

Discussion document: A personal take on the ICT challenges and barriers in Linguistics

Paul Rayson

Director of UCREL, Lancaster University, UK

9th July 2006

1. Introduction

This document has been prepared to stimulate debate and discussion for the E-Science Scoping Survey Subject Expert Seminar for Linguistics. It is based on my viewpoint as a computer scientist looking at Linguistics having worked in the computational and corpus linguistics fields for a number of years. Therefore, it should not be taken as representative of all aspects of Linguistics. Feel free to disagree with any or all of the contents, in fact that is the main point. I present some of the key issues that I think we should focus on in the seminar. Please read this document before the first session on Tuesday 18th July!!



This document was written during the Digital Humanities 2006 (ALLC/ACH) conference in Paris. The first discussion question that occurs to me is whether we need a Digital Linguistics discipline, or does it already exist?

2. Impact of ICT on linguistics

Information and communications technology (ICT) and, in particular, computing technology has already had a huge impact on linguistics. Empirical linguistics experienced a revival in the late 1970s and early 1980s and this coincided with wider availability of personal computers, corpus software, the creation of electronic corpora and the increased use of corpus-based techniques. The impact of ICT is still ongoing alongside new developments in ICT itself, such as processing power, data storage, and network connectivity.

3. Linguistics research process

In order to place the challenges and barriers in context, I will first describe what I see as the typical linguistic research methodology using the corpus paradigm. We can define the term corpus linguistics as a methodology that can be applied to a wide range of linguistic study (McEnery and Wilson, 1996). This is ably demonstrated by Biber, Conrad and Reppen (1998) who describe corpus-based approaches in different areas of linguistics, including lexicography, grammar, discourse, register variation, language acquisition and historical linguistics. In all of these various areas of linguistic study, I claim that there are five core steps when we examine the corpus-based approach:

1. **Question:** A research question or model is devised
2. **Build:** Corpus design and compilation
3. **Annotate:** Computational analysis of the corpus
4. **Retrieve:** Quantitative and qualitative analyses of the corpus
5. **Interpret:** Manual interpretation of the results or confirmation of the accuracy of the model

There are at least three further stages to the research process that are typical across many, if not all, disciplines:

6. **Output:** Distil the research into a paper or presentation
7. **Dissemination:** Pass the paper to a publisher for printing, or submit to a conference for presentation
8. **Feedback:** Reviews of papers or presentations, citation practice influence the direction of future research

As outlined above in these five core steps, the process model of the corpus-based methodology typically used by researchers is as follows: it would begin with the identification of a research question (1), continue with building (2) and annotating (3) a corpus with which to investigate the topic, and finish with the retrieval, extraction (4) and interpretation (5) of information from the corpus which may help the researcher to answer the research question or confirm the parameters of the model. In some cases, the process may be an iterative one, where, following the interpretation of the results some refinement is needed on the research question or annotation of the corpus.

My process model, as described above, is in line with Leech's (1992) view of the corpus linguistic paradigm. Leech argues that the corpus-based methodology conforms to standards commonly ascribed to 'the scientific method': falsifiability, completeness, simplicity, strength, and objectivity.



Discussion point – how suitable is this process model if we are using usage-based models or a theoretical linguistics paradigm?

4. Where does ICT fit into this research process?

ICT has already had a major impact on linguistics and on corpus linguistics in particular. ICT is used in every stage of the above process model. Increasingly, researchers do not need to build and annotate (steps 2 and 3 above) their own corpus material. Instead they can use precompiled and annotated corpora that are available 'off-the-shelf' (Meyer 1991). Types of software tools to carry out compilation and annotation and retrieval from corpora (steps 2 and 3 in our process model) will be briefly reviewed in section 4.1. There are two major methods used in corpus linguistics to retrieve and interpret (steps 4 and 5) data from corpora. These are frequency profiling and concordancing, examined in section 4.2.

4.1 Corpus building and annotation

Current work: This stage may involve transcription, scanning and OCR, alignment of audio or video to transcripts, whether it is orthographic, phonetic, prosodic or other types of transcription. For example, the SCOTS project¹ uses Praat software to time align orthographic transcription with audio and video recordings, and the resulting corpus is loaded into a relational database

¹ <http://www.scottishcorpus.ac.uk/>

with a web front end for retrieval. By contrast, other texts may be born digital, e.g. student essays in the ICLE corpus², selected and collected for inclusion. Related to annotation, many tools have been developed in computational linguistics and natural language processing that are used for automatic annotation of data at the word-class, syntactic and semantic levels. Manual annotation, or manual correction of automatic annotation, extends this to the discourse and pragmatic analysis areas.

