

Review of CL2003, the International Conference on Corpus Linguistics
by Bayan Abu Shawar, School of Computing, University of Leeds, England

Lancaster is a small city in northwest England. The university campus is set amidst hills, woods, and green grass, and you can enjoy watching rabbits, ducks, and squirrels. It was sunny during the day and cold in the night for all participants who attended the CL2003 conference this year in Lancaster.

CL2003 is the International Conference on Corpus Linguistics, held at Lancaster University between 28 and 31 March 2003; 210 participants from many countries were there. The first International Conference on Corpus Linguistics, in 2001 (Rayson et al 2001), was organised to honour Geoffrey Leech on his 65th birthday; CL2003 was a repeat by popular demand. The conference included four one-day workshops, and four days of main papers sessions; one pre-conference workshop produced separate Proceedings (Simov and Osenova 2003), other papers were included in the main conference Proceedings (Archer et al 2003). Talks were presented in three different parallel sessions, classified according to topics such as: corpus building, parsing techniques, translation studies, corpus tools, taggers, morphosyntactic annotation, information extraction, semantics, text comparison, European minority languages, word frequency studies, internet English, and speech annotation. A wide range of domains related to corpus-based natural language processing was covered.

Five invited talks were presented; the first was ‘What can corpus linguistics tell us about linguistic creativity?’ by Michael Hoey. The second was ‘Everything you wanted to know about the American National Corpus..but weren’t afraid to ask’ by Nancy Ide. The third was ‘Frame, phrase, or function: a comparison between frame semantics and local grammars’ by Susan Hunston. Geoffrey Sampson addressed the issue “Are we nearly there yet, mum?” The last was ‘Corpora and the lexicon’ by Nicoletta Calzolari. All five talks were exciting, discussing new issues and ideas related to language and corpora.

There were too many presentations to discuss in a short review; the Proceedings (Archer et al 2003) ran to two thick volumes, so that most participants chose the CD-ROM version. Instead, I will present a summary of some papers relevant to my work:

Stefan Grondelaers presented “ a corpus-based approach to informality: the case of Internet chat”. He described the language of the Internet Relay chat as an example of “spoken language in written form”, as it shares with spoken language the informality characteristic. He classifies the informality characteristics into four types. The first is the dialogue character represented in the higher frequency of 2nd person pronouns and vocatives. The second is that the users type a lot of abbreviation and ellipses. A third source is speaker related: age, gender, and the topic of chatting. The fourth factor is register-related: chatters might choose to sound colloquial, or to maintain a more formal standard. Then he illustrated two methods to identify the informality, stylometric approaches and the onomasiological profile. Stylometric approaches are based on the calculation of isolated variables, used to identify the first three types of informality characteristics. Onomasiological profiles were used to compare lexical, morphological, syntactic and phonological preferences in Belgian IRC logs and other modes of written communication. The calculations revealed that the four types of informality do not coincide in Belgian Dutch IRC.

He concluded that the linguistic specificity of chat is determined in the first place by the production demands (speed and turn-taking efficiency) of spoken conversation.

Jun Takahashi examined interaction via computer mediated communication (CMC) within his paper "Do we talk or write differently over the net?" He reviewed previous work that described the characteristics of the language of CMC and the relationship between CMC and spoken and written language. These characteristics describe CMC as dynamic communication, classified into two types: synchronous such as IRC, which maintains strong coherence in turn-taking, like face-to-face conversation; and asynchronous such as emails and BBS. CMC looks like writing because users type to enter their discourse and deliver textual information; but the language is informal: a lot of pronouns and modal auxiliaries arise as in spoken language.

He investigated English as a multi-national language (EML) in the Internet discussion forum based on computer mediated communication. He gathered data from different corpora such as: the NET-EN Corpus (NC) which represented discussion in various domains, such as political, social, hobbies and sport; the Base text Corpus (BC), and Response text Corpus (RC). All these corpora show EML in use by Japanese English users.

Then he illustrated the analysis method he used by comparing the 50 most frequent lexical items between written, spoken, interviews & letters (Int&Lt) English from the native-speaker British National Corpus (BNC) and Net_EN corpus (NC).. Another aim was to find words with unique uses and investigate these in terms of sense and patterns of use. Finally he concluded that this study shows the unique use of English produced by Japanese users of English as multi-national language (EML) in today's context, computer mediated communication (CMC).

