Coling 2008

# 22nd International Conference on Computational Linguistics

# Proceedings of the workshop on Cognitive Aspects of the Lexicon

Workshop chairs:
Michael ZOCK and Chu-Ren HUANG

24 August 2008
Manchester, UK

Order copies of this and other Coling proceedings from:

> Association for Computational Linguistics (ACL)
> 209 N. Eighth Street
> Stroudsburg, PA 18360
> USA
> Tel: +1-570-476-8006
> Fax: +1-570-476-0860
> `acl@aclweb.org`

*Design by Chimney Design, Brighton, UK*
*Production and manufacture by One Digital, Brighton, UK*

# Preface

Information access and exchange play a major role in our globalized world. Hence, building resources (lexica, thesauri, ontologies or annotated corpora) and providing access to words become an important goal. The lexicon is a vital resource for building applications. It is also a crucial element in the study of human language processing.

The spirit of this workshop multidisciplinary, the goal being to gather experts with various backgrounds and to allow them to exchange ideas, to compare their methodologies and theoretical perspectives, to create synergy, and to encourage future collaborations. In sum, the participants will be discussing questions concerning the **cognitive aspects** of the lexicon, and their answers should guide the design of on-line dictionaries.

While completeness is a virtue, the quality of a dictionary depends not only on coverage (number of entries) and granularity, 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 (word, concept, sound). Unlike *readers*, who look for meanings, *writers* start from them, searching for the 'right' 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 word) and via diverse access routes. Navigation takes place in a huge conceptual-lexical space, and the results are displayable in a multitude of forms (as trees, as lists, as graphs, or sorted alphabetically, by topic, by frequency).

Many lexicographers work nowadays with huge digital corpora, using language technology to build and to maintain the resource. But access to the potential wealth in dictionaries remains limited for the common user. Yet, the new possibilities of electronic media in terms of comfort, speed and flexibility (multiple inputs, polymorph outputs) are enormous and probably beyond our imagination. More than just allowing electronic versions of paper-bound dictionaries, computers provide a freedom for rethinking dictionaries, thesauri, encyclopedia, etc., a distinction necessary in the past for economical reasons, but not justified anymore.

The goal of this workshop is to perform the groundwork for the next generation of electronic dictionaries, that is, to study the possibility of integrating the different resources, as well as to explore the feasibility of taking the users' needs, knowledge and access strategies into account.

To reach this goal we have asked authors to address one or more of the following:

1. **Conceptual input of a dictionary user:** what is present in speaker's/writer's minds when they are generating a message and looking for a (target) word? Does the user have in mind conceptual primitives, semantically related words, some type of partial definition, something like synsets, or something completely different?

2. **Access, navigation and search strategies:** how can search be supported by taking into account prior, i.e. available knowledge? Entries should be accessible in many ways: by word forms, by meaning, by sounds (syllables), or in a combined form, and this even if input is given in an incomplete, imprecise or degraded form. The more precise the conceptual input, the less

navigation should be needed and vice versa. How can we create manageable search spaces, and provide a user with the tools for navigating within them?

3. **Indexing words and organizing the lexicon:** Words and concepts can be organized in many ways, varying according to typology and conceptual systems. For example, words are traditionally organized alphabetically in Western languages, but by semantic radicals and stroke counts in Chinese. The way words and concepts are organized affects indexing and access. Indexing must robustly allow for multiple ways of navigation and access. What efficient organizational principles allow the greatest flexibility for access? What about lexical entry standardization? Are universal definitions possible? What about efforts such as the Lexical Markup Framework (LMF) and other global structures for the lexicon? Can ontologies be combined with standards for the lexicon?

4. **NLP Applications:** Contributors can also address the issue of how such enhanced dictionaries, once embedded in existing NLP applications, can boost performance and help solve lexical and textual-entailment problems such as those evaluated in SEMEVAL 2007, or, more generally, generation problems encountered in the context of summarization, question-answering, interactive paraphrasing or translation.

We've received 18 papers, of which 6 were accepted as full papers, while 8 were chosen as poster presentations. While we did not get papers on all the issues mentioned in our call, we did get a quite rich panel on ideas as divers as use of ontologies; sense extraction; computation of associative responses to multi-word stimuli; saliency relations; lexical relationships within collocations and word association norms; cognitive organization of dictionaries; user-adapted views on a lexicographic database; access based on conceptual input; search in onomasiological dictionaries, access based on underspecified input; dictionary use for authoring aids or MT, use of feature vectors, corpora and machine learning, etc..

