shake and quoted its definitions in three dictionaries: The Longman Dictionary of Contemporary English, The Oxford Advanced Learner's Dictionary and The Collins COUBUILD Dictionary. They found that both the Longman and COBUILD dictionaries list the verbs quake and quiver as being intransitive, while the Oxford dictionary lists quake and quiver as being transitive.

In calling up all the examples of these verbs in a corpus of 50,000,000 words, they found that both quiver and quake are used both transitively and intransitively. Thus, the dictionaries have got these verbs wrong. As far as the creation of Arabic dictionaries is concerned, Al-Eryaan (1984) discussed the stages of collecting

Arabic dictionaries manually. The first stage that the Arab lexicographer can do is to gather the words from people living in different regions depending on hearing these words such as ‘rain’, ‘sword’, etc. The second stage is to categorize the words under separate headings. The result will be a book for ‘rain’ and another for ‘sword’, for instance. The third stage is to gather all this information in a complete dictionary that includes all the words of Arabic.

Thus we have Al-Sahah Dictionary, Al-Waseet Dictionary, Al-Kabeer Dictionary, etc. This is the traditional way of creating Arabic dictionaries. The prospective Arabic National Corpus will make the process of creating dictionaries easier, improve the kinds of information contained in them, and address some deficiencies inherent in many of these dictionaries. This can be done by going through a huge number of computerized examples of Arabic that will be included in the prospective corpus.

2.5. Stylistics

If the stylist wants to offer confident stylistic studies, he has to analyze linguistic features of the texts that are computer-readable form. This requires analyzing the literary works to compare between the use of different linguistic devices not only in one’s own work but also with other authors’ works. This leads to a quantitative analysis of the work – an area where corpora play an important part. Leech and Short (1981) pointed out that stylistics often demands the use of quantification to back up judgments.

Arab stylists who study the stylistic features of the works of some Arabic writers go through their works and write the linguistic features manually – a very tedious and time consuming process. For instance, Al-Trabulsi (1996) analyzed the Anthology “Al-Shawqiyat” written by Ahmad Shawqi, the prince of poets, stylistically. Citing, manually, 11, 320 lines of poetry that cover 370 poems, he studied different linguistic aspects of Shawqi’s poetry.

However, creating an Arabic National Corpus will help researchers of stylistics to easily and objectively examine the linguistic features of Arabic writers. For example, converting the works of different Arab writers into a computer-readable form will provide the stylistic researchers not only with an effective means of studying the linguistic features of these writers but also comparing their style with that of other foreign writers.

2.6. Pragmatics

Pragmatics means language in context. Corpora are a plentiful source of studying pragmatics. For example, After Stenstöm (1987) examined what he termed “carry on signals” in a corpus, he was able to classify these signals according to their functions, e.g. right has been used in all functions, but especially in a response, to evaluate a previous response or terminate an exchange.

All right has been used to mark a boundary between two stages in a discourse. That’s right has been used as an emphasizer. And it’s alright and that’s alright have been responses to apologies. Creating an Arabic pragmatic corpus that is a part of the Arabic National Corpus will help linguists to study effectively Arabic pragmatics.

2.7. Contrastive Linguistics

Corpora can be used to facilitate contrastive linguistic analysis. For example, the English-Norwegian Parallel Corpus contains examples of English and Norwegian fiction and non-fiction that are 10,000 – 15,000 words in length. Such a corpus has been used to compare structures in both languages to allow a range of different contrastive studies (Johansson and Ebeling 1996). The appearance of such bilingual corpora led to the invention of the ParaConc program (Barlow 1999) which is used to align sentences in any two languages. Most, if not all, of the contrastive studies done by Arab linguists are not corpus-based.

The unavailability of computerized bilingual corpora has led those linguists to contrive an introspective linguistic data then subject this data to linguistic analysis. For example, Mahmoud (1989) studied the morphological, syntactic and semantic features of middle and inchoative verbs of Arabic and English, Gadalla (1999) gave a morphological and phonological analysis of Standard Arabic and Cairene Arabic, and Mansour (1999) gave a contrastive analysis of the morphosyntax of English and Arabic verbs.

These studies are valuable contributions and a step forward in the study of Arabic linguistics. However, had these studies been corpus-based, their findings might have been much different. A Section of the Arabic National Corpus can include a bilingual corpus. That is, collecting and computerizing some foreign texts to be included in the corpus. Creating such a bilingual corpus will facilitate such contrastive studies. Such a corpus might generate an impressive amount of research in the field of contrastive linguistics.

