NEW PERSPECTIVES ON CORPUS LINGUISTICS

KEITH STUART
Dpt. de Lingüística Aplicada
Universidad Politécnica de Valencia

Abstract: Corpus linguistics has developed a battery of sophisticated linguistic and statistical techniques as the basis for an empirical approach to language study. This paper argues that these techniques may be applicable to other areas such as knowledge discovery in text. This paper firstly describes how corpus linguistics works and, secondly, discusses new perspectives on corpus linguistics in relation to the areas of knowledge discovery in text, terminology extraction and ontology building. Most explicit knowledge is written down in text. This seemingly obvious observation means that most explicit knowledge (and, perhaps, novel implicit knowledge) is to be found in collections of texts or corpora.

Key words: corpus linguistics, knowledge discovery in text, terminology extraction, ontology

Resumen: La lingüística de corpus ha desarrollado una batería de técnicas lingüísticas y estadísticas sofisticadas como la base para un acercamiento empírico al estudio de la lengua. Este trabajo argumenta que dichas técnicas pueden ser aplicadas a otras áreas tales como el descubrimiento del conocimiento en texto. En primer lugar, este trabajo describe las técnicas de la lingüística de corpus y, en segundo lugar, presenta nuevas perspectivas sobre la lingüística de corpus en áreas relacionadas con el descubrimiento del conocimiento en texto, la extracción de la terminología y la construcción de ontologías. La mayoría del conocimiento explícito se encuentra en textos. Esta observación aparentemente obvia significa que la mayoría del conocimiento explícito (y quizás, el nuevo conocimiento implícito) se encuentra en recopilaciones de textos o corpora.

Palabras claves: lingüística de corpus, descubrimiento del conocimiento en texto, extracción de terminología, ontología

1. INTRODUCTION

Corpus Linguistic techniques (for example, extracting word frequency statistics) provide quantitative approaches to the analysis of texts and enable comparisons of language patterns in texts: phonological, morphological, lexico-grammatical, discoursal, etc. Moreover, corpus analysis can offer linguistic data about register, genre and ideological underpinnings of texts. Corpus linguistics affords a more objective view of language than that of introspection and intuition because, as Sinclair (1998) has pointed out, speakers do not have access to the subliminal patterns which run through a language. Despite the fact that Sinclair's team on the COBUILD project have thrown light on many linguistic issues using corpus linguistic techniques, most of their research effort has been applied to producing

English language teaching materials. In general, the limitation of the potential of corpus linguistics to English Language Teaching has obviated that many computer-based tasks in the knowledge economy are using corpus linguistic techniques.

Some of these computer-based tasks involve choosing a piece of text (information retrieval, question-answering, and summarization) from a larger body of text. Some involve tagging a piece of text with a system of labels (part-of-speech tagging, genre recognition). Other tasks involve linking up two distinct pieces of texts (anaphora resolution, syntactic dependency extraction, clustering). Some other computer processes try to discover regularities (collocations, terminology extraction). In this article, we discuss approaches to corpus linguistics that mimic meaning-laden tasks that humans perform on text and are applicable in areas related to knowledge discovery in text. Before analysing what knowledge discovery in text means, we need to know how corpus linguistics works.

2. How Corpus Linguistics Works

The fairly recent availability of large quantities of digitized text and other data is changing the way many disciplines, from linguistics to genetics, are thinking about and practicing scientific research. Tognini-Bonelli (2001: 1) states that "what we are witnessing is the fact that corpus linguistics has become a new research enterprise and a new philosophical approach to linguistic enquiry" driven by massive amounts of data. "It is strange to imagine that just more data and better counting can trigger philosophical repositionings, but […] that indeed is what has happened" (Tognini-Bonelli 2001: 48). Empirical data about language has the ability to confirm or deny what until then may have only been hypothesized.

