The Workshop on the Cognitive Aspects of the Lexicon

Proceedings of the Workshop

December 12, 2020
Barcelona, Spain (Online)
Copyright of each paper stays with the respective authors (or their employers).

Introduction

Supporting us in many tasks (thinking, searching, memorizing, categorizing, communicating) words are a key aspect of natural language and human cognition. Yet they are large in number, and complex in form, hence the question, how are they learned, accessed and used? These are typical questions addressed in this kind of workshop that looks at words from a cognitive perspective.

More specifically we are interested in the creation, use and enhancements of lexical resources (representation, organization of the data, etc.). What are their limitations, and how can they be overcome? What are the users’ and the engineers’ needs (computational aspects)?

Given these goals, the CogALex workshop series, which this year has reached the sixth edition, has become a venue to discuss these issues from multiple viewpoints: linguistics (lexicography, computational- or corpus linguistics), neuro- or psycholinguistics (tip-of-the-tongue problem, word associations), network-related sciences (vector-based approaches, graph theory, small-world problem).

Also, just like in previous workshops (CogALex IV and V) we proposed a ‘shared task’. This year the goal was to provide a common benchmark for testing lexical representations for the automatic identification of lexical semantic relations (synonymy, antonymy, hypernymy) in various languages (English, Chinese, etc.). Discovering whether words are semantically related or not – and, if so which kind of relation holds between them – is an important task both in language (writing, production of coherent discourse) and cognition (thinking, categorization). Semantic relations also play a central role in the organization of words in the lexicon and their subsequent retrieval (word access, navigation).

We received 25 submissions of which we accepted 7 for oral presentation, 6 as posters, and 3 as shared task papers (which means, we had overall an acceptance rate of about 65%, and 32% for oral presentations). We were pleasantly surprised to see that a growing number of authors combine work done in linguistics and psychology. Actually, the most highly rated paper came from two teams favorably complementing each other in this respect, which shows that it is possible to work together, and to produce excellent results which may benefit not only the respective teams but also the rest of the community. We hope that such results encourage other researchers or research teams to build on the strengths of other disciplines.

Obviously, none of the results here presented would be possible without the dedication of the authors and the efforts of the reviewers, who have considerably contributed to helping the authors to improve their work. We would like to thank both of them.

The CogALex Organizing Committee
Michael Zock
Emmanuele Chersoni
Alessandro Lenci
Enrico Santus
Organizers:

Zock, M. (CNRS, LIS-AMU, Marseille, France)
Chersoni, E. (The Hong Kong Polytechnic University, China)
Lenci, A. (University of Pisa, Italy)
Santus, E. (Bayer Pharmaceuticals, U.S.A)

Program Committee:

Arenas, A. (Universidad Rovira i Virgili, Tarragona, Spain)
Barbu Mititelu V. (RACAI, Bucharest, Romania)
Biemann, C. (Language Technology group, Universität Hamburg, Germany)
Brysbaert, M. (Experimental Psychology, Ghent University, Belgium)
Camacho-Collados, J. (Cardiff University, Cardiff, UK)
Carl, M. (Kent State University, Cleveland, Ohio, U.S.A.)
Chersoni, E. (The Hong Kong Polytechnic University, China)
Chiarcos, C. (Goethe-Universität Frankfurt am Main, Germany)
Cristea, D. (Al. I. Cuza University, Iasi, Romania)
de Melo, G. (Hasso Plattner Institute, University of Potsdam, Germany)
deDeyne, S. (School of Psychological Sciences, University of Melbourne, Australia)
Delmonte, R. (Department of Computer Science, Università Ca’ Foscari, Italy)
Erik, K. (University of Texas, U.S.A.)
Espinosa-Anke, L. (COMSC, Cardiff University, UK)
Evert, S. (University of Erlangen, Germany)
Fairon, C. (Catholic University of Louvain, Belgium)
Ferret, O. (CEA LIST, France)
François, T. (Catholic University of Louvain, Belgium)
Fellbaum, C. (Princeton University, U.S.A.)
Gala, N. (Laboratoire Parole et Langage, Aix Marseille University)
Hirst, G. (University of Toronto, Canada)
Hovy, E. (CMU, Pittsburgh, U.S.A.)
Hsieh, S.K. (National Taiwan University, Taipei, Taiwan)
Iacobacci, I. (Huawei Noah’s Ark Lab, London, UK)
Kenett, Y. (University of Pennsylvania, Philadelphia, PA, U.S.A.)
Langlais, F. (DIRO, University of Montreal, Canada)
L’Homme M.C. (University of Montreal, Canada)
Lafourcade, Matthieu (LIRMM, University of Montpellier, France)
Lenci, A. (University of Pisa, Italy)
McCrae, J. (Data Science Institute, NUI Galway, Ireland)
Padó, S. (University of Stuttgart, Germany)
Philevar, T. (Iran University of Science and Technology, Teheran, Iran)
Pirrelli, V (ILC, Pisa, Italy)
Polguère, A. (ATILF, CNRS and University of Lorraine, Nancy, France)
Purver, M. (King’s College, London, UK)
Raganato, A. (University of Helsinki, Helsinki, Finland)
Saggion, H. (Universitat Pompeu Fabra, Spain)
Sahlgren, M. (SICS, Smart AI, Sweden)
Santus, E. (Bayer Pharmaceuticals, U.S.A.)
Schulte im Walde, S. (University of Stuttgart, Germany)
Schwab, D. (LIG, Grenoble, France)
Stella, M. (Complex Science Consulting, Lecce, Italy)
Tiberius, C. (Institute for Dutch Lexicology, Leiden, The Netherlands)
Tufis, D. (RACAI, Bucharest, Romania)
Zock, M. (CNRS, LIS-AMU, Marseille, France)