2.8. Language Pedagogy

One of the strategies that can be used in teaching a foreign language is to expose students to extensive training using a corpus. Using a concordancing program to investigate such a corpus will give students real examples of language usage rather than contrived ones that are often found in grammar books.

This inductive exploration of different linguistic constructions on vast amount of data will allow students to practice with concordance programs to generate so much data. Creating an Arabic National Corpus will help teachers and students alike to practice Arabic and English grammatical structures by themselves using language in context. Such a process, Gavioli (1997: 84) claims, is an effective "language-learning activity".

4.9. Translation Bilingual corpora that contain translated texts from two or more languages can facilitate translation studies, train translators and advance linguistic translation theories. Moreover, using such information in translation can be used to create bilingual dictionaries (Schmied and Schäffler 1996). One section of the suggested Arabic National Corpus can include translated works from English to Arabic and vice versa. Gadalla (2003), for instance, studied translating Arabic perfect verbs into English through analyzing manually two Arabic novels by Naguib Mahfouz.

A corpus of 250 sentences was randomly and manually chosen from the two novels, 125 sentences from each novel. In fact dealing with such an issue and other similar ones in translation through a computer-readable corpus of a large number of texts may make the task easier and more effective.

2.9. Grammar studying grammatical structures

Whether in morphology or syntax, it yields linguistic information on these structures and their frequency. In the area of qualification, Arts (1992) used the London Corpus to analyze “small clauses” in English. He was able to provide a complete description of the small clauses. In the area of quantification, Collins (1991), in a corpus study, compared the relative frequency of modals in four genres of Australian English: press reportage, conversation, learned prose, and parliamentary debates to test whether modals of necessity and obligation are more suitable for some contexts than others.

For more recent corpus study of the history of English syntax, see Rissanen 2012) In the field of Arabic linguistics, Kebbe (2000), for example, gave a transformational analysis of modern written Arabic based on the transformational theory as formulated by Chomsky. Though the majority of works on the Arabic language concentrated on regional dialects and this book fulfils a long-felt need by focusing on modern written Arabic, it is not corpus-based – a prerequisite that might render the book more realistic. Moreover, Fischer (2000) offered a book which is though unquestionably considered the most useful reference grammar of the classical Arabic language.

However, it is not corpus-based because “the examples cited are for the most part borrowed from the standard grammatical treatises (Wright, Nöldeke, Reckendorf, Bbrokelmann, Wehr, Spitaler) and to a smaller extent are supplemented from my own stock)” Fischer (2000: xiii). In the field of Arabic morphology, Abd-Elghany (1970) studied the morphological units and their role in Arabic word formation, not through a corpus, rather using examples that he obtained either introspectively or cited from other books.

Better results might be obtained if his morphological analysis were based on a computer corpus. Creating the Arabic National Corpus will facilitate the study of Arabic grammatical and morphological structures instead of studying them in contrived contexts or depending on the manual corpus gathered by the researcher.

4.11. Geolinguistics

This branch of linguistics studies different linguistic aspects of the languages and dialects in terms of regional distribution. Geolinguistics provides us with maps that present different linguistic features of the different dialects by region. There are different atlases in the world that give dialectical maps for different languages in Europe and America. We do not, as Arabs, have one. The only preliminary effort that was made for Arabs was at the hands of Bergsträsser, the German Orientalist, who applied this idea on the Arabic language for Syria and Palestine (AbdEltawab 1997).

In fact, creating the Arabic National Corpus can contain a section that includes a survey of the different dialects not only between Arab countries but also in the same country, i.e. creating an atlas for the dialects in Saudi Arabia. . This may help to form the Arabic national geolinguistics atlas which classifies, by region, similarities and differences between different Arabic dialects in phonetics, morphology, syntax and semantics.

5. Designing an Arabic National Corpus

After showing how most of the Arabic linguistic studies cannot dispense with corpora if they are aiming at offering a real picture of a real use of a real language, the next step is how to plan the Arabic National Corpus effectively because as Meyer (2002: 53) states that “well planned corpora are the most effective tools possible for linguistic research”. Following Meyer (2002), creating such a corpus goes through four processes: planning the corpus, collecting the data, computerizing the data, and analyzing the data.

Collecting Data Two kinds of samples are collected to be incorporated in the corpus: speech samples and written samples. The first thing that the corpus linguist can do is to collect speech.