Corpora for linguistic research include *general corpus*, *monitor corpus*, *text collections*, and *data sets*. McEnery and Wilson (2001: 32) provide what they call a prototypical definition of a systematically designed and collected corpus: "a finite-sized body of machine-

readable text, sampled in order to be maximally representative of the language variety under consideration". Lindquist (1999) describes two types of corpora used in multilingual translation: *parallel corpora* and *translation corpora*. Parallel corpora consist of source texts and similar or related texts in target languages, while translation corpora are source texts and their translations into one or more target languages. Lindquist (1999: 182) argues that the parallel corpora model is especially powerful for translation because the translator can see "the words and collocations in actual use in the appropriate type of text", and thus the resulting translation "is likely to sound more natural than it would have done otherwise".

Corpora are built for a purpose. The nature of a corpus will be determined by its purpose. "A corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kind of research questions that can be addressed" (Biber *et al.* 1998: 246). A corpus that is representative is essential in corpus design so as to be able to make generalisations about the language of the target population the corpus aims to represent.

However, there are some research questions that are easier to resolve with a tagged or annotated corpus. In fact, tagging or encoding is one way of making a corpus more useful. As yet, there is no standard way to encode text corpora (however, see www.bnc.ac.uk). We can distinguish four levels of document and text mark-up:

Level 1: General Document mark-up:
- genre identification
- bibliographic description of the document: author, title, date of publication, journal name, author affiliation etc.

Level 2: General textual and structural mark-up:
- structural units of text, such as volume, chapter, etc., down to the level of paragraph
- quotations, footnotes, headings, subheadings, tables, figures, graphs etc.

Level 3: Contextual and linguistic annotation (partially language dependent):
- discourse annotation (anaphora resolution; cohesive devices)
- pragmatic annotation (speech act type)

- semantic annotation (semantic category of word)

Level 4: Language dependent annotation and mark-up for sentence level structures:
- sentences
- words
- abbreviations, names, dates, etc.
- morphological information
- syntactic information (part-of-speech tagging)
- prosodic annotation

An example of part-of-speech annotation is given below:

She_PNP told_VVD him_PNP to_TO0 clean_VVI his_DPS boots_NN2 as_CJS he_PNP was_VBD playing_VVG in_PRP the_AT0 match_NN1 that_DT0 afternoon_NN1 ._.

Level 1 provides global information about the text and its content. Level 2 includes universal text elements down to the level of paragraph, which is the smallest unit that can be identified language independently. Sentences can be identified by computers using the full stop as the criterion for a sentence. Paragraphs are identified by carriage returns or marked up as …. . Level 3 enriches the text with contextual information and linguistic analysis that may be intersentential. Level 4 enriches the text with the results of sentence-level linguistic analyses. Various corpus annotation tools (sentence segmenters, text tokenisation tools, morphological analysers, POS taggers, etc.) are available to make these tasks easier (see http://www.ling.ohio-state.edu/~dickinso/corpus.html).

There are also various tools to exploit or search corpora whether they are annotated or raw text. These include indexing tools which construct indexes for fast access to data; search and retrieval tools: concordancing, retrieval of collocations, etc., based on a given word or words, or on a lexico-grammatical pattern; statistical and quantitative tools that generate wordlists and statistics (basic frequency statistics for words, collocates etc.). The statistical and linguistic analysis of text that we have summarised here is the basis of much work being done on knowledge discovery in text.

3. KNOWLEDGE DISCOVERY IN TEXTS

There is, in addition to corpora for linguistic research, a need for consciously created and organized collections of data and information that can be used to carry out "knowledge discovery in texts" and to evaluate the performance and effectiveness of the tools for these tasks. According to Karanikas and Theodoulidis (2005: 2), "Knowledge Discovery in Text (KDT) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured textual data". I use *Knowledge Discovery in Text* (KDT) to express the search for knowledge or information that is unknown, hidden or implicit in semi-structured (documents or text that have been annotated or encoded with a mark-up language) and unstructured collections of text. Below are some of the kinds of KDT tasks that many subject disciplines are interested in:

- Identification and retrieval of relevant documents from one or more large collections of documents;
- Identification of relevant sections in large documents (*passage retrieval*);
- *Co-reference resolution*, i.e., the identification of expressions in texts that refer to the same entity, process or activity;
- *Extraction* of entities or relationships from text collections;
- Automated *characterization* of entities and processes in texts;
- Automated construction of *ontologies* for different domains (e.g., characterization of medical terms related to cancer, see Oncoterm);
- Construction of *controlled vocabularies* from fixed sets of documents for particular domains.