Challenges and barriers: Since the first electronic corpora were collected in the 1960s and 1970s, the size of the largest corpora has roughly increased by a factor of 10 every decade. We can see this trend from the Brown and LOB corpora (1 million words in the 1960s and 1970s), through the British National Corpus (100 million in the 1990s), to the Oxford English Corpus (1 billion words in 2006). A key challenge is making this amount of data usable for individual scholars, and facilitating such collection activities by those scholars. The use of the web as a corpus is becoming more common place³. How can such large amounts of data be cleaned, encoded, annotated, stored, and shared are all key questions. Clearance of copyright for web data as well as other corpus data is a vital issue. Multi-media corpus linguistics poses great technical challenges but also great opportunities to ask new research questions that haven't been possible before, e.g. investigating the storage of multiword expressions (MWE) by native and non-native speakers by examining the placement of pauses near the MWEs in speech (Second Language Speech Fluency and Multi-word Units Project⁴). Challenges for annotation are related to problems of tagsets and scale. *How acceptable is it to reuse standard tagsets on new data⁵? How can we avoid the annotation bottleneck for the web as corpus (Rayson et al, 2006)? Can we undertake collaborative annotation and share the effort and results via the internet?*



4.2 Corpus retrieval

Current work: In the area of retrieval software, I will focus on frequency profiling and concordancing tools. Search and retrieval software is more familiar to the general corpus user than any other: anyone who wishes to make use of a corpus is inevitably going to look for means to extract linguistic information from it. Associated with the concordancer will often be other facilities: providing frequency lists of word types, listing collocations based on mutual information or other measures, and furnishing information about subdivisions of the corpus, together with the incidences of linguistic phenomena in these. Many packages of this kind are available, some more advanced than others, and each tending to have its own special features.

Challenges and barriers: In addition, other disciplines such as digital humanities (humanities computing, literary studies) have their own well known

² <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm>

³ http://sslmit.unibo.it/~baroni/web_as_corpus_eacl06.html

⁴ <http://www.nottingham.ac.uk/english/research/cral/adolphs.htm>

⁵ <http://www.lboro.ac.uk/research/mmethods/research/software/dictionary.html>

tools (such as TACT) that offer similar functionality (Hockey, 2000). In the area of Computer-Assisted Qualitative Data Analysis Software (CAQDAS) similar tools are available for content analysis, discourse analysis and frame analysis methodologies⁶. Social science researchers have yet another view of the tools available for text statistics and concordancing⁷.



Can we open a dialogue with these other disciplines to produce better applications suitable for multiple users, or will general purpose tools be of any use for a specific purpose? Do you first need to decide on which methodology you want to use to analyse your data, and then match your analysis with the appropriate software?

We can distinguish three different generations of text retrieval software:

1. **Dos and Command-line:** e.g. WordCruncher, OCP, Childes-Clan, LEXA, TACT, MicroConcord
2. **Graphical user interfaces:** IMS-CorpusWorkBench, ICECUP, Xaira, WordSmith, MonoConc
3. **Web-based:** VIEW, BNCweb, SketchEngine, Wmatrix, Phrases in English (PIE)



What will the next generation of retrieval software be? Grid based? What are the challenges here? Can we share the data via the internet within the grey area of copyright restrictions?

4.3 Research question

Current work: There are two main kinds of research question (step 1 in the research process model) that can be investigated. Firstly, we can focus on the use of a particular linguistic feature, possibly a word or grammatical construction. I will call this type 1. Secondly, we can examine the characteristics of whole texts or varieties of language, and I will call this type 2. These two types are sometimes referred to as microscopic (type 1) and macroscopic (type 2), for example see Biber (1988: 61). Traditionally, studies tend to focus on type 1 and examine linguistic (lexical or grammatical associations of the feature), and non-linguistic aspects (distribution of the feature across different types of text or speech). Type 2 inverts this relationship in investigating, for example, register variation across text, by examining how certain features or groups of features characterise a text.

There are many examples of both types of research question in the many conference publications, journals and edited collections that have appeared. Common to both is the prior selection of which linguistic features to study.

Challenges and barriers: An alternative method could be proposed with appropriate ICT support: decisions on which linguistic features are important or should be studied are made on the basis of information extracted from the data itself; in other words, it is *data-driven*. I will call this type 3. It combines

⁶ <http://www.lboro.ac.uk/research/mmethods/research/software/caqdas.html>

⁷ <http://www.lboro.ac.uk/research/mmethods/research/software/stats.html>

For a wider review of text analysis software see <http://www.textanalysis.info/>

the approaches of types 1 and 2 by first focussing on whole texts and then suggesting specific linguistic features to study in further detail. In other words, the ordering of the five main steps above will change to the following (with iteration back from step 4 to step 3, which enables refinement of the research question following a retrieval step):

1. **Build:** Corpus design and compilation
2. **Annotate:** Computational analysis of the corpus
3. **Retrieve:** Quantitative and qualitative analyses of the corpus
4. **Question:** A research question or model is devised (iteration back to step 3)
5. **Interpret:** Manual interpretation of the results or confirmation of the accuracy of the model

The type 3 process model shown here is similar to that of *corpus-driven linguistics* as presented by Tognini-Bonelli (2001: 85), in which the corpus is the main informant. However, I decided to use the term data-driven to distinguish our approach from that of Sinclair, presented by Tognini-Bonelli (ibid: xi). The corpus-driven approach questions the “underlying assumptions behind many well established theoretical positions” (ibid: 48) stating that they need to be re-established or replaced based on evidence from corpora. For example, it proposes a “new unit of meaning” (ibid: 85) and states that there is “no such thing as a synonym”. In the corpus-driven approach, Stubbs (1993: 17) notes that even the traditional POS system “is under attack”.