Collecting Speech Sample

In collecting the speech samples in the British National Corpus, for example, the participants, not the corpus linguist, in the project were given portable tape recorders and instructed to record all the conversations for a period ranging from 2 – 7 days (Crowdy 1993). Digital recorders are sometimes preferred particularly when we work on the computer to edit unwanted background. The individuals record speech in different social contexts such as conversations over dinner, informal conversations among friends, co-workers speaking at work, teachers and students in class discussions, etc.

Those individuals can use different of microphones that are suitable for the situation. According to Meyer (2002) three types of speech are collected: direct speech, telephone conversations, and radio and television broadcasts. Different microphone types can be used with the first type of speech:

1. Uni-directional microphones are used to record single individuals;

2. Omni-directional microphones are used for larger groups;

3. Wireless microphones and laviere microphones (worn around the neck) are used by persons who are moving around and giving speech; 4. Extra-sensitive microphones can be used for recording individuals who are not close to the microphones.

As for recording telephone conversations of individuals talking over the telephone, adaptors can be used to record directly over the telephone. The third source of speech is the radio and television broadcast. In recording this type of speech, one either puts the audio input plug on the tape recorder or connecting the TV to a video cassette recorder by running a line from the audio output plug on the recorder. In collecting spoken samples for the suggested Arabic National Corpus there are some steps to be followed. First, participants are given digital recorders and instructed to collect speech from different social occasions. Those people are also provided with different types of microphones. Secondly, some people are given adaptors that can be used to record directly over the telephone. Thirdly, the TV will be a plentiful resource of the Arabic National Corpus. For example, corpus linguists can record the news to study Modern Standard Arabic, to record plays and serials to study different Arabic dialects. Also, we can record sports comments, discussions, commercials, etc.

Collecting Written Samples

Collecting samples of writing is the second main task. The first step is to obtain permission from the authors that their writings are included in the corpus. The second step is to gather the written texts, 1,000 to 2,000 word samples that will be computerized.

Computerizing the Corpus

After collecting the written and spoken texts, it can be entered into a database. First of all, we need to transcribe the collected speech. That is representing the oral form of language in a written form. As for English, there are software programs designed to transcribe English speech that has been digitized such as “Voice Walker 2.0”. As for Arabic, we need a program to transcribe Arabic speech and turn it into a spoken form. Concerning computerizing written texts, the first step is to convert written texts into electronic format either with retyping texts or with optical scanners.

Analyzing the Corpus

After computerizing the corpus, the next step is to find the appropriate software programs as well as the appropriate statistical tests for both quantitative and qualitative analysis. These programs and tests are used by linguistic research to analyze the data. As for programs, software programs can be used in corpus analysis.

The most common software program to be used with a corpus is the Concordancing program. Kettemann (1995: 4) argues that the concordancing program is “an extremely powerful hypothesis testing device”. This program can be used by researcher to conduct searches for words, group of words, suffixes, prefixes and calculating the frequency. The next step is to subject the information obtained through the programs to some kind of statistical analysis to make frequency counts as well as determining the similarities and differences and to show to what extent they are statistically significant.

- Concluding Remarks

Creating the Arabic National Corpus is not an effort of an individual or even a group of individuals, rather it is a national project that needs the collaboration of many institutions in different Arabic countries such as the Arabic language academies, Arab Scientific Research Councils Unions, King Abdul Aziz City For Science and Technology, Arab universities including faculties of arts, faculties of education, and faculties of computers and information systems.

Moreover, since it is a national project it needs the governmental support particularly for funding. Creating an Arabic National Corpus will be rewarding and will help in advance the study of the Arabic language and linguistics. Although this work, if taken seriously, will be in its infancy in the Arab world and requires methodological refinement, it seems to be an interesting and promising area of studying Arabic linguistics.

References

Arts, B. (1992). Small clauses in English: the nonverbal types. Berlin and New York: Mouton de Gruyter.

Abd-Elghany, A. (1970). Al-wahadaat al-sarfiyya wa dawraha fi binaa’ al-kalima al-Arabiyya [Morphological units and their roles in Arabic word-formation]. A Published M.A. Dissertation. Cairo: Dar El-Nashr Press.

Abd-Eltawab, R. (1997) Al-madkhal ila ‘lm al-lugah wa manahij al-bahth al-lughawy [An Introduction to linguistics and methods of linguistic research]. 3rd edn. Cairo: Maktabt Al-Khanji.