The need to construct controlled vocabularies for subject domains has meant that terminological extraction from corpora has become an important process in tasks related to knowledge discovery in text.

4. FROM CORPUS LINGUISTICS TO TERMINOLOGY EXTRACTION

In many disciplines, it has now been established that the maintenance of unambiguous terminologies or the comparison and aggregation of different terminologies goes through the building of formal specialized terminologies. Specialized terminologies are normally

obtained by building a corpus of domain knowledge and then using statistical and linguistic analysis to extract the necessary terms.

A first analysis has to be done to determine the optimal number of texts in order to cover most concepts and terms pertaining to the knowledge domain. There are no fixed rules for determining the overall size of any corpus for a particular purpose, and especially for terminology extraction. It is important to remember that if a corpus is smaller, it is intrinsically less reliable (Sinclair 1991: 13–20). Equally important in designing a corpus for extracting terminology is the question of representativeness. The representativeness of a corpus will affect the validity and reliability of the research and this will depend on the quality of the composition of the corpus. The composition should be determined by the purpose of the research (Biber *et al.* 1998: 246-250). A corpus for term extraction must represent the domain knowledge and language that is being investigated.

Once corpus design has been determined, criteria for eliminating non-technical terms have to be established. The simplest way of selecting terms is to use two steps. As a first step, general language items, for example function or grammar words, are removed, and then terms are selected according to frequency. Stop word lists can also be used to establish technical terms. More sophisticated systems involve the use of computer programmes to distinguish terms in tagged corpora, based on linguistic attributes such as word forms, parts of speech and syntactic structures of possible terms, and the statistical contrast between the frequencies of words in general texts and specialised texts (Kageura and Umino 1996: 259–289).

The latter (the statistical contrast between the frequencies of words in general texts and specialised texts) takes a corpus comparison approach between, for example, a general corpus such as the British National Corpus and a specialized corpus such as the PASTA project (see (Demetriou and Gaizauskas 2002). The starting point for the PASTA project was the creation of two corpora (a corpus of abstracts and a corpus of journal papers). The abstracts corpus

consists of about 1500 abstracts from a variety of relevant molecular biology journals. The full paper corpus consists of 300 journal papers, again from relevant molecular biology journals.

An important aspect to note is that these general and specialized corpora are different sizes; as a result the frequencies would need to be adjusted to make them directly comparable. This is likely to be fairly common when comparing general and specialized corpora. As general corpora represent the language as a whole, it will be bigger than a corpus of specialist knowledge.

The PASTA project extracts information on the roles of amino acid residues in protein active sites. The project consists of several components or modules that combine with each other to perform the following text processing tasks:

- Text Pre-processing
- Lexical and Terminological Analysis
- Syntactic Analysis and Meaning Representation
- Domain Modelling and Discourse Processing
- Template Extraction and loading into databases
- Design and implement a Web-based interface to the extracted protein structure database and to the original text sources

As can be seen from the above description of the project, the project goes beyond what is strictly a linguistic analysis. It is clear that lexical and terminological processes of analysis provide the basis for working towards the modelling of a knowledge domain. Much modelling of knowledge domains takes the form of ontologies.