Shared Task Organizers:
Chersoni, E. (The Hong Kong Polytechnic University, China)
Iacoponi, L. (Amazon)
Santus, E. (Bayer Pharmaceuticals, U.S.A.)
Xiang, R. (The Hong Kong Polytechnic University, China)
Best Paper Award:
Markus J. Hofmann, Lara Müller, Andre Rölke, Ralph Radach and Chris Biemann.
Individual corpora predict fast memory retrieval during reading

Keynote Talk

Vito Pirrelli (ILC, Pisa, Italy)
Emerging words in a vanishing lexicon: prospects of interdisciplinary convergence

In the wake of the ‘cognitive revolution’ (Miller 2003), scholars believed in a parsimonious form of direct correspondence between grammar rules, their organization and processing principles on one hand, and psychological and even neurological processes on the other. Grammatical rules and syntactic structures were claimed to have psycholinguistic reality, i.e. to be mentally represented, and speakers were believed to use the representations offered by linguists (e.g., Clahsen 2006, Jackendoff 1997, Levelt et al. 1999, Marslen-Wilson and Tyler 1998, Miller and Chomsky 1963). This straightforward assumption has been shared and popularized by Steven Pinker (Pinker 1999, Prasada and Pinker, 1993).

According to the Declarative/Procedural model (Pinker and Ullman 2002), the distinction drawn by linguists between regular and irregular morphology is not just a matter of classificatory convenience, but is motivated by the way linguistic information is processed in the human brain. Different cortical and subcortical areas are recruited to process and retrieve word forms like walked and sang. The former is decomposed into its sublexical constituents (walk and –ed), while the latter is stored as a whole. Likewise, the classical organization of the language architecture into two sharply compartmentalized modules, namely a word-based lexicon and a rule-based grammar, was viewed as reflecting the neuro-functional divide between a long-term store of static units (or knowledge of “what”) and a procedural system (or knowledge of “how”) (Ullman 2002). Accordingly, the irreducible building blocks of language competence are stored in and accessed from a redundancy-free mental lexicon. Rules, in turn, are responsible for the assembly and disassembly of these blocks when complex structures (e.g. morphologically complex words, phrases or sentences) are processed in production or recognition. Despite this seemingly clear division of labour, researchers disagreed considerably on matters of detail. Some of them assume that full words are the minimal building blocks in the mental lexicon (Butterworth 1983, Manelis and Tharp 1977), others claim that only sublexical units are stored (Taft 1979, 2004, Taft and Forster 1975), yet others propose a combination of the two (Baayen 1992, Frauenfelder and Schreuder 1992, Caramazza, Laudanna and Romani 1988, Laudanna and Burani 1985). Nonetheless, the largely dominant view was that storage and computation are distinct processes, subserved by different brain areas, in line with what Harald Baayen (2007) humorously dubbed the ‘pocket calculator metaphor’.