Al-Eryaan, M. A. (1984). Al-maajim al-Arabiyya al-mujanasah [The hybrid Arabic dictionaries]. Cairo: Dar AlMuslim. Al-Trabulsi, M. A. (1996). Khasa's al-'uslup fi al-shawqiyyat [Stylistic features of al-shawqiyyat]. Cairo: AlMajlis Al-Ala Lil-thaqafa.

Aston, G. and L. Burnard (1998). The BNC handbook: exploring the British national corpus with SARA. Edinburgh: Edinburgh University Press.

Atkins, B., T. Sue and B. Levin (1995). Building on a corpus: a linguistic and lexicographical look at some nearsynonyms. International Journal of Lexicography 8.2: 85-114.

Barlow, M. (1999). MonoConc 1.5 and PraConc. International Journal of Corpus Linguistics 4.1: 319-27.

Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing 8: 241-57. --- and J. Burgues. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. Journal of English Linguistics 28.1: 21-37.

Collins, P. (1991). The modals of obligation and necessity in Australian English. In Aijmer and Altenberg. 145- 65. C

ICL/M2

Dr. HAMZAOUI

12. Debating and Looking to the Future of Corpus Linguistics

1. Debates in corpus linguistics

It was previously mentioned that corpus linguistics is viewed primarily as a methodology, not a theory. However, this should not be understood to imply that corpus linguistics is theory-free. The focus and method of research, as well as the type of corpus selected for a study, is influenced by the theoretical orientation of the researchers, explicit or implicit. Kennedy’s statement that corpus linguistics has “a tendency sometimes to focus on lexis and lexical grammar rather than pure syntax” (1998:8) is a case in point.

Methodologically, corpus linguistics is equally diverse and encompasses different approaches to corpus building and use. The main points of tension in corpus linguistics, which are interconnected, concern the relation between theory and data, the utility of corpus annotation, and the role of intuitions. These tensions have been formalised in the distinction between corpus-based and corpus-driven approaches to linguistics (e.g. Tognini-Bonelli, 2001).

This distinction is not acknowledged by all corpus linguists, and it has been felt by some to be overstated (Aarts, 2002: 121), as "the worlds of the corpus-based and of the corpus-driven linguist may not be all that far apart as they are made out to be" (ibid.: 123). However, since at the centre of this distinction lie the central issues outlined above, the definitions of the corpus-based and corpus-driven approaches can serve as a springboard for the discussion of these issues.

In the corpus-based approach, the corpus is mainly used to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli, 2001: 65). Although the intuitive basis of the theories being tested is seen as a weakness of this approach, it is not as much the target of criticism as the attitudes to, or techniques for, dealing with discrepancies between theoretical statements and corpus data that are supposed to characterise corpus-based linguists.

Corpus annotation is a central feature of all three techniques. The first is to "insulate the data",3 that is, either to dismiss data that do not fit the theory, or to make the data fit the theory, for example, by annotating the corpus according to the theory (ibid.: 68-71). The second technique is to reduce the data to "a set of orderly categories which are tractable within existing descriptive systems" (ibid.: 68), again by annotating the corpus.

The criticism here is two-pronged: the annotation scheme is based on a pre-conceived theory, and the manual annotation of the training corpus is influenced by both the theory and the annotator's intuitions. The third approach is "building the data into a system of abstract possibilities, a set of paradigmatic choices available at any point in the text" (ibid.: 74), and is strongly associated with Halliday's probabilistic view of grammar (e.g. 1991, 1992).

This stance is criticised mainly on two related grounds: its focus is predominantly paradigmatic rather than syntagmatic, that is, it is concerned with grammar rather than lexis (Tognini-Bonelli: 75-77), and, consequently, requires an annotated corpus, since "grammatical patterns … are not easily retrievable from a corpus unless it is annotated" (ibid: 77). The basic tenet of the corpus-driven approach is that any "theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus" (ibid.: 84).

Corpus-driven research aims at discovering facts about language free from the influence of existing theoretical frameworks, which are considered to be based on intuitions, and, therefore, are not comprehensive or reliable. Consequently, research is carried out on unannotated corpora, as annotation would impose a restrictive theoretical taxonomy on the data. A further characteristic of this approach is that it makes no distinction between lexis and grammar, as that, too, would require using existing distinctions, which may not be supported by the corpus data. Finally, in the corpus-driven approach the starting point of research is the patterning of orthographic words.

