COLING 2012

24th International Conference on Computational Linguistics

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)

Workshop chairs:
Michael Zock and Reinhard Rapp

15 December 2012
Mumbai, India

Diamond sponsors

Tata Consultancy Services
Linguistic Data Consortium for Indian Languages (LDC-IL)

Gold Sponsors

Microsoft Research
Beijing Baidu Netcon Science Technology Co. Ltd.

Silver sponsors

IBM, India Private Limited
Crimson Interactive Pvt. Ltd.
Yahoo
Easy Transcription & Software Pvt. Ltd.

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)
Michael Zock and Reinhard Rapp (eds.)
Revised preprint edition, 2012

Published by The COLING 2012 Organizing Committee
Indian Institute of Technology Bombay,
Powai,
Mumbai-400076
India
Phone: 91-22-25764729
Fax: 91-22-2572 0022
Email: pb@cse.iitb.ac.in

Also available online in the ACL Anthology at http://aclweb.org

Introduction to the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)

Encouraged by the enthusiasm and interest expressed by the participants of COGALEX-I (co-located with COLING 2008 in Manchester)[1] and COGALEX-II (co-located with COLING 2010 in Beijing)[2] it was natural to come up with a follow-up workshop. As with the preceding events (including the workshop "*Enhancing and Using Electronic Dictionaries*" held in conjunction with COLING 2004 in Geneva),[3] our aim was to provide a forum for computational lexicographers, researchers in NLP, and industrial practitioners to share their knowledge concerning the construction, organisation and use of a lexicon by people (lexical access) and machines (NLP, IR, data-mining). However, given the progress in various fields outside of linguistics (biology, psycholinguistics, neuro-sciences, network sciences, etc.) we decided to broaden the scope by inviting researchers from other domains, as we believe their work to be relevant.

Dictionaries store knowledge concerning words. Obviously, they should be comprehensive and complete enough to reveal the meaning of words (analysis), their form or other related information relevant for language producers (speakers, writers). Yet, the quality of a dictionary depends not only on *coverage*, but also on *accessibility* of information. Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the moment of consultation (words, concepts, speech sounds). Unlike readers who look for meanings, writers start from them, searching for the corresponding words. While paper dictionaries are static, permitting only limited strategies for accessing information, their electronic counterparts promise dynamic, proactive search via multiple criteria (meaning, sound, related words) and via diverse access routes. Navigation takes place in a huge conceptual lexical space, and the results are displayable in a multitude of forms (e.g. as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

The way we look at dictionaries (their creation and use) has changed dramatically over the past 30 years. While being considered as an appendix to grammar in the past, they have in the meantime moved to centre stage. Indeed, there is hardly any task in NLP which can be conducted without them. Also, rather than being static entities (data-base view), dictionaries are now viewed as graphs, whose nodes and links (connection strengths) may change over time. Interestingly, properties concerning topology, clustering and evolution known from other disciplines (society, economy, human brain) also apply to dictionaries: everything is linked, hence accessible, and everything is evolving. Given these similarities, one may wonder what we can learn from these disciplines. In the 3rd edition of the CogALex workshop we therefore intended to also invite scientists working in these fields, our goals being to broaden the picture, i.e. to gain a better understanding concerning the mental lexicon and to integrate these findings into our dictionaries in order to support navigation. Given recent advances in neurosciences, it appears timely to seek inspiration from neuroscientists studying the human brain. There is also a lot to be learned from other fields studying graphs and networks, even if their object of study is something else than language, for example biology, economy or society.

[1] Workshop proceedings (in ACL anthology): http://www.aclweb.org/anthology/W/W08/#1900
[2] Workshop proceedings (in ACL anthology): http://aclweb.org/anthology-new/W/W10/#3400
[3] Workshop proceedings (in ACL anthology): http://aclweb.org/anthology-new/W/W04/#2100

We agree with van Deemter and colleagues[4] when they write "... computational and psycholinguistic approaches to reference production can benefit from closer interaction, and this is likely to result in the construction of algorithms that differ markedly from the ones currently known in the computational literature.". One might add that the same is true for many areas of NLP, including the lexicon. This is in line with Krahmer's[5] inspirational paper 'What computational linguists can learn from psychologists (and vice versa)' which was published in the Computational Linguistics journal.

This workshop is about possible enhancements of existing electronic dictionaries. To perform the groundwork for the next generation of electronic dictionaries we invited researchers involved in the building of such dictionaries. The idea is to discuss modifications of existing resources by taking the users' needs and knowledge states into account, and to capitalize on the advantages of the digital media. For this workshop we invited papers including but not limited to the following topics which can be considered from various points of view: linguistics, neuro- or psycholinguistics (tip of the tongue problem, associations), network related sciences (sociology, economy, biology), mathematics (vector-based approaches, graph theory, small-world problem), etc.