5. From Terminology to Ontology

Another project of a similar nature is the Genia Project (see http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/). In the case of the Genia Project, the research project members refer to their investigation as ‘corpus-based knowledge acquisition’. Much work done in corpus annotation can be regarded as identifying and classifying the terms that appear

in the texts according to a pre-defined classification. For this purpose, the Genia Project first built a conceptual model (ontology) of substances (Proteins, DNAs, RNAs etc.) and sources (body part, tissues, cells etc.). Based on this ontology, the names of proteins, DNAs etc. and sources that appear in their corpus of abstracts are tagged. “These names are considered to be relevant to the description of biological processes, and recognition of such names is necessary for understanding higher level ‘event’ knowledge” (Ohta *et al.* 2001: 469). In other words, the detection and extraction of terminology from text is a first step towards knowledge acquisition (discovery) and organisation. Their strategy is to build language resources simultaneously by providing a mechanism for extending the pre-conceptualised existing ontological model and lexicon while annotating corpora. This process they call ‘Ontology-based corpus annotation’ (Ohta *et al.* 2001).

Ontological classification works well when you are dealing with domain-specific knowledge. Corpus-based ontological classifications are likely to work best on small very specific corpora that have formal categories and restricted entities. The design of the corpus will be such that it will have clear boundaries and represent precisely the knowledge that a research team is trying to capture. It is also likely that corpus design and development as well as the processing and analysis of the corpus will be informed by expert users from the knowledge domain. It may even be the case that there is a need to coordinate expert users from linguistics, computer science and, for example, genetics (if the corpus is in the area of genome research where the extraction of information about the micro-biology domain is the objective).

6. THE WEB AS CORPUS

However, you can also turn all this around. Up to now, we have been maintaining a fairly technical discussion laying the emphasis on corpus annotation as a means to making a

corpus useful for knowledge discovery and information extraction; what Leech (2004) calls giving 'added value' or enrichment of a corpus. For Sinclair (2004), the unannotated corpus or raw text is the 'pure' corpus. Likewise, he considers that the whole point of assembling a corpus is to gather data in quantity. The default value of quantity for a corpus is large. A corpus is assumed to contain a large number of words. This is a good description of the Web - the largest free and searchable corpus of maybe 10 trillion words that is available to anyone with an internet connection. Nobody knows exactly how many words are on the World Wide Web, but this figure of 10 trillion words appeared in *The Economist* on January 25th, 2005. The more you push in the direction of scale, spread, fluidity, flexibility, the harder it becomes to manage a classification system such as an ontology and yet there is so much knowledge waiting to be discovered on the web.

The World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette 2003). Some researchers collect frequency data directly from commercial search engines. Others use a search engine to find relevant pages, and then retrieve the pages to build a corpus. Others build a corpus by spidering the web and then go about managing the data that they have collected (Terra and Clarke 2003). Similarly, some prototypes of Internet search engines for linguists have been proposed (Elkiss and Resnik 2004). For example, Webcorp (http://www.webcorp.org.uk/) is capable of creating word concordances of website pages. Among the obvious uses of the web as corpus are the following:

- access to dictionaries, glossaries, thesaurus
- access to ontologies (e.g. WORNET)
- analysis of collocations
- analysis of noun groups (phrases) through a search engine such as Google
- comparative analysis of news reporting on the web
- construction of parallel corpora (many web pages are translated into various languages)
- mining the web to create minority language corpora
- study of emerging new lexical items (new uses of the language)
- study of knowledge construction on the web

- study of social networks on the web
- study of specialized corpora (academic, business, news, etc. corpora)
- study of web genres

On a theoretical level, there are aspects of the web as corpus that relate strongly to ideas proposed by Hoey (1991) in relation to the organisation of lexis in text. Hoey (1991: 31-32) observed that "if it is reasonable to describe a text in terms of something smaller than a text, then it might also be helpful to describe a text in terms of something larger than a text – a collection of texts". It is becoming clear from research by Barbasi (1999) and others on power law distribution of node linkages in a scale-free network that the way web pages are linked is similar to Hoey's description of the patterns of lexis in text. Hoey envisages meaningful choices organizing text as a network of links and bonds (lexical items form links, sentences sharing three or more links form bonds between different parts of text no matter how far apart they may be). It just may be that laws governing the web (the web seems to be self-organising through links, nodes and hubs rather than apparently unstructured) are applicable to language organisation in text and vice versa (a text as a network of links with nodes and hubs).