As these issues are interrelated, their discussion will overlap to some extent. At one end the corpus is used to find evidence for or against a given theory, or one or more theoretical frameworks are taken for granted;5 at the other, the observed patterns in the corpus data are used as a basis from which to derive insights about language, independent of pre-existing theories and frameworks, with a view to developing a purely empirical theory.

Of course this distinction begs the question of whether data observation and analysis can ever be atheoretical. It is interesting to note that the corpusbased approach, which is criticised, is associated with corpus research influenced by the work of Leech (e.g. 1991) or Halliday (e.g. 1991), and is presented as typically prioritising "the information yielded by syntactic rather than lexical patterns" (Tognini-Bonelli, 2001: 81), whereas the corpus-driven approach, which is proposed, is associated with corpus research influenced by the work of Sinclair (e.g. 1991) and Firth's contextual theory of meaning, and favours a focus on lexical patterning.

This indicates that the distinction is not only methodological, but also theoretical. Hunston & Francis (2000: 250), who have located their study of pattern grammar within the corpus-driven paradigm, state that their method "is indeed theory-driven", as "theories are, in a sense, constructed by methods". Our view is that an atheoretical approach is not possible and hence the idea of corpus-driven approaches to language must be seen as an idealized extreme, for, as Stubbs (1996: 47) notes, “the concept of data-driven linguistics must confront the classic problem that there is no such thing as pure induction. … The linguist always approaches data with hypotheses and hunches, however vague”. Sampson (2001: 124) shifts the focus from the formulation of hypotheses to their testing: We do not care how a scientist dreams up the hypotheses he puts forward in the attempt to account for the facts - he will usually need to use imagination in formulating hypotheses, they will not emerge mechanically from scanning the data.

What is crucial is that any hypothesis which is challenged should be tested against interpersonally observable, objective data. For example, Biber et al. (1999) make use of some frameworks used in Quirk et al. (1985), but they are also influenced by research in lexicogrammar (Biber et al., 1999: viii, 13). The testing of hypotheses on corpus data is related to the use of intuitions and the annotation of corpora. Sinclair (2004: 39) contrasts two attitudes in corpus linguistics research in a manner which reveals that, for those working within the corpus-driven paradigm, the use of annotation is seen as interconnected with the use of intuition.

Some corpus linguists prefer to research using plain text, while others first prepare the texts by adding various analytic annotations. The former group express reservations about the reliability of intuitive "data", whereas the latter group, if obliged, will reject corpus evidence in favour of their intuitive responses. One explanation for this connection is that adherence to a given theory is expected to have influenced the linguist to such an extent that the categories and structures recognised by the theory have become part of his/her intuitions. Sampson (2001: 135) highlights the role of schooling in the forming of intuitions: "Certainly we have opinions about language before we start doing linguistics. … In some cases our pre-scientific opinions about language come from what we are taught in English lessons, or lessons on other languages, at school".

Similarly, Sinclair (2004: 40) sees intuition not as a "gut reaction to events, [but] educated in various ways, and sophisticated". It can be argued that the influence of education on intuitions about language is more pronounced in linguists, whose education and training involves familiarisation with a number or theories, and, not uncommonly, in-depth study of a specific theoretical framework. Although the usefulness of intuitions in the forming of hypotheses has been challenged by corpus-driven linguists, there seems to be a consensus that intuitions are unavoidable in the interpretation of corpus data (e.g. Hunston, 2002: 65). However, Sinclair (2004: 47) has argued that there is a way for "keeping … intuition temporarily at bay".

The technique seems to involve the decontextualisation of the observed patterns and a temporary disassociation of form and meaning, and is aided by examining the vertical patterns of the key word in a concordance, or slotting in alternative words in a frame (e.g. on the __ of). Sinclair (ibid.: 47- 48) argues that Since the essence of finding the meaning-creating mechanisms in corpora is the comparison of the patterns - as physical objects and quasi-linguistic units - with the meanings, it is valuable to be able at times to study one without the other. This takes a little skill and practice, but to my mind should be an essential part of the training of a corpus linguist.

One criticism of annotation is that it imposes the categories of a theoretical framework on the data, a practice which may interfere with finding evidence against the theory, or with discovering language features that the theory does not predict. There is also disagreement on whether annotation adds information, and therefore "value", to the corpus (Leech, 1997: 2), or whether it "loses information" (Sinclair, 2004: 52), because it assigns only one unalterable tag, when the word may not clearly belong to one existing category.