This general picture was challenged by the ‘connectionist revolution’ (Medler 1998, Pirrelli et al. 2020). Multi-layered perceptrons proved to be able to process both regularly and irregularly inflected words with a unique underlying mechanism (Rumelhart and McClelland 1986). Accordingly, morphological structure was not modelled as an all-or-nothing issue. Rather, it was an emergent property of the dynamic self-organization of subsymbolic, distributed representations, contingent on the processing history of input forms. In fact, artificial neural networks appear to mark an even more radical departure from traditional language processing architectures. First, in neural networks, lexical representations are not given but learned. Thus, aspects of how they are acquired and eventually represented are not taken for granted, but lie at the core of connectionist modelling. Secondly, and most importantly, lexical representations and processing routines are not assigned to different components, but they both rest on the same level of weighted connections.
On the one hand, storage implies processing. Network nodes that have been repeatedly activated in processing an input word are the same units representing this word in long-term memory. On the other hand, processing implies storage. The online processing of an input word consists in the short-term re-activation of processing routines that were successfully triggered by the same word in the past. Ultimately, processing and storage only designate two different points in time (i.e. immediate response to stimulus, and response consolidation) of the same underlying learning dynamics.

McClelland and Rumelhart (1986) published their pioneering book on Parallel Distributed Processing 35 years ago. Strange as it may seem, the theoretical consequences of this radical shift of paradigm have not been fully appreciated. In my talk, I will try to reappraise their contribution by showing how it can deal with issues like word representation and processing. In doing so, I will consider evidence from several interrelated lines of research. The first one revolves around evidence from human word processing and lexical acquisition, suggesting that a lot of lexical information is inextricably related to processing. The second line of research focuses on recent advances in the neurobiology of human memory and its tight connection with language processing. Last but not least, I will consider neurobiologically inspired computer models of the language architecture.

Drawing on an analogy with recent developments in discriminative learning and morphological theory (Baayen et al. 2011; Blevins 2016, Ramscar and Gitcho 2007, Ramscar and Yarlett 2007), I will suggest that speakers’ knowledge about words is the resulting state of a dynamic, self-organizing process. According to this view, words are abstractions emerging from interrelated patterns of sensory experience, communicative and social interaction and psychological and neurobiological mechanisms (Elman 2009). The information associated with them is hardly ever stable, time-independent or context-independent. Their content is continuously updated and reshaped as a function of the moment (when), the reason (why) and the frequency (how often) of its access and processing. Such flowing activation state is reminiscent of the wave/particle duality in quantum physics (Libben 2016) or the inherently adaptive, self-organising behaviour of biological dynamic systems (Beckner et al. 2009). Hence, if we look for particles (i.e. individual word representations) we may not be able to find them. It all depends on the task or the context.

This discussion leads to an apparently paradoxical state of affairs. We know that the linguists’ view of the lexicon as a redundancy-free container of word representations fails to capture our current understanding of how words are accessed, used and processed. Can then we talk about the reality of words in any non-metaphorical or non-epiphenomenal sense? I believe that we can. After all, speakers have lexical knowledge. However, the mental reality of words resides in the collective behaviour of nodes and connections in a lexical network engaged in a particular word processing task. It cannot be deduced from any individual and stable representation. As the context or task changes, also the pattern of these connections will change, and so our perception of individual representations.

Language sciences address different Marr’s (1982) levels of understanding of a complex cognitive system. Theoretical linguistics addresses Marr’s level one (i.e. what speakers do when they use language), while psycholinguistics is mostly concerned with level two (i.e. how speakers use language), and neurolinguistics with level three (i.e. where in the brain language processes take place). Due to the dominant focus of theoretical linguistics on the basic units of language and the principles of their combination, linguists have laid nearly exclusive emphasis on representation issues. Conversely, psycholinguists and neurolinguists have mainly been concerned with behavioural and physiological evidence of the human processor. As suggested by Anderson (1972), inter-level mapping rarely implies the extrapolation to level Y of properties holding at level X. It is thus not surprising that moving from Marr’s level one to another level is often a matter of discovering new laws and concepts, and requires a creative shift from quantitative to qualitative differentiation. Decade-long developments in recurrent neural networks have proved instrumental
in shedding light on the psychological nature of classical linguistic categories and basic units (Elman 2009, Marzi and Pirrelli 2015). Ultimately, they appear to lend support to Poggio’s (2010) claim that (language) learning is key to the appropriate methodological unification of Marr’s epistemological levels: units in language crucially depend on the way they are acquired, organized and used by the speakers. From this perspective, any attempt to put all these units directly into the speaker’s mind is dubious, if not futile.