It was also interesting to see the variety of languages in which these issues are addressed. The proposals range from Japanese, English, German, Russian, Dutch, Bulgarian, Romanian, Spanish, to French and Chinese. 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 express here our sincerest thanks to all the specialists who have assisted us to assure a good selection of papers, despite the very tight schedule. Their reviews were helpful not only for us as decision makers, 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 Chu-Ren Huang

**Organizers:**

Michael Zock,LIF, CNRS, Marseille,(France)
Chu-Ren Huang, Sinica, (Taiwan)

**Programme Committee:**

Slaven Bilac, Google-Tokyo, (Japan)
Pierrette Bouillon, ISSCO, Geneva, (Switzerland)
Dan Cristea, University of Iasi, (Romania)
Christiane Fellbaum, Princeton, (USA)
Olivier Ferret, CEA LIST, (France)
Thierry Fontenelle, Microsoft, Redmont, (USA)
Gregory Grefenstette, CEA LIST, (France)
Graeme Hirst, University of Toronto, (Canada)
Ed Hovy, ISI, Los Angeles, (USA)
Chu-Ren Huang, Sinica, (Taiwan)
Terry Joyce, Tama University, Kanagawa-ken, (Japan)
Adam Kilgarriff, Brighton, Lexical Computing Ltd, (UK)
Philippe Langlais, University of Montreal, (Canada)
Dekang Lin, Google, Mountain View, California, (USA)
Rada Mihalcea, University of North Texas, (USA)
Alain Polguère, University of Montreal, (Canada)
Reinhard Rapp, University of Tarragona, (Spain)
Sabine Schulte im Walde, University of Stuttgart, (Germany)
Gilles Serasset, Imag, Grenoble, (France)
Anna Sinopalnikova, FIT, BUT, Brno, (Czech Republic)
Takenobu Tokunaga, Titech, Tokyo, (Japan)
Dan Tufis, RACAI, Bucharest, (Romania)
Jean Véronis, Université d'Aix-Marseille, (France)
Yorick Wilks, Oxford Internet Institute, (UK)
Michael Zock, LIF, CNRS, Marseille, (France)
Pierre Zweigenbaum, Limsi, Orsay, (France)

# Table of Contents

# Conference Programme

**Sunday, August 24, 2008**

9:00–9:10      Opening Remarks

**Session 1: Regular Talks**

9:10–9:50      *Comparing Lexical Relationships Observed within Japanese Collocation Data and Japanese Word Association Norms*
Terry Joyce and Irena Srdanović

10:50–10:30    *Lexical access based on underpecified input*
Michael Zock and Schwab Didier

10:30–11:00    Cofee Break + Poster Installation

11:00–11:40    *Accessing the ANW Dictionary*
Fons Moerdijk, Carole Tiberius and Jan Niestadt

**Session 2: Poster Presentations (8 minutes each)**

11:40–11:48    *ProPOSEL: a human-oriented prosody and PoS English lexicon for machine-learning and NLP*
Claire Brierley and Eric Atwell

11:48–11:56    *Natural Language Searching in Onomasiological Dictionaries*
Gerardo Sierra

11:56–12:04    *First ideas of user-adapted views of lexicographic data exemplified on OWID and elexiko*
Carolin Möller-Spitzer and Christine Möhrs

12:04–12:12    *Multilingual Conceptual Access to Lexicon based on Shared Orthography: An ontology-driven study of Chinese and Japanese*
Chu-Ren Huang, Ya-Min Chou, Chiyo Hotani, Sheng-Yi Chen and Wan-Ying Lin

12:12–12:20    *Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing*
Neculai Curteanu, Alex Moruz and Diana Trandabăţ

12:20–12:28    *Lexical-Functional Correspondences and Their Use in the System of Machine Translation ETAP-3*
Andreyeva Sasha

**Sunday, August 24, 2008 (continued)**

12:28–12:36   *The "Close-Distant" Relation of Adjectival Concepts Based on Self-Organizing Map*
Kyoko Kanzaki, Noriko Tomuro and Hitoshi Isahara

12:36–12:45   *Looking up phrase rephrasings via a pivot language*
Aurelien Max and Michael Zock

12:45–14:00   Lunch

**Session 3: Regular Talks**

14:00–14:40   *Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy*
Bruno Gaume, Karine Duvignau, Laurent Prévot and Yann Desalle

14:40–15:20   *Cognitively Salient Relations for Multilingual Lexicography*
Gerhard Kremer, Andrea Abel and Marco Baroni

15:20–15:50   Coffee Break + Poster Session

15:50–16:30   *The Computation of Associative Responses to Multiword Stimuli*
Reinhard Rapp

**Session 4: Poster Session + Wrap Up Discussion**

16:30–17:00   Poster Session

17:00–17:30   Wrap Up Dission

17:30–17:30   End of the Workshop