Finally, reservations have been expressed regarding the degree to which corpus researchers are aware of the theoretical assumptions underlying different annotation schemes (e.g. Hunston, 2002: 67; Sinclair, 2004: 55-56) Leech (1997: 6-8) outlines three "practical guidelines, or standards of good practice" (ibid.: 6) for the annotation of corpora, and three further "maxims [applicable] both to the compilers and users of annotated corpora" (ibid.: 6-7), which partly address these reservations.

1. The raw corpus should be recoverable.

2. The annotation should be extricable.

3. The corpus user should have access to documentation providing information about the annotation scheme, the rationale behind it, the annotators, the place of annotation, and comments on the quality of annotation.

4. The annotation scheme "does not come with any 'gold standard' guarantee, but is offered as a matter of practical usefulness only" (ibid.: 6)

5. The annotation scheme should be "based as far as possible on consensual or theoryneutral analyses of the data" (ibid.: 7) [boldface in original].

6. "No one annotation scheme should claim authority as an absolute standard" (ibid.) There is agreement on the necessity for the unannotated version of a corpus to be available to researchers (Leech, 1997: 6; Sinclair, 2004: 50-51). There also seems to be an area of consensus on the need for researchers to be aware of the theoretical principles behind the annotation scheme.

Although Leech's point (3) above does not include the explicit statement of the theory informing the annotation, it can be argued that the theoretical framework is inferable from the information given in the documentation. The main point of concern, that of the imposition of a theory on the data, seems to be largely unresolved. Linguists of the corpus-driven persuasion would find existing annotation schemes influenced by intuition-based theories, and, therefore, restricting. Proponents of annotation 8 would see the annotated corpus as "a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation" (McEnery & Wilson, 2001: 32).

However, some consensus, albeit implicit, regarding the categories used in annotation schemes seems to exist, as corpus-driven studies do make use of what might be called traditional categories, such as ‘verb’, ‘preposition’, 'object', ‘clause’ and 'passive', without a definition (e.g. Hunston & Francis, 2000; TogniniBonelli, 2001), which indicates that they are treated as given. Furthermore, if, as Sinclair (2004: 47-48) proposes, it is feasible for linguists to distance themselves from their intuitions, it can be argued that it is also feasible to adopt an informed and critical approach towards the annotation.

Finally, irrespective of the perceived usefulness of the annotated corpus as a product, the annotation process can reveal the strengths and limitations of the theory informing the annotation scheme and lead to its modification - a process which is consistent with an empirical approach. Aarts (2002: 122) argues that "the only way to test the correctness and coverage of an existing description is to formalize it into an annotation system and test it on a corpus. … It is the annotation process, rather than its result (i.e. an annotated corpus) that matters" (see also Leech, 1992: 112).

Although, within the corpus-driven paradigm, annotation is seen as counterproductive when the corpus is used for theoretically-oriented research, it is deemed acceptable when the corpus is annotated with a view to be used in an "application" (Sinclair, 2004: 50-56), that is, "the use of language tools in order to achieve a result that is relevant outside the world of linguistics … [such as] a machine that will hold a telephone conversation, or a translating machine or even a dictionary" (ibid.: 55). An argument that can be advanced on the basis of this view is that if applications relying on a corpus which has been annotated according to a theoretical framework are successful, then this can be regarded as an indication that the theory affords helpful insights into actual language use. Undoubtedly, there are pitfalls and limitations in uncritically using an annotated corpus.

However, the use of an unannotated corpus has its own pitfalls and limitations. An unannotated electronic corpus lends itself to the examination of forms and their patterns, as the software exists that will produce a concordance of a word-form for manual examination, or statistical measures of the strength of its collocation patterns, from an unannotated corpus. However, an unannotated corpus is of little, if any, use if the research focus is upon grammatical categories, semantic notions or pragmatic functions. Tognini-Bonelli (2001: 89- 90) concedes that "while collocation is instantly identifiable on the vertical axis of an alphabetical concordance, colligation represents a step in abstraction and is therefore less immediately recognisable unless the text is tagged with precisely the required grammatical information". Sampson (2001: 107) agrees that, "in general, more complex forms of investigation may only be possible if the computer has access to some form of detailed linguistic analysis of the text".

Also, the interpretation of concordance lines (e.g. Hunston, 2002: 38-66), that is the manual examination of concordances in order to identify patterns, which is a frequently used technique of corpus-driven linguists, is open to what we might call 'implicit annotation'. That is, while examining concordance lines, researchers may assign grammatical or semantic roles to words or configurations of words, either unwittingly, influenced by tradition or their education, or consciously, refraining from using established roles and patterns.