Another interesting paper was "*Relating lexical items to sociolinguistic features in a spontaneous speech corpus of Spanish*" by Jose Maria Guirao Miras et al. They presented the correlation between linguistic and socio-contextual features using the C-ORAL-ROM corpora, a multilingual corpus of spontaneous speech for four basic Romance languages; French, Italian, Portuguese and Spanish. Nine participants are working on this project, which is funded by the EU. The main goal of C-ORAL-ROM is to compare the four languages, on the same grounds, and provide comparative studies in different linguistic levels. In order to do the comparison, each subcorpus followed the same text distribution (sampling design and text size) and standard format (using XML to denote header and transcript). New computational tools give you analysis of the linguistic features of each sub-corpus in three levels: word, lemma, and POS using log-likelihood ratio. In the POS level, every word is related with speaker and the text, all socio-contextual information is stored in the file header, and depending on different header features many sub-corpora can be derived from the corpus, for example a male sub-corpus, a telephone sub-corpus, etc.

They illustrated a new idea based on an n-grams algorithm that is used at word level to extract multi-words candidates. All n-grams with three or more occurrences, for n=4, 3, and 2 were removed. Then a filter process is applied that discards all n-grams which start or end with determiner or auxiliary. Finally multi-words are selected by hand and each one is treated as a single lexical unit. They showed some results of this tool, for example according to POS analysis men prefer to use nouns and women prefer pronouns in the POS level with respect to the Spanish corpus.

The intention is to focus on comparison between the four romance languages after finishing the morphosyntactic annotation. This work will be published as a chapter in a book "Corpus Linguistics around the world".

Prof. Geoffrey Leech and Martin Weisser presented the SPAAC tool within their paper "Generic speech act annotation for task-oriented dialogues". SPAAC (Speech Act Annotated Corpus) is the XML semi-automatic tool developed to annotate dialogues in terms of speech acts. They applied this tool within a pilot project, which has attempted to achieve a middle ground between the aim to of general coverage of dialogues and the aim of annotating specific tasks or domains. Dialogue corpora were collected from two sources: first, BT OASIS dialogues which involves "100" calls to the operator and "150" calls to BT residential enquiries; second, the Train-Line dialogues which involves calls to a call centre providing railway timetable information and tickets, and seat reservation services.

They illustrated the main jobs of SPAAC as follows: automatic conversation of the text files containing the transcription to XML mark-up, interactive segmentation of the dialogues into utterance-units (called C-units), automatic assignment of speech-act categories, together with other categories giving information on form (declarative, yes-no question, imperative), polarity (positive, negative), topic (relating to train journey location, name, day, date, time) and mood (semantic categories such as probability, reason), and finally manual post-editing and correction of speech-act tags. One of the possible uses of speech act annotated corpora is as a training corpus for the modelling of dialogue behaviour, using either statistical, rule/structure-based or hybrid language model.

The conference also had several software demo sessions. For example, Paul Rayson and Olga Moudraia demonstrated the W-matrix tool. W-matrix was implemented by Dr. Rayson to compare different sized corpora at three levels: word, POS and semantic-tagged. The comparison results are viewed as frequency lists ordered by log-likelihood ratio, which indicates the most important differences between corpora.

One of the most important achievements for my research group and myself was, that *10 papers from University of Leeds were presented*. Eight students and postdocs of my group (Natural Language Processing) presented papers, in addition to our supervisor Eric Atwell who presented two papers. We have collected extended abstracts in a separate research report on "*Corpus Linguistics, Machine Learning and Evaluation: views from Leeds*" (Atwell et al 2003).

Belinda Maia and Luis Sarmiento won the prize for the best poster design., on "*Constructing comparable and parallel corpora for terminology extraction*"

Overall, the CL2003 conference was a nice opportunity to see the work of other people in different countries and to chat directly with them. Participants enjoyed the friendly, familiar environment, the delicious food and professional organization. Many thanks to: Tony McEnery, Dawn Archer, Paul Rayson, Andrew Wilson, Paul Baker, and all assistants from Lancaster University who organized the CL2003.

References

Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) 2003. Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University

Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds) 2001. Proceedings of the Corpus Linguistics 2001 conference. UCREL technical paper number 13. UCREL, Lancaster University

Kiril Simov and Petya Osenova (eds.) 2003. Proceedings of the The Workshop on Shallow Processing of Large Corpora (SProLaC 2003) held in conjunction with the Corpus Linguistics 2003 conference. UCREL technical paper number 17. UCREL, Lancaster University.