1st Call for Papers

4th Workshop on **Cognitive Aspects of the Lexicon** (CogALex)

together with a *shared task* concerning the 'lexical access-problem'

Pre-conference workshop at COLING 2014 (August 23d, Dublin, Ireland)

(!) *Submission deadline*: May 25, 2014 (!)

Invited speaker : *Roberto Navigli* (Sapienza University of Rome)

http://pageperso.lif.univ-mrs.fr/~michael.zock/ColingWorkshops/CogALex-4/CogALex-IV/cogalex-webpage/cfp.html
(page under construction)

GOAL

The aim of the workshop is to bring together researchers involved in the construction and application of electronic dictionaries to discuss modifications of existing resources in line with the users' needs, thereby fully exploiting the advantages of the digital form. Given the breadth of the questions, we welcome reports on work from many perspectives, including but not limited to: computational lexicography, psycholinguistics, cognitive psychology, language learning and ergonomics.

MOTIVATION

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, by now they have 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 dynamic networks, i.e. 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 this 4th edition of the CogALex workshop we therefore also invite scientists working in these fields, with the goal 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.

TOPICS OF INTEREST

This workshop is about possible enhancements of lexical resources and electronic dictionaries. To perform the groundwork for the next generation of such resources we invite researchers involved in the building of such tools. 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 solicit papers including but not limited to the following topics, each of 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.

1) **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)?

2) **The meaning of words**

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

3) **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 (reading between the lines, subliminal communication);
- 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)

4) **Methods for crafting dictionaries or indexes**

- Manual, automatic or collaborative building of dictionaries and indexes (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, ...)
- Use of corpora and patterns (data-mining) for getting access to words, their uses, combinations and associations

5) **Dictionary access** (navigation and search strategies, interface issues,...)

- Search based on sound, meaning or associations
- 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

6) **Dictionary applications**

- Methods supporting vocabulary learning (for example, creation of data-bases showing words in various contexts)
- Tools for supporting Human translation

IMPORTANT DATES

- Deadline for paper submissions: May 25, 2014
- Notification of acceptance: June 15, 2014
- Camera-ready papers due : July 7, 2014
- Worskhop date: August 23, 2014

SUBMISSION INFORMATION

Papers should follow the COLING main conference formatting details (http://www.coling-2014.org/call-for-papers.php) and should be submitted as a PDF-file via the START workshop manager at https://www.softconf.com/coling2014/WS-1/ (you must register first).

Contributions can be short or long papers. Short paper submission must describe original and unpublished work without exceeding six (6) pages (references included). Characteristics of short papers include: a small, focused contribution; work in progress; a negative result; a piece of opinion; an interesting application nugget. Long paper submissions must describe substantial, original, completed and unpublished work without exceeding twelve (12) pages (references included).

Reviewing will be double blind, so the papers should not reveal the authors' identity. Accepted papers will be published in the workshop proceedings.

For further details see: http://pageperso.lif.univ-mrs.fr/~michael.zock/ColingWorkshops/CogALex-4/CogALex-IV/cogalex-webpage/index.html

SHARED TASK

We invite participation in a shared task devoted to the problem of lexical access in language production, with the aim of providing a quantitative comparison between different systems.

Motivation of shared task

The quality of a dictionary depends not only on coverage, but also on the accessibility of the information. That is, a crucial point is dictionary access. Access strategies vary with the task (text understanding vs. text production) and the knowledge available at the very 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).

To bring some structure into this multitude of possibilities, the shared task will concentrate on a crucial subtask, namely multiword association. What we mean by this in the context of this workshop is the following. Suppose, we were looking for a word expressing the following ideas: ‘superior dark coffee made of beans from Arabia’, but could not remember the intended word ‘mocha’. Since people always remember something concerning the elusive word, it would be nice to have a system accepting this kind of input, to propose then a number of candidates for the target word. Given the above example, we might enter ‘dark’, ‘coffee, beans, and Arabia, and the system would be supposed to come up with lists of associated words such as mocha, espresso, or cappuccino.

Procedure

The participants will receive lists of five given words (primes) such as 'circus', 'funny', 'nose', 'fool', and 'fun' and are supposed to compute the word which is most closely associated to all of them. In this case, the word 'clown' would be the expected answer. Here are some more examples:

given words: gin, drink, scotch, bottle, soda
expected answer: whisky

given words: wheel, driver, bus, drive, lorry
expected answer: car

given words: neck, animal, zoo, long, tall
expected answer: giraffe

given words: holiday, work, sun, summer, abroad
expected answer: vacation

given words: home, garden, door, boat, chimney
expected answer: house

given words: blue, cloud, stars, night, high
expected answer: sky

We will provide a training set of 2000 sets of five input words (multi
word stimuli), together with the expected target words (associative response). The participants will have several weeks to train their systems on this data. After the training phase, we will release a test set containing another 2000 sets of five input words, but without providing the expected target words.

