Workshop on Advanced Corpus Solutions *

Janne Bondi Johannessen
The Text Laboratory, Deaprtment of Linguistics and Nordic Studies, University of Oslo,
P.O.Box 1102 Blindern, N-0317 Oslo, Norway

jannebj@iln.uio.no

Abstract. The initiative for this Workshop on Advanced Corpus Solutions has been taken in order to focus on the need for corpora that take into account that many users are linguists and philologists who do not have an interest in technical matters.

Keywords: Corpora, user-friendly interfaces

1 Background

One of the main uses of language corpora is to assist linguists and language technologists in identifying the correct or representative language data within a certain domain. Linguists, unlike the other group, however, cannot be expected to be computationally advanced, and yet their research needs as complex data as the technological group does. This is not least true since one of the types of output of linguistic research often as a next step will be input to language technology systems.

While corpora used to be relatively simple and straightforward, perhaps varying along one dimension, such as written text genre (the Brown Corpus, the LOB corpus), and annotated according to one type, such as part of speech, the needs of linguists have risen in accordance with the possibilities that the technology offers.

Linguists have higher expectations nowadays, as they would like corpora to, for example, contain audio files of spoken language, dialect data, or videos, or to be multilingual. All content should be searchable, with possibilities for searching in the spoken language files for ways of expressing some word or grammatical category, or in the video files for types of gestures, or in the multilingual files, to see how one word or category in one language corresponds to another in a different language.

With more advanced corpora, users also expect annotations to be included. Part of speech is still an issue, but syntactically parsed corpora are also desired, as well as annotations relating to gestures, speech events, emotions etc. Spoken corpora should be transcribed, and there are expectations as to type of transcription (orthographic, phonetic). Sociolinguistic, geographical and historical variables are also on the agenda: sex, age and education are background variables that may distinguish linguistic types, and are therefore important factors that one might want to use as filters for different searches.

The human-machine interface is important. Few linguists accept search expressions to being produced in a language of regular expressions. The options should be clickable or be presented as choices from a menu. For larger areas, querying via maps would be a desirable option.

Corpus search issues are not the only important ones. Results handling is also something that researchers want. The results should be exportable straight to a database, statistics should be calculated, further annotations should be possible, maps should illustrate geographic distribution of hits.

While the list of desiderata is long, it turns out that few of the points are fulfilled in actual corpora. For example, spoken language corpora are often represented by transcriptions (even orthographic ones), but very rarely come with audio or video possibilities. Maps are still uncommon in connection with corpora.

In addition to the need for advanced individual corpora, there is also a growing interest in interoperability between corpora (as stated explicitly, for example, by the European CLARIN inititative).

* We are grateful to the organizers of PACLIC 24, who have kindly let us arrange this workshop in connection with the conferance.

2 About the papers at the workshop

When announcing this workshop we were hoping to get papers on corpora that would address one or more of the issues above, either because they provide principled solutions to some of the challenges, or because they have implemented exciting solutions to specific topics mentioned above within these areas:

- Corpus tools: corpus search, results presentation, results handling, linguistic annotation, text annotation
- Corpus types: monolingual corpora, parallel corpora, spoken language corpora, multimedia corpora

We were also hoping to get papers on tools for different languages, including, of course, Asian languages, given that the PACLIC conference is the Pacific Asia Conference on Language, Information and Computation. We are pleased to see that our hopes were fulfilled. We received papers from a variety of locations on a variety of topics. We give a very short presentation of each below. It will be seen that many of our formerly desired topics are also the focus of the workshop papers. Several papers deal with multilingual corpora, with speed in connection with very large corpora, with user-driven corpus features, and with annotation of advanced variables.

Eckhard Bick (University of Southern Denmark): 'Degrees of orality in speech-like corpora. Comparative annotation of chat and e-mail corpora.' This paper describes and evaluates the automatic grammatical annotation of a chat and an e-mail corpus of altogether 117 million words, using a modular Constraint Grammar system. It discusses a number of genre-specific issues, such as emoticons and personal pronouns, and offers a linguistic comparison of the two corpora with corresponding annotations of the Europarl corpus and the spoken and written subsections of the BNC corpus, with a focus on orality markers such as linguistic complexity and word class distribution.

