Aims, tools and practices of Corpus Linguistics

The discipline of Corpus Linguistics, in essence, entails the compilation of very large databases or archives of texts for subsequent linguistic analysis. The final ends of Corpus Linguistics lie in the scope of Artificial Intelligence, that is, teaching machines to comprehend and produce natural language. A vital correlated aim is to improve translation techniques, both human and machine. Intermediate ends include furthering our knowledge of how language is structured and how humans use it to communicate thought, to express evaluations and to influence the behaviour of their interlocutors (i.e. persuasion).

Corpora can be either heterogeneric or monogeneric, that is, they may contain texts of many different types, generally as many different types as the compilers can practically and legally obtain, or they may contain texts of a single type. The former, heterogeneric corpora are thus intended to be in some way representative of the language in question as a whole. The latter, on the other hand, are compiled as a means of studying one particular text-type, for example, the language of law, of economics, of Parliamentary debates, and so on. Heterogeneric corpora tend to be very large, nowadays typically from 100million to a billion words in size. Their compilation is complex and expensive and tends to be carried out by special organizations attached to Universities or large institutions, such as publishing houses. Monogeneric corpora, on the other hand, can be relatively easy to compile and are often created by individual researchers with a special interest in a particular text-type. The favourite source for accessing texts today is the Internet.

Editing

Corpora are sometimes edited, either by the compilers or by third-party users. There are two principal forms of editing known respectively as *part-of-speech*(or *POS*) *tagging* and *mark-up*. In the first of these each lexical element in the corpus of segment thereof is assigned a *tag* or label indicating its grammatical status (noun, determiner, qualifier and so on) in the context in which it appears. This is usually performed semi-automatically; the software makes a preliminary assignment but human post-editing is normally essential. Tagging is generally carried out for linguistic purposes, either as a precursor to parsing the text or to the check the accuracy (and therefore grammatical understanding) of the tagging system.

Editors may choose to mark-up an almost infinite variety of items. They may wish to indicate structural units of texts, such as introductions and closing sequences, or passages of transaction and interaction, or even shifts in the topic of discussion. In spoken texts they may wish

to add information about the sex, age, occupation, and so on, of speakers. Or they may wish to indicate the occurrence of foreign words, slang, personal names, place names, dates, or almost anything an analyst might conceivably be interested in. Standardized editing protocols have been devised which enable marked-up texts to be machine-read in any platform environment. The most commonly used and the one adopted for use in the current research is the Text Encoding Initiative (T.E.I.).

Such editing is clearly highly painstaking and requires considerable investments of time and financial resources.

Instruments

A corpus by itself it simply an inert archive. However, it can be 'interrogated' using dedicated software. The most important interrogation tools include, first of all, the *concordancer*, then calculators of *frequency*, *keywords*, *clusters* and *dispersion*.

The concordancer extracts as many examples as the analyst wishes of the word or expression under analysis and arranges them in a KWIC (KeyWord In Context) concordance, which is thus a list of unconnected lines of text that have been summoned by the concordance program from a computer corpus. At the centre of each line is the item being studied (keyword or node). The rest of each line contains the immediate co-text to the left and right of the keyword. Such a list enables the analyst to look for eventual patterns in the surrounding co-text, which proffer clues to the use of the keyword item. It allows the observer to discover patterns of *collocation*, that is, how any particular word or expression cooccurs with other words/expressions with particular frequency. These patterns are often not available to unassisted introspection.

The frequency calculator supplies a list of the words in the corpus in order of frequency. The frequency lists of two or more corpora can also be compared using the *Keyword* facility to show up *relative* frequency, or *key*-ness of vocabulary in a corpus (it should be noted that this is a different use of *key* from that used in concordancing). In practice, this tool produces lists (one alphabetical and one ordered by significance) of all words which are significantly *more* frequent in the first corpus than the second and also of those which are significantly *less* frequent.

Clusters are sequences or strings of words (generally from two to a maximum of eight items) which occur 'with a particular frequency fixed by the inquirer in the set of texts being examined'. They are a kind of 'extended collocation'. Clusters are an intriguing phenomenon in themselves. Partington and Morley (2004) suggest they 'constitute "missing links" on the chain or cline from the linguistic morass to the abstraction we call grammar' and their study will 'tell us a great deal

about how speakers go about the construction of discourse'. In discourse terms, they reveal typical ways of saying things and therefore typical author/speaker messages. The software generally allows the user to cluster items in three ways, either from the Concordance programme by clicking directly on the cluster menu option, or cluster lists can be prepared from *WordList* (by activating and specifying cluster length in the *settings* menu option) and finally key-cluster lists can be compiled by comparing cluster lists. These latter become efficient when very large corpora are being examined.