What becomes evident from the discussion of tensions in corpus linguistics is that theoretical and methodological issues are interconnected. Therefore, these issues will, inevitably, be revisited in the remainder of this chapter. In sum, when considered from specific theoretical or methodological viewpoints, different approaches to corpus linguistics appear to have merits, as well as problems and limitations. However, when considered from the viewpoint of linguistics in general, the current diversity in corpus research can only be seen as an indication of health, and should be welcomed. The next section examines in some detail the theoretical assumptions and methodological positions of what has been termed the lexical approach, and which lies behind the corpus-driven approach to linguistic research.

2. Looking to the Future of Corpus Linguistics

Given the enormous changes in the world of technology over the last five years, it is difficult to imagine the scope of changes that might take place in the area of corpus construction and tools. However, making a wish list for the future is always a delightful task. One of the changes that we will see in the near future is greater availability of spoken corpora. This could be a result of two factors. First, researchers may be more able and willing to share the spoken corpora that they have assembled. Second, hopefully, creating spoken corpora will benefit from technological advances in speech recognition, thus making the task of transcribing spoken language to text files a much more efficient process and more automated task. Perhaps digital sound files will be fed through a conversion program and then the researcher can go through to edit any areas that are problematic. This would be a tremendous boost to spoken language researchers.

The development and use of video and multi-modal corpora is another area that will probably change dramatically in the next decade. Some research is already being done in this area (Carter and Adolphs 2008; Knight and Adolphs 2008; Dahlmann and Adolphs 2009) and given how quickly technology can advance, this seems to be the next area that can provide new levels of corpus building and analysis, allowing us to ask and answer questions that are not even imagined at this point in time.

As far as the field of sociolinguistics is concerned, there is also a need for better standardisation of sociolinguistic and corpus-linguistic methods for annotating and accessing language data. There is a need for harmonisation of annotation schemes, particularly at the levels of discourse, speech acts and interactional structure. This will enable easier cross-corpora comparison, for the benefit of both sociolinguistics and language technology purposes.

Given the range of studies which have successfully applied corpus-linguistic methods for sociolinguistic purposes, it is beyond doubt that corpus-linguistic methods have a lot to offer to sociolinguistic research. As we have seen, corpora have been used to study linguistic variation between different varieties of a language and between different groups of speakers, by pointing at manifest differences in language use within a corpus or between corpora. Corpus linguistics provides efficient tools and methods for the collection, annotation and study of spoken data. This makes spoken corpora ideal for sociolinguistics.

References/ Further reading

Andersen, G. (2001) Pragmatic Markers and Sociolinguistic Variation. Amsterdam: John Benjamins.

Reppen, Fitzmaurice, R. and Biber, D. (eds) Using Corpora to Explore Linguistic Variation. Amsterdam:John Benjamins.

Schneider, K. P. and Barron, A. (2008) Variational Pragmatics: A Focus on Regional Varieties in Pluricentric

Languages. Amsterdam: John Benjamins. (This book is purposefully designed to establish and explore the field of variational pragmatics and contains several corpus-based studies.)

Tagliamonte, S. (2006) Analysing Sociolinguistic Variation. Cambridge: Cambridge University Press.

-Where to find further information on corpus linguistics

Other web-based introductions to corpus linguistics: http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html (Concordances and Corpora Tutorial by Catherine Ball at Georgetown University.) http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm (By McEnery & Wilson, intended as a supplement to their 2001 book.)

General books on corpus linguistics:

Biber, Douglas; Conrad, Susan; Reppen, Randi (1998). Corpus Linguistics. Investigating Language Structure and Use. Cambridge: CUP.

Kennedy, Graeme (1998). An Introduction to Corpus Linguistics. London & New York: Longman.

McEnery, Tony; Wilson, Andrew (2001). Corpus Linguistics. Edinburgh: EUP. Meyer, Charles F. (2002). English Corpus Linguistics. An Introduction. Cambridge: CUP.

Partington, Alan (1998). Patterns and Meanings. Using Corpora for Language Research and Teaching. Amsterdam: Benjamins.

Sinclair, John (1991). Corpus, Concordance, Collocation. Oxford: OUP. Tognini-Bonelli, Elena (2001). Corpus Linguistics at Work. Amsterdam: Benjamins.