Jan Pieter Kunst and Franca Wesseling (Meertens Institute, Amsterdam): 'Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools'. The paper discusses the DynaSAND database as a case study of a corpus tool. It also focusses on its implementation in other search engines, thereby illustrating how the underlying data is detached from their original interface and used in new ways.

Janne Bondi Johannessen, Joel Priestley and Anders Nøklestad (University of Oslo): 'A multilingual speech resource: The Nordic Dialect Corpus.' This paper describes the Nordic Dialect Corpus, a corpus that consists of transcribed spoken dialects, with audio and video, from five North European languages (Danish, Faroese, Finnish, Icelandic, Norwegian and Swedish). The paper focuses on recent developments that have been added as a result of wishes expressed by the linguist users. These include map views of various selections of search results, English translations of every dialect concordance, and search possibilities and presentation of both orthographic and phonetic transcriptions.

Johannes Goller (University of Munich): 'Parallel Suffix Arrays for Corpus Exploration.' This paper describes how recently developed techniques for suffix array construction and compression can be expanded to bring a new data structure, called a parallel suffix array, into existence, which is suitable as an in-memory representation of large annotated corpora, enabling complex queries and fast extractions of the context of matching substrings. It is also shown how parallel suffix arrays are superior to existing corpus search engines, in particular when sequential queries and corpora that are hard to tokenize are involved.

James Wilson, Anthony Hartley, Serge Sharoff and Paul Stephenson (University of Leeds). 'Advanced corpus solutions for humanities researchers'. This paper describes the design and implementation of an interface to corpora in 12 languages, stemming from the analysis of the needs of a diverse group of users: language teachers and language students, (non-computational) linguists, researchers in history and translation studies. The authors identified a set of requirements shared across the disciplines, as well as more specific requirements from the targeted user groups. The interface is designed to handle large-scale corpora of 20-500 million words.

Mitsuko Yamura-Takei, Miho Fujiwara and Etsuko Yoshida (Hiroshima Shudo University, Willamette University, Mie University). 'Entity Coherence in Comparable Learner Corpora: Seeking Pedagogical Insights'. This paper describes an ongoing collaborative project, between Japanese and U.S.

universities, that aims to build, analyze and use comparable learner corpora in an attempt to promote discourse-level proficiency in foreign language learning contexts. The focus is placed on discourse coherence created by reference to nominal and clausal entities. The corpus analysis results, within the framework of Centering Theory are presented, along with some pedagogical insights that teachers can utilize.

Milos Jakubicek, Adam Kilgarriff, Diana McCarthy and Pavel Rychlý (Masaryk University, Brno, and Lexical Computing Ltd., UK). ’Syntactic searching in very large corpora for many languages.’ For many linguistic investigations, the first step is to find examples. In the 21st century, the authors believe that they should all be found, not invented. Thus linguists need flexible tools for finding even quite rare phenomena. To support linguists well, they need to be fast even where corpora are very large and queries are complex. They present extensions to the CQL 'Corpus Query Language' for intuitive creation of syntactically rich queries, and demonstrate that they can be computed quickly within their own tool even on multi-billion word corpora.

3 The reviewers

We should mention that we received more papers than we could accept at the workshop, and had to reject some. In order to choose the best papers, we had to rely on the judgements of the reviewers. We are therefore extremely grateful for the work they have done. Below is the list of our helpful reviewers:

Wirote Aroonmanakun, Chulalongkorn University, Thailand
Emily M. Bender, University of Washington, USA
Eckhard Bick, University of Southern Denmark, Denmark
Francis Bond, Nanyang Technological University, Singapore
Lars Borin, Gothenburg University, Sweden
Ying Chen, China Agriculture University, China
Stefan Th. Gries, UCSB, St Barbara, USA
Stefan Evert, University of Osnabrück, Germany
Jan Pieter Kunst, Meertens Institute, Netherlands
Shoushan Li, Suzhou University, China
Kikuo Maekawa, The National Institute for Japanese Language, Japan
Adam Przepiórkowski, Polish Academy of Sciences, Poland
Franca Wesseling, Meertens Institute, Netherlands

They are worth the utmost gratitude.

References

Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press. (On the Brown Corpus)

Johansson, Stig, Geoffrey N. Leech and Helen Goodluck. Manual of information to accompany the Lancaster.Oslo/Bergen corpus of British English, for use with digital computers. University of Oslo. (On the LOB Corpus)