Finally, the dispersion tool plots where an item occurs within a text. It can display in graphic fashion where an item or set of items typically occur in a large number of texts. Thus it is possible, for instance, to observe where certain types of modals typically appear in newspaper editorials (Morley 2004) or when during press briefings particular issues tend to be discussed, which may well reflect the relative degree of importance the participants endow them with.

Heterogeneric corpora and the study of language

Heterogeneric corpora, by enabling researchers to take into account vast quantities of language data and therefore obtain an overview of the authentic behaviour of language users not otherwise readily available to the 'naked ear', have helped provide a mass of new information about the grammar and lexis of languages, and have led to the compilation of a new generation of dictionaries, of grammatical descriptions, as well as language-teaching materials. Given its global importance, the lion's share of corpus research has been conducted into English, but the field is in expansion as regards other languages, especially German, Portuguese, French, Polish, Japanese and the Scandinavian languages.

Examples of lexicological and grammatical research which heterogeneric corpora have enabled include the following.

Grammar: before the advent of corpora, grammarians had a good idea of the grammatical structures possible in a language, but it was impossible to judge their relative frequency. Using corpora we can see which structures are fairly common and which are extremely rare in the language as a whole (very useful information in language pedagogy) and also how different types of discourse 'prefer' different modes of grammatical expression (for example, transitivity).

Several competing grammatical descriptions, particularly of English, have been tested using computers. These include Transformational Grammar (TG), Valency Grammar (VG) and Systemic-Functional Grammar (SFG). In this process, the system is first 'taught' the rules of the grammar and

then exposed to actual sentences to ascertain whether it responds appropriately (the working definition of 'understanding'). The grammatical description which has so far proved the most powerful is SFG and this will be the grammar we privilege in our analyses in this project.

Corpus study has also shown how grammatically 'creative' human language users are. A study of the use of *if* constructions in newspapers, for example, showed how they were used in a far greater variety of forms than listed in any current grammar of English.

Synonymy: corpora can shed light on the precise relations and subtle distinctions of use among members of a set of similar items, at first glance synonymous, such as, for instance, *completely, entirely, utterly, absolutely, perfectly* etc. This is important information, especially for non-native speakers.

False or true friends: for translation purposes, by interrogation parallel corpora of two languages, it is possible to test the reliability as translation equivalents of cognate items, such as, - taking English and Italian - *just* and *giusto*, *correct* and *corretto* etc.

Evaluative language: lexical items do not have just denotational meaning but also connotational or evaluative meaning. Corpus research is revealing that many more items than was previously suspected express a speaker's favourable or unfavourable attitude to the object of discourse, often unbeknown to the user. This can only be seen in the combinatorial behaviour of items, the kinds, the sets of other words/phrases it collocates with. It has been suggested that the study of these so-called hidden *semantic prosodies* can reveal instances of both irony and insincerity in the user, particularly in suasive discourses such as advertising and politics (Louw 1994).

Monogeneric corpora and the study of discourse

Research of this types generally entails the comparison of two or more corpora of a particular text-type and very often also the comparison of the contents of a monogeneric corpus with that of a heterogeneric one. In fact, discourse study is necessarily comparative in two separate but related ways. Firstly, within an individual discourse type, only by comparing the choices being made by speakers or writers at any point in a discourse with those which are normal, that is, usual within the genre, can we discover how *meaningful* those choices are. Testing observations and findings against corpus data can provide 'background information' against which particular events can be judged.

Secondly, if we are also interested in the characteristics and content of the discourse type itself, it is vital to be able to compare its particular features and patterns with those of other discourse types. In this way we discover *how* it is special, and can go on to consider *why*. All genre or discourse-type analysis is thus properly comparative. In the wider field of discourse studies, this requirement has unfortunately not always been observed in practice. Corpora provide the means and methodology to enable rigorous and principled comparative study to be performed.

The types of research possible using monogeneric corpora include the following:

Style and authorship studies: These generally attempt to identify distinctive characteristics of a particular author's writings. A recent development in this area is forensic linguistics which analyses written documents or transcripts in the attempt to provide evidence in legal cases of disputed authorship (Coulthard 1993, 1996).

Historical studies: So-called diachronic linguistics compares language from different periods in time to gather information on language change (Kytö and Rissanen 1990; Mair 1993).