References


# Table of Contents

**Individual corpora predict fast memory retrieval during reading**
Markus J. Hofmann, Lara Müller, Andre Rölke, Ralph Radach and Chris Biemann .......................... 1

**Investigating Rich Feature Sources for Conceptual Representation Encoding**
Lu Cao, Yulong Chen, Dandan Huang and Yue Zhang ................................................................. 12

**General patterns and language variation: Word frequencies across English, German, and Chinese**
Annika Tjuka .................................................................................................................................... 23

**Less is Better: A cognitively inspired unsupervised model for language segmentation**
Jinbiao Yang, Stefan L. Frank and Antal van den Bosch ............................................................... 33

**The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations**
Rong Xiang, Emmanuele Chersoni, Luca Iacoboni and Enrico Santus ........................................... 46

**Extracting meaning by idiomaticity: Description of the HSemID system at CogALex VI (2020)**
Jean-Pierre Colson .......................................................................................................................... 54

**CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification**
Lennart Wachowiak, Christian Lang, Barbara Heinisch and Dagmar Gromann .............................. 59

**CogALex-VI Shared Task: Bidirectional Transformer based Identification of Semantic Relations**
Saurav Karmakar and John P. McCrae ............................................................................................. 65

**Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions**
REYHANEH HASHEMPOUR and Aline Villavicencio ................................................................. 72

**Definition Extraction Feature Analysis: From Canonical to Naturally-Occurring Definitions**
Mireia Roig Mirapeix, Luis Espinosa Anke and Jose Camacho-Collados ................................. 81

**Speech Disfluencies occur at Higher Perplexities**
Priyanka Sen ...................................................................................................................................... 92

**Bilingual Lexical Access and Cognate Idiom Comprehension**
Eve Fleisig ......................................................................................................................................... 98

**Schwa-deletion in German noun-noun compounds**
Tom S Juzek and Jana Haeussler ................................................................................................. 107

**Translating Collocations: The Need for Task-driven Word Associations**
Oï Yee Kwong ................................................................................................................................. 112

**Characterizing Dynamic Word Meaning Representations in the Brain**
Nora Aguirre-Celis and Risto Miikkulainen ................................................................................. 117

**Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge**
Sathvik Nair, Mahesh Srinivasan and Stephan Meylan ................................................................. 129

**Automatic Word Association Norms (AWAN)**
Jorge Reyes-Magaña, Gerardo Sierra Martínez, Gemma Bel-Enguix and Helena Gomez-Adorno 142
Conference Program

14:00–14:10  Introduction

14:10–15:10  Keynote Talk

Emerging Words in a Vanishing Lexicon: Prospects of Interdisciplinary Convergence
Vito Pirrelli

15:10–15:30  Break

15:30–16:30  Session 1: Oral Presentations

Individual corpora predict fast memory retrieval during reading
Markus J. Hofmann, Lara Müller, Andre Rölke, Ralph Radach and Chris Biemann

Investigating Rich Feature Sources for Conceptual Representation Encoding
Lu Cao, Yulong Chen, Dandan Huang and Yue Zhang

General patterns and language variation: Word frequencies across English, German, and Chinese
Annika Tjuka

Less is Better: A cognitively inspired unsupervised model for language segmentation
Jinbiao Yang, Stefan L. Frank and Antal van den Bosch
16:30–16:50  Poster Booster and Shared Task Introduction

*The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations*
Rong Xiang, Emmanuele Chersoni, Luca Iacoponi and Enrico Santus

16:50–17:50  Poster Session

*Extracting meaning by idiomaticity: Description of the HSemID system at CogALex VI (2020)*
Jean-Pierre Colson

*CogALex-VI Shared Task: Transrelation - A Robust Multilingual Language Model for Multilingual Relation Identification*
Lennart Wachowiak, Christian Lang, Barbara Heinisch and