Collections on studies and issues in corpus linguistics:

Aarts, Jan; de Haan, Pieter, Oostdijk, Nelleke, eds. (1993). English Language Corpora: Design, Analysis and Exploitation. Amsterdam: Rodopi.

Aijmer, Karin; Altenberg, Bengt, eds. (1991). English Corpus Linguistics. London: Longman.

Leitner, Gerhard, ed. (1992). New Directions in English Language Corpora. Berlin: de Gruyter.

Svartvik, Jan, ed. (1992). Directions in Corpus Linguistics. Berlin: de Gruyter. Thomas, Jenny; Short, Michael, eds. (1996). Using Corpora for Language Research. London: Longman.

A few examples of (paper-length) corpus-based studies:

Altenberg, Bengt (2002). “Modality in advanced Swedish learners‟ written interlanguage.” In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, eds., Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: Benjamins, 55-76. [ICLE]

Berglund, Ylva (2000). “Utilising present-day English corpora: a case study concerning expressions of future.” ICAME Journal 24, 25-63. (available at http://nora.hd.uib.no/journal.html) [BNC, LLC, LOB, FLOB]

Gotti, Maurizio (2003c). “Shall and will in contemporary English: a comparison with past uses.” In Roberta Facchinetti, Manfred Krug, Frank Palmer, eds, Modality in Contemporary English. Berlin / New York: Mouton de Gruyter, 267-300. [Helsinki corpus, diachronic part]

Hundt, Marianne (1998). “It is important that this study (should) be based on the analysis of parallel corpora: On the use of the mandative subjunctive in four major varieties of English.” In Hans Lindquist et al., eds., The Major Varieties of English. Papers from MAVEN ’97. Växjö: Acta Wexionensia. [LOB, Brown, FLOB, Frown, ACE, WCNZE]

Hundt, Marianne (2004). “Animacy, agentivity, and the spread of the progressive in Modern English.” English Language and Linguistics 8 (1), 47-69. [ARCHER]

Leech, Geoffrey (2003). “Modality on the move: the English modal auxiliaries 1961-1992.” In Roberta Facchinetti, Manfred Krug, Frank Palmer, eds, Modality in Contemporary English. Berlin / New York: Mouton de Gruyter, 223-240. [LOB, Brown, FLOB, Frown, ICE-GB, Survey of English Usage]

Skandera, Paul (2000). “Research into idioms and the International Corpus of English.” In Christian Mair & Marianne Hundt, eds., Corpus Linguistics and

Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999. Amsterdam: Rodopi, 339-353. [ICE-East Africa]

Information on certain corpus resources:

Biber, Douglas; Edward Finegan, Dwight Atkinson (1994). “ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers.” In Udo Fries, Gunnel Tottie and Peter Schneider, eds., Creating and Using English Language Corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993. Amsterdam & Atlanta: Rodopi, 1-13.

Granger, S., E. Dagneaux & F. Meunier, eds. (2002). International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Greenbaum, Sidney, ed. (1996). Comparing English Worldwide: The International Corpus of English. Oxford: Clarendon.

On corpus linguistics versus introspection:

Fillmore, Charles (1992). “‟Corpus-linguistics‟ vs. „computer-aided armchair linguistics‟”. In Jan Svartvik, ed., Directions in Corpus Linguistics. Berlin: de Gruyter, 35-60.

On statistics and corpus linguistics:

Oakes, Michael P. (1998). Statistics for Corpus Linguistics. Edinburgh: EUP.

On corpus linguistics and linguistic theories:

Halliday, M.A.K. (1991). “Corpus studies and probabilistic grammar.” In Karin Aijmer & Bengt Altenberg, eds., English Corpus Linguistics. London: Longman, 30-43.

Halliday, M.A.K. (1992). “Language as system and language as instance: the corpus as a theoretical construct.” In Jan Svartvik, ed., Directions in Corpus

Linguistics. Berlin: de Gruyter, 61-77. Schönefeld, Doris (1999). “Corpus linguistics and cognitivism.” International Journal of Corpus Linguistics 4, 137-171.

On the use of corpora for diachronic analyses:

Rissanen, Matti (1992). “The diachronic corpus as a window to the history of English.” In Jan Svartvik, ed., Directions in Corpus Linguistics. Berlin: de Gruyter,185-209.

On the analysis of learner corpora:

Nesselhauf, Nadja (2004). “Learner corpora and their potential for language teaching.” In John Sinclair, ed., How to Use Corpora in Language Teaching. Amsterdam & Philadelphia: Benjamins, 125-152.