Political science: Corpora have been used to study, *inter alia*, the following:

The typical language, metaphors and motifs used by participants in a number of political/institutional spheres, e.g. journalists and spokespersons in US press conferences, and how they reflect their respective world-views (Partington 2003);

The rhetoric of Berlusconi's electoral speeches (Garzone and Santulli 2004);

Differences among the UK quality newspapers in their stance on European Monetary Union (Vaghi and Venuti 2004);

How prediction is effected in economic texts, that is, how economic forecasts are presented and hedged (Walsh 2004);

The language of representative assemblies, or parliaments and the question of special discourse communities working within specific political institutions (Bayley 2004);

A comparison of the language of studio reporters, correspondents and embedded journalists from the BBC, CBS and the RAI during the conflict in Iraq to investigate various accusations of pro- and anti-war bias and pro- or anti-government leaning on the part of broadcasters (Clark forthcoming; Haarman 2004).

Corpus-Assisted Discourse Studies

The most relevant branch of Corpus Linguistics to the current *Intune* project is the nascent interdisciplinary school known as Corpus-Assisted Discourse Studies (CADS). This arose from the realisation that some of the methodology and instruments commonly used in corpus linguistics might be adapted for the study of features of discourse. In other words, that it was possible to combine the *quantitative* types of analysis used in corpus linguistics, which generally take into consideration large quantities of texts and subject them to statistical analysis, with the *qualitative* methods more typical of discourse studies which examine in detail much smaller amounts of discourse, frequently single texts (Haarman *et al* 2002). In this school of thought, research is 'a dynamic process which links together problems, theories and methods' (Bryman and Burgess 1994:4) and the researcher is free to shunt back and forth among hypotheses, data-collection, analysis, evaluation and even speculation, as long as these phases are kept separate and the movements among them are closely chartered.

References

Bayley, P. ed. 2004. *Cross-Cultural Perspectives on Parliamentary Discourse*. Amsterdam: Benjamins.

Bryman, A. and Burgess, R. eds 1994. *Analyzing Qualitative Data*. London: Routledge.

Clark, C. (forthcoming) Dodging the bullets and the brickbats: the embedded voice in the Iraq war. In José Maria Bernardo, Guillermo López y Pellegrì Sancho eds., *Análisis critico del discurso de los medios de comunicación de massas*, Valencia, Universidad de Valencia.

Coulthard, D. 1993. Beginning the study of forensic texts: Corpus, concordance and collocation. In Hoey, M. ed. *Data, Description Discorse*. London: Harper Collins, 86-97.

Coulthard, M. 1996. The official version: Audience manipulation in police records of interviews with suspects. In Caldas-Coulthard, C. and Coulthard, M. eds. *Texts and Practices*. London: Routledge, 166-178.

Garzone, G. and Santulli, F. 2004. What can Corpus Linguistics do for Discourse Analysis? In Partington, A., Morley, J. and Haarman, L. eds, 351-368.

Haarman, L. 2004. "What's going on, John?". Some features of live exchanges on television news. In Partington, A., Morley, J. and Haarman, L. eds, 71-88.

Haarman, L., Morley, J. and Partington, A. 2002. *Habeas Corpus*: methodological reflections on the creation and use of specialized corpora. In Gagliardi, C. ed. *Quality and Quantity in English Linguistics Research*. Pescara: Libreria dell'Università, 55-120.

Kytö, M. and M. Rissanen. 1990. The Helsinki Corpus of English texts: Diachronic and dialectal. *Medieval English Studies Newsletter* 23: 11-14.

Louw, B. 1993. Irony in the text or insincerity in the writer – The diagnostic potential of semantic prosodies. In Baker, M., Francis, G. and Tognini-Bonelli, E. eds. *Text and Technology*. Philadelphia and Amsterdam : John Benjamins, 157-176.

Mair, C. 1993. A corpus-based study of grammaticalization in present-day English: The case of *help*. Talk given at ESSE/2. Bordeaux: Université de Bordeaux II.

Morley, J. 2004. — 2004. The sting in the tail: persuasion in English editorial discourse. In Partington, A., Morley, J. and Haarman, L. eds., 239-255.

Partington, A. 2003. *The Linguistics of Political Argument*. London: Routledge.

Partington, A. and Morley, J. 2004. At the heart of ideology: Word and cluster / bundle frequency in political debate. In Lewandowska-Tomaszyk, B. ed. *Practical Applications in Language and Computers (PALC 2003)*. Bern: Lang, 1179-192.
Partington, A., Morley, J. and Haarman, L. 2004. *Corpora and Discourse*. Bern: Lang.
Vaghi, F and Venuti, M. 2004. Metaphor and the Euro. In Partington, A., Morley, J. and Haarman, L. eds., 369-382.
Walsh, P. 2004. Throwing light of prediction: Insights from a corpus of financial news articles. In Partington, A., Morley, J. and Haarman, L. eds., 335-348.