To combine a practical and theoretical note about the use of web as corpus, we can take on Leibniz's classical assumption about synonymy, according to which "two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made" (Miller 1990: 241). The web is a good source of evidence to demonstrate that synonymy, once defined in this way, is rare or does not exist at all.

7. CONCLUSION

While it may be argued that corpus linguistics is not really a domain of research but only a methodological basis for studying language, one can in fact use corpora as the basis

for an empirical approach to linguistics, while the same techniques are applicable to other areas of knowledge. Corpora are “reservoirs of evidence” (Tognini-Bonelli 2001: 55) that can be used in the scientific study of natural phenomena, phenomena ranging from natural human language to natural genetic language. Most explicit knowledge is written down in text. This seemingly obvious observation means that most explicit knowledge is to be found in collections of texts or corpora. What corpus linguistics might also be able to do is to mine for implicit knowledge by such a simple technique as concordance lines of a specific knowledge domain. There is really no discipline that cannot make use of corpora.

BIBLIOGRAPHY

Biber, D., S. Conrad, and R. Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Demetriou G. y R. Gaizauskas. 2002. “Utilizing Text Mining Results: The Pasta Web System”. Proceedings of the Association for Computational Linguistics Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, US, July 11. 77-84.

Elkiss, A., and P. Resnik. 2004. *The Linguist's Search Engine User's Guide*. [Internet document available at: http://lse.umiacs.umd.edu:8080/lseuser]

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Kageura, K., and Umino, B. 1996. “Methods of Automatic Term Recognition”. *Terminology* 3 (2): 259-289.

Karanikas, H., and B. Theodoulidis. 2005. “Knowledge Discovery in Text and Text Mining Software”. UMIST: Department of Computation Technical Report. Manchester. [Internet document available at: http://www.crim.co.umist.ac.uk/parmenides/internal/docs/Karanikas_NLDB2002%20.pdf]

Kilgarriff, A., y G. Grefenstette. 2003. “Introduction to the special issue on the Web as corpus”. *Computational Linguistics* 29(3): 333-347.

Leech, G. 2004. **“**Adding Linguistic Annotation”. Developing Linguistic Corpora: a Guide to Good Practice. Ed. M. Wynne. AHDS Literature, Languages and Linguistics, Oxford: Oxford University Computing Services. [Internet document available at: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm]

Lindquist, H. 1999. “Electronic corpora as tools for translation”. Word, Text, Translation. Eds. G. Anderman y M.Rogers. Clevedon: Multilingual Matters.

McEnery, T., and A.Wilson. 2001. *Corpus Linguistics: An Introduction* (2nd Ed). Edinburgh: Edinburgh University Press.

Miller, A. 1990. "WordNet: An On-line Lexical Resource". *Journal of Lexicography* 3(4): 235-244.

Ohta, T., Y. Tateisi, J. Kim, H. Mamma, and J. Tsujii. 2001. "Ontology Based Corpus Annotation and Tools". *Genome Informatics* 12: 469–470.

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

Sinclair, J. 1998. "Large corpus research and foreign language teaching". Language Policy and Language Education in Emerging Nations. Eds. R. de Beaugrande, M. Grosman y B. Seidlhofer. Volume LXIII in the series Advances in Discourse Processes. Stamford: Ablex. 79-86.

Sinclair, J. 2004. **"**Corpus and Text: Basic Principles". Developing Linguistic Corpora: a Guide to Good Practice. Ed. M. Wynne. AHDS Literature, Languages and Linguistics, Oxford: Oxford University Computing Services. [Internet document available at: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm]

Terra E., and C. Clarke. 2003. "Frequency estimates for statistical word similarity measures". In Proceedings of the Human Language Technology and North American Chapter of Association of Computational Linguistics Conference 2003, 244–251.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.