Participants will have five days to run their systems on the test data, thereby predicting the target words. For each system, we will compare the results to the expected target words and compute an accuracy. The participants will be invited to submit a paper describing their approach and the results.

For the participating systems, we will distinguish two categories: (1) *Unrestricted systems*. They can use any kind of data to compute their results. (2) *Restricted systems*: These systems are only allowed to draw on the freely available ukWaC corpus (comprising 2 billion words) in order to extract information on word associations. Participants are allowed to compete in either category or in both.

Schedule for Shared Task

- Training Data Release: March 25, 2014

- Test Data Release: May 5, 2014
- Final Results: May 9, 2014
- Deadline for Paper Submission : May 25, 2014
- Reviewers' feedback: June, 15, 2014
- Camera-Ready Version : July 7, 2014
- Workshop date: August 23, 2014

All data releases can be found on the workshop website.

PROGRAMME COMMITTEE

- Bel Enguix, Gemma (LIF-CNRS, France)
- Chang, Jason (National Tsing Hua University, Taiwan)
- Cook, Paul (University of Melbourne, Australia)
- Cristea, Dan (University A.I.Cuza, Iasi, Romania)
- De Deyne, Simon (Experimental Psychology, Leuven, Belgium) and (Adelaide, Australia)
- De Melo, Gerard (IIIS, Tsinghua University, Beijing, China)
- Ferret, Olivier (CEA LIST, Gif sur Yvette, France)
- Fontenelle, Thierry (CDT, Luxemburg)
- Gala, Nuria (LIF-CNRS, Aix Marseille University, Marseille, France)
- Granger, Sylviane (Université Catholique de Louvain, Belgium)
- Grefenstette, Gregory (Inria, Saclay, France)
- Hirst, Graeme (University of Toronto, Canada)
- Hovy, Eduard (CMU, Pittsburgh, USA)
- Hsieh, Shu-Kai (National Taiwan University, Taipei, Taiwan)
- Huang, Chu-Ren (Hongkong Polytechnic University, China)
- Joyce, Terry (Tama University, Kanagawa-ken, Japan)
- Lapalme, Guy (RALI, University of Montreal, Canada)
- Lenci, Alessandro (CNR, university of Pisa, Italy)
- L'Homme, Marie Claude (University of Montreal, Canada)
- Mihalcea, Rada (University of Texas, USA)
- Navigli, Roberto (Sapienza, University of Rome, Italy)
- Pirrelli, Vito (ILC, Pisa, Italy)
- Polguère, Alain (ATILF-CNRS, Nancy, France)
- Rapp, Reinhard (LIF-CNRS, France) and (Mainz, Germany)
- Rosso, Paolo (NLEL, Universitat Politècnica de València, Spain)
- Schwab, Didier (LIG-GETALP, Grenoble, France)
- Serasset, Gilles (IMAG, Grenoble, France)
- Sharoff, Serge (University of Leeds, UK)
- Su, Jun-Ming (University of Tainan, Taiwan)
- Tiberius, Carole (Institute for Dutch Lexicology, The Netherlands)
- Tokunaga, Takenobu (TITECH, Tokyo, Japan)
- Tufis, Dan (RACAI, Bucharest, Romania)
- Valitutti, Alessandro (Helsinki Institute of Information Technology, Finland)
- Wandmacher, Tonio (IRT SystemX, Saclay, France)
- Zock, Michael (LIF-CNRS, Marseille, France), currently (University of Tainan, Taiwan)

WORKSHOP ORGANIZERS and **CONTACT PERSONS**

- Michael Zock (LIF-CNRS, Marseille, France), michael.zock AT lif.univ-mrs.fr
- Reinhard Rapp (University of Aix Marseille (France) and Mainz (Germany), reinhardrapp AT gmx.de
- Chu-Ren Huang (The Hong Kong Polytechnic University, Hong Kong), churen.huang AT inet.polyu.edu.hk

For more details see :

http://pageperso.lif.univ-mrs.fr/~michael.zock/ColingWorkshops/CogALex-4/CogALex-IV/cogalex-webpage/index.html