International Journal of Lexicography, Vol. 23 No. 2, pp. 249–255
doi:10.1093/ijl/ecq014

SUMMER 2010

EURALEX NEWSLETTER

Editor: Paul Bogaards
Email: p.bogaards@hum.leidenuniv.nl

The EURALEX Newsletter

This quarterly Newsletter is intended to include not only official announcements but also news about EURALEX members, their publications, projects, and (it is hoped) their opinions, and news about other lexicographical organizations. Please try to support this by sending newsletter contributions to Paul Bogaards at the email address above. The deadlines for spring (March), summer (June), autumn (September), and winter (December) issues are respectively 15 January, 15 April, 15 July, and 15 October annually.

The EURALEX Web Site

The URL of the EURALEX web site is www.euralex.org

Lexicography and LanguageTechnology in the Nordic countries. Report from a Symposium in Copenhagen January 29 to 31, 2010

The annual symposium arranged by The Nordic Association of Lexicography was held at Schæffergården outside Copenhagen, Denmark. The symposium was hosted by The Nordic Language Council represented by Rikke Hauge, Norway, and by the editors of LexicoNordica Ruth Vatvedt Fjeld, Norway, and Henrik Lorentzen, Denmark. The primary aim of the symposium was to get an overview of existing computational, lexical resources in the Nordic languages and to discuss new initiatives to be carried out in the field. By bringing together researchers from lexicography and language technology, respectively, the intention was to encourage more synergy and cooperation between the related fields. An additional aim of the event was to address the use of language technology and corpus tools in the dictionary making process. 28 participants from Denmark, Finland, Iceland, Norway and Sweden attended the symposium. 11 presentations within the fields of lexicography and language technology were given, all followed by a lively discussion.

Eckhard Bick, University of Southern Denmark, opened the symposium by presenting a new type of lexical resource, DeepDict (at www.gramtrans.com), built from grammatically analyzed corpus data. Co-occurrence strength between mother-daughter dependency pairs is used to automatically produce dictionary entries of typical complementation patterns and collocations, in the fashion of an instant monolingual usage dictionary. DeepDict is capable of abstracting lemma relations and semantic classes from inflected surface forms, and provides concordances and statistics for the relations found. Entries are supplied to the user in a graphical interface with various thresholds for lexical frequencies as well as absolute and relative co-occurrence frequencies. DeepDict draws its data from Constraint Grammar-analyzed corpora, ranging between tens and hundreds of millions of words, covering the major Germanic and Romance languages, among them Swedish, Danish and Norwegian. Apart from its obvious lexicographical uses, DeepDict also targets teaching environments and translators.

Trond Trosterud, University of Tromsø, Norway, stated in his talk that small language societies are forced to find synergy between the fields of formal grammar, lexicography and language technology in order to make best use of the scanty available labour resources. He illustrated this point by means of two practical examples; that of (i) Komi, a Finno-Ugrian language spoken in the Northwestern part of Russia, and that of (ii) Faroese, a Scandinavian language spoken in the Faroe Islands. For both languages, initiatives have been taken with the aim of combining lexicographical and language technology efforts into joint language resources.

This talk was followed by *Lars Borin*, Språkbanken, University of Gothenburg, Sweden, who presented an integrated lexical resource for Swedish language technology. More often than not, digital resources that result from individual research projects will languish once the project ends and the publications are out. Lack of funding for resource maintenance, non-interoperability due to lack of standards, and closed-content license formats are probably the main reasons for this. The Swedish Language Bank, University of Gothenburg, is embarking upon a project to integrate a number of existing free lexical resources into a new open-content resource for Swedish language technology applications, in the process of salvaging some existing resources from undeserved non-use. Interoperability among resources is ensured by linking them using the standardized persistent sense and lemgram identifiers of the SALDO lexicon project (spraakbanken.gu.se/sal). The resource integration process will be largely automatic, with some manual post-processing. On top of the integrated resource, a Swedish framenet will be defined (spraakbanken.gu.se/swefn), through a considerable amount of automatic language processing and reuse of linguistic knowledge already encoded in the existing resources.

Anna Björk Nikulásdóttir, University of Iceland, presented a semantic database for Icelandic language technology. Developing semantic networks with semantic mining is one of the topics of a project funded by the Icelandic research fund RANNÍS called *Viable Language Technology beyond English – Icelandic as a Test Case*. The research topic is divided into two tasks: (i) collection of data through automatic extraction of semantic relations and (ii) exploiting of possibilities to semi-automatically combine the extracted relations and other available resources to build a semantic database. In the presentation several approaches to automatic extraction of semantic relations were discussed, including extraction from dictionaries, pattern based extraction from text, computing of semantic similarity and clustering. Finally, possible combination methods to extend and validate results were touched upon.