Examples of type 3 research process are Ringbom (1998)⁸, Hoffmann and Lehmann (2000)⁹, the keywords method by Scott implemented in his WordSmith software (1996-onwards), Rayson (2003) and finally Leech and Fallon (1992) who also describe a two-stage research process which I would categorise as type 3.



How widespread is the use of the type 3 research process in linguistics, both corpus-based and otherwise?

5. Conclusion and e-science opportunities

Corpus linguistics has in the past focussed on English and in particular modern standard varieties of English. Challenges that have started to be addressed in the last five to ten years are the adaptation of the techniques and tools to non-English, historical and dialectal corpora (Archer et al, 2003;

⁸ Ringbom (1998) investigated advanced-learner language in the International Corpus of Learner English (ICLE) by comparing the essays produced by learners to those of native speakers. I have identified this as type 3 since Ringbom selected two verbs (get and think) for further study based on their overuse in frequency terms in the non-native speaker corpora when compared to the native speaker data.

⁹ I have classified Hoffmann and Lehmann (2000) as type 3 since they used collocational evidence from the British National Corpus to select pairs of related words that were then used in a study to discover native and non-native speakers’ familiarity with the word pairs. Due to the large size of the corpus, they selected collocation pairs with less than 100 occurrences to avoid problems of excessive computation. They used the log-likelihood statistic to select 150 collocations.

Beal et al, 2006). Variation in spelling is one of the biggest challenges facing corpus linguistic techniques over the next few years. Frequency profiling, concordancing, n-grams and keyword methods all suffer from problems of unreliability when applied to historical or dialectal corpora.

A vast amount of digitisation activity is being undertaken by commercial organisations: (e.g. Open Content Alliance, Google Print, Early English Books Online). However, it is unclear whether linguistics as a discipline can take advantage of this new resource. Individual scholars and research projects still need to digitise specific data sets for analysis. Sharing the pockets of expertise gained from individual projects is of vital importance, see for example the ICT Guides and case studies provided by the AHDS¹⁰.



Perhaps the largest challenge for linguistics as a discipline is to open or keep alive discussions with the other disciplines mentioned above and those of computer science, e-science and the semantic web. In order for academics involved in ICT developments to understand the requirements of linguistic research, they need to understand the language of linguists. Discuss!

References

- Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.
- Beal, J., Corrigan, K., Rayson, P. and Smith, N. (2006) Writing the Vernacular: Transcribing and Tagging the Newcastle Electronic Corpus of Tyneside English (NECTE). *Pre-conference workshop on corpus annotation, ICAME-27*, University of Helsinki, Finland, 24 May 2006.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.
- Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: investigating language structure and use*. Cambridge University Press, Cambridge.
- Hockey, S. (2000). *Electronic texts in the humanities*. Oxford University Press.
- Hoffmann, S. and Lehmann, H. M. (2000). Collocational evidence from the British National Corpus. In Kirk, J. M. (ed.) *Corpora galore: analyses and techniques in describing English*. Rodopi, Amsterdam, pp. 17 – 32.
- Leech, G. (1992). Corpus linguistics and theories of linguistic performance. In Svartvik, J. (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 – 8 August 1991*. Mouton de Gruyter, Berlin, pp. 105 – 122.
- Leech, G. and Fallon, R. (1992). Computer corpora – what do they tell us about culture? *ICAME Journal*, 16, Norwegian Computing Centre for the Humanities, Bergen, Norway, pp. 29 – 50.
- McEnery, T., and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

¹⁰ <http://www.ahds.ac.uk/creating/case-studies/index.htm> and <http://www.ahds.ac.uk/ictguides/>

- Meyer, C. F. (1991). A corpus-based study of apposition in English. In Aijmer, K. and Altenberg, B. (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. Longman, London, pp. 166 – 181.
- Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. *Ph.D. thesis*, Lancaster University.
- Rayson, P., Walkerdine, J., Fletcher, W.H. and Kilgarriff, A. (2006) Annotated Web as corpus. In proceedings of the 2nd Web as Corpus Workshop held in conjunction with the *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, April 3, 2006, pp. 27 - 33.
- Ringbom, H. (1998). High-frequency verbs in the ICLE corpus. In Renouf, A. (ed.) *Explorations in corpus linguistics*. Rodopi, Amsterdam, pp. 191 – 200.
- Stubbs, M. (1993). British traditions in text analysis: from Firth to Sinclair. In Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair*. Benjamins, The Netherlands, pp. 1 – 33.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Benjamins, The Netherlands.