Analysis of the conceptual input of a dictionary user

- What does a language producer start from (bag of words)?
- What is in the authors' minds when they are generating a message and looking for a word?
- What does it take to bridge the gap between this input and the desired output (target word)?

The meaning of words

- Lexical representation (holistic, decomposed)
- Meaning representation (concept based, primitives)
- Revelation of hidden information (vector-based approaches: LSA/HAL)
- Neural models, neurosemantics, neurocomputational theories of content representation.

Structure of the lexicon

- Discovering structures in the lexicon: formal and semantic point of view (clustering, topical structure)
- Creative ways of getting access to and using word associations
- Evolution, i.e. dynamic aspects of the lexicon (changes of weights)
- Neural models of the mental lexicon (distribution of information concerning words, organisation of words)

Methods for crafting dictionaries or indexes

- Manual, automatic or collaborative building of dictionaries and indexes (distributional semantics, crowd-sourcing, serious games, etc.)
- Impact and use of social networks (Facebook, Twitter) for building dictionaries, for organizing and indexing the data (clustering of words), and for allowing to track navigational strategies, etc.
- (Semi-) automatic induction of the link type (e.g. synonym, hypernym, meronym, association, collocation, ...)

[4] van Deemter, K., Gatt, A., van Gompel, R. & Krahmer, E. (2012). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4 (2), 166–183.

[5] Krahmer, E. (2010). What computational linguists can learn from psychologists (and vice versa). *Computational Linguistics*, 36 (2), 285–294.

- Use of corpora and patterns (data-mining) for getting access to words, their uses, combinations and associations

Dictionary access (navigation and search strategies), interface issues

- Semantic-based search
- Search (simple query vs multiple words)
- Context-dependent search (modification of users' goals during search)
- Recovery
- Navigation (frequent navigational patterns or search strategies used by people)
- Interface problems, data-visualisation

We received 22 submissions, of which ten were accepted as full papers, while six were chosen for poster presentation. While we did not get papers on all the issues mentioned in our call, we did get a quite rich panel of topics including cognitive approaches to lexical access, considerations on word meaning and ontologies, manual and automatic approaches for constructing lexicons, as well as pragmatic aspects.

It was also interesting to see the variety of languages in which these issues are addressed. The proposals range from European languages such as Bulgarian, Dutch, English, French, German, Italian, Polish, Romanian, Russian, and Spanish to Asian languages including Assamese, Bangla, Bodo, Chinese, Hindi and Japanese. In sum, the community working on dictionaries is dynamic, and there seems to be a growing awareness of the importance of some of the problems presented in our call for papers.

We would like to thank Alain Polguère for having accepted to be our invited speaker, and the COLING organizers, in particular publication chair Roger Evans, for providing the framework and for their support. We would also like to express our sincerest thanks to all the members of the Programme Committee whose expertise was invaluable to assure a good selection of papers, despite the very tight schedule. Their reviews were helpful not only for us to make the decisions, but also for the authors, helping them to improve their work. In the hope that the results will inspire you, provoke fruitful discussions and result in future collaborations.

Michael Zock and *Reinhard Rapp*

Organizers:

Michael Zock (LIF-CNRS, Marseille, France)
Reinhard Rapp (LIF, Marseille, France & University of Mainz, Germany)

Invited Speaker:

Alain Polguère (Université de Lorraine, ATILF, France)

Programme Committee:

Eduard Barbu (Universidad de Jaén, Spain)
Alain Barrat (Centre de physique théorique, CNRS & Aix-Marseille Université, France)
Gemma Bel-Enguix (LIF, Aix-Marseille Université, France)
Pierrette Bouillon (TIM, Faculty of Translation and Interpretating, Geneva, Switzerland)
Paul Cook (The University of Melbourne, Australia)
Dan Cristea (University of Iasi, Romania)
Cedrick Fairon (CENTAL, Université catholique de Louvain, Belgium)
Afsaneh Fazly (University of Toronto, Canada)
Christiane Fellbaum (University of Princeton, USA)
Olivier Ferret (CEA LIST, Palaiseau, France)
Thierry Fontenelle (Translation Centre for the Bodies of the European Union, Luxemburg)
Sylviane Granger (Université Catholique de Louvain, Belgium)
Gregory Grefenstette (3DS Exalead, Paris, France)
Silvia Hansen-Schirra (University of Mainz, FTSK, Germany)
Ulrich Heid (University of Hildesheim, Germany)
Graeme Hirst (University of Toronto, Canada)
Ed Hovy (ISI, Los Angeles, USA)
Terry Joyce (Tama University, Kanagawa-ken, Japan)
Olivia Kwong (City University of Hong Kong, China)
Marie Claude L'Homme (OLST, University of Montreal, Canada)
Guy Lapalme (RALI, University of Montreal, Canada)
Verginica Mititelu (RACAI, Bucharest, Romania)
Vito Pirrelli (ILC, Pisa, Italy)
Alain Polguère (Université de Lorraine, ATILF, France)
Reinhard Rapp (LIF Marseille, France & University of Mainz, Germany)
Tom Ruette (KU Leuven, Belgium)
Didier Schwab (LIG, Grenoble, France)
Gilles Sérasset (IMAG, Grenoble, France)
Serge Sharoff (University of Leeds, UK)
Anna Sinopalnikova (FIT, BUT, Brno, Czech Republic)
John Sowa (VivoMind Research, LLC, USA)
Carole Tiberius (Institute for Dutch Lexicology, The Netherlands)
Takenobu Tokunaga (TITECH, Tokyo, Japan)
Dan Tufis (RACAI, Bucharest, Romania)
Alessandro Valitutti (University of Helsinki and HIIT, Finland)
Piek Vossen (Vrije Universiteit, Amsterdam, The Netherlands)
Eric Wehrli (LATL, University of Geneva, Switzerland)
Michael Zock (LIF, CNRS & Aix-Marseille Université, France)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay & ERTIM-INALCO, Paris, France)

Table of Contents

3rd Workshop on Cognitive Aspects of the Lexicon

Program

Saturday, 15 December 2012

09:00–0905	**Opening Remarks**
09:05–1000	**Invited Presentation** *Like a Lexicographer Weaving Her Lexical Network* Alain Polguère
10:00-11:30	**Session 1: Cognitive Approaches**
10:00–10:30	*Long Tail in Weighted Lexical Networks* Mathieu Lafourcade and Alain Joubert
10:30–11:00	*On discriminating fMRI representations of abstract WordNet taxonomic categories* Andrew Anderson, Tao Yuan, Brian Murphy and Massimo Poesio
11:00–11:30	*Automatic index creation to support navigation in lexical graphs encoding part_of relations* Michael Zock and Debela Tesfaye
11:30–12:00	Tea break
12:00-13:30	**Session 2: Word Meaning and Ontological Considerations**
12:00–12:30	*Modeling Word Meaning: Distributional Semantics and the Corpus Quality-Quantity Trade-Off* Seshadri Sridharan and Brian Murphy
12:30–13:00	*Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes* Francesca Frontini, Irene de Felice, Fahad Khan, Irene Russo, Monica Monachini, Gloria Gagliardi and Alessandro Panunzi
13:00–13:30	*Dictionary-ontology cross-enrichment* Emmanuel Eckard, Lucie Barque, Alexis Nasr and Benoît Sagot
13:30–14:30	Lunch

Saturday, 15 December 2012 (continued)

14:30–15:30	**Session 3: Crafting Lexicons, Manual and Automatic Approaches**
14:30–15:00	*Automatic Construction of a MultiWord Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective* Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum
15:00–15:30	*Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor* Nabil Gader, Veronika Lux-Pogodalla and Alain Polguère
15:30–16:30	**Session 4: Posters with Booster Session**
15:30–15:35	*A Procedural DTD Project for Dictionary Entry Parsing Described with Parameterized Grammars* Neculai Curteanu and Mihai Alex Moruz
15:35–15:40	*Automatic Generation of the Universal Word Explanation from UNL Ontology* Khan Md Anwarus Salam, Hiroshi Uchida and Tetsuro Nishino
15:40–15:45	*Towards merging common and technical lexicon wordnets* Raquel Amaro and Sara Mendes
15:45–15:50	*Building Multilingual Lexical Resources using Wordnets: Structure, Design and Implementation* Shikhar Kr. Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Himadri Bharali, Mayashree Mahanta and Utpal Saikia
15:50–15:55	*A New Semantic Lexicon and Similarity Measure in Bangla* Manjira Sinha, Abhik Jana, Tirthankar Dasgupta and Anupam Basu
15:55–16:00	*Where's the meeting that was cancelled? existential implications of transitive verbs* Patricia Amaral, Valeria de Paiva, Cleo Condoravdi and Annie Zaenen
16:30-17:00	Tea break
17:00–1800	**Session 5: Pragmatic Aspects**
17:00–17:30	*SEJFEK - a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units* Agata Savary, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek and Filip Makowiecki
17:30–18:00	*The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database* Ekaterina Manicheva, Maria Petrova, Elena Kozlova and Tatiana Popova
18:00–18:15	**Session 6: Wrap-up Discussion and Closing Address**
18:15	**End of the Workshop**