Bolette Sandford Pedersen, University of Copenhagen, Denmark, spoke about semantic language resources for language technology purposes and started out by stating that lexicography and language technology have not always followed the same lines, the two disciplines stemming from very different traditions. The discipline of language technology originates from the generative language paradigm where lexicon was originally not in focus. With lexicalist theories like Lexical Functional Grammar and Head-Driven Phrase Structure Grammar, and with recent more corpus-based approaches to language technology, however, the focus has changed, giving the lexicon a more central position in this field. As an example of a project which fully utilizes the synergy between lexicography and language technology, she presented DanNet (wordnet.dk), a wordnet which has been semi-automatically produced on the basis of Den Danske Ordbog (a medium-sized, corpus-based dictionary of modern Danish). She discussed several issues in relation to this reuse, such as the adjustment of underspecified hyponymies as well as the insertion of information not required in the source dictionary meant for human users, but necessary for a resource meant for computational use.

Krister Lindén and *Lauri Carlson*, University of Helsinki, Finland, presented a wordnet for Finnish through translation of Princeton WordNet. A wordnet consists of sets of synonyms, i.e. synsets, with the same part of speech that can be interchanged in a given context. The context of the synset is provided in a gloss exemplifying or describing the meaning of the synset. Synsets in a wordnet have hierarchical partial orderings according to semantic relations, e.g. hypernyms. These parameters fixate the meaning of a word and constrain the possible translations of a word in a given synset. University of Helsinki has opted for translating Princeton WordNet 3.0 synsets wholesale into Finnish, because the translation process is controlled with respect to quality, coverage, cost and speed, yielding results within six months. Several professional translators were used in parallel. According to the evaluation results, the translation process was diligent and the translators did their best. In the beginning, the translators provided fairly complete sets of synonyms. Towards the

end of their synset collections, the translators tended to revert to a standardized single-best translation mode perhaps prompted by the fact that the majority of English synsets are terms.

Anders Nøklestad from Oslo University, Norway, gave a talk on the use of a Norwegian lexicon for tagging and other language technology purposes. Norsk ordbank (the Norwegian Word Bank) is an electronic lexicon for the two Norwegian written standards, Bokmål and Nynorsk. It forms the basis of many, probably most, of the existing language technology tools for Norwegian. The lexicon is based on the entries and inflectional information found in the dictionaries Bokmålsordboka and Nynorskordboka as well as word lists and inflectional patterns developed by IBM Norway. The speaker presented some background information about the lexicon and showed how it has been applied to a variety of language technology tools and various applications for end users. Since the lexicon was developed from resources meant for use by human readers, much work has been devoted to modifying the lexicon to make it better suited for use in language technology.

Jakob Halskov, Danish Language Council, Denmark, presented a study on semiautomatic selection of lemma candidates for a dictionary of neologisms. One of the key tasks of the Danish Language Council is to monitor, record and document linguistic changes in the Danish language. Since the volume of language production has increased dramatically with the digital revolution and the advent of the Internet, a prototype neologism detector called the Ordtrawler (Word Trawler) is being developed. The system processes vast amounts of text and extracts candidate neologisms using a combination of simple filters (e.g. head words and inflected forms from existing dictionaries), collocational statistics (versus a reference corpus predating the analysis corpus) and so-called neology markers (e.g. inverted commas). A thorough evaluation of the Ordtrawler indicates that neology markers are good for optimizing system precision, and although they drastically reduce recall, this is less important when vast amounts of text is available. While certain types of noise such as technical terms, occasionalisms and semantically transparent compounds remain to be tackled, introducing diachronic frequency profiling appears to be a promising solution.

Christian Sjögreen and *Emma Sköldberg*, University of Gothenburg, Sweden, presented a study on dictionary writing systems for monolingual Swedish dictionaries, i.e. software for writing and producing dictionaries. First, the necessary features of such systems were discussed followed by a discussion of which systems are used in Sweden as well as internationally. Due to the fact that dictionary writing systems are expensive to develop and maintain, the situation in Swedish publishing houses is unsatisfactory. In Swedish lexicographic institutes though, tailor-made systems have been developed. The rest of the presentation concerned a comparison between the two Swedish systems used at Centre for Lexicology and Lexicography at the University of Gothenburg.

One of them is used to compile Swedish words in the Lexin-project, the other in the project Svensk Ordbok published by the Swedish Academy. The two systems are rather different, although both are developed in close co-operation between lexicographers and developers.

Eiríkur Rögnvaldson, University of Iceland, gave an overview of Icelandic language technology since its inception ten years ago. In 2000, the Government launched a special Language Technology Program with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in the creation and development of several important resources and tools which have had profound influence on Icelandic language technology, and are also valuable for Icelandic lexicography and linguistic research in general. Some of the most important of these products were discussed, such as a morphological database (260,000 lemmas), a 25 million word balanced and PoS tagged corpus, a lemmatizer, a rule-based tagger, and a shallow parser. Finally, it was pointed out that all the tools that the Icelandic Language Technology Community has developed in the past few years have been made open source, and the importance of adopting open source policy for small language communities was emphasized.

Viggo Kann, Kungliga Tekniska Högskolan (KTH), Sweden, talked about morphological and lexicographical tools and resources. During the last 15 years the human language technology group at KTH has developed tools and resources that may have an interest to the lexicographical community. Several tools were developed as part of the group's research on Swedish authoring tools: spelling error detection and correction, grammar checking, part-of-speech tagging, lemmatization, compound splitting, and an interactive learning environment called Grim. Most of the tools are made open source and may be downloaded from www.csc.kth.se/theory/humanlang. The group has also made several dictionaries available on the web: the Lexin series of dictionaries for 15 languages, the Scandinavian Dictionary, the Tvärslå dictionary collection, the Swedish Hyphenation Dictionary and the two crowd-sourced resources The People's Dictionary of Synonyms and The People's English-Swedish Dictionary.

The final discussion focused on synergy and standardization among Nordic lexical databases as well as on the availability and openness of the given resources and tools. One hindrance to harmonization seems to rely on the fact that several of the Nordic language technology resources and tools are developed in relation to time-limited projects with very specific aims and duration. In relation to the further development of *semantic* lexical resources, it was decided to oppose this obstacle by starting a collaborative initiative on evaluation and harmonization. A first step in order to achieve funding for this initiative was taken by applying for a NordCorp project under The Joint Committee for Nordic Research Councils for the Humanities and the Social Sciences (NOS-HS). The next symposium will take place in the beginning of 2011 and

will further take up the topic of semantic lexical resources by addressing in particular onomasiological dictionaries in the Nordic countries.

Bolette Sandford Pedersen
Denmark

Forthcoming events

2010

July
6–10, Fryske Akademy, Leeuwarden (The Netherlands), 14th EURALEX International Congress. Information at http://www.euralex2010.eu.

July
19–21, Gaborone, Botswana, AFRILEX 2010 (15th International Conference of the *African Association for Lexicography*). Pre-conference workshop theme: Traditions, trends and changes in lexicography (Presenters: Prof. D. J. Prinsloo, M. S. Mogano et al.) Conference keynote speaker 1: Prof. Robert Lew (Poznan, Poland). Conference keynote speaker 2: Dr Anderson Chebanne (Gaborone, Botswana). Latest info: http://afrilex.africanlanguages.com/

September
15–17, Universität Leipzig, Leipzig, Germany: SprachRäume (Annual Conference of German Applied Linguistics Society, GAL 40, with Section on Lexicography). Information at: www.gal-ev.de/jahrestagung-2010-leipzig.html

September
20–22, Campus Catalunya, Universitat Rovira i Virgili, Terragona, Spain: Congreso Internacional de Lexicografía Hispánica (AELex 4). Information at: www.urv.net/congressos/lexicografia_hispanica/es_indice.html

2011

January
17–21, Santiago de Cuba. Twelfth international symposium on social communication. Information at: www.santiago.cu/hosting/linguistica/Index.html

May
24–27, University of Lund, Lund, Sweden: Biennial Conference of the Nordic Association for Lexicography (NFL 11). Information at: www.nordisk-sprakrad.no/nfl.htm

June
8-11, McGill University, Montréal, Quebec, Canada: Biennial Meeting of the Dictionary Society of North America (DSNA 18). Information at: www.dictionarysociety.com/

August
Kyoto, Japan: Biennial Conference of the Asian Association for Lexicography (ASIALEX 7). Information at: www.asialex.org/

August
23–28, Beijing Foreign Studies University, Beijing, China: Harmony in Diversity: Language, Culture, Society (16th World Congress of Applied Linguistics, with Research Network No. 22 on Lexicography & Lexicology: Online Dictionaries in Linguistics and Communication Science). Information at www.aila2011.org/en/

November
4–10, Barcelona, Spain: Triennial Congress of the International Council of Onomastic Sciences (ICOS 24). Information at: www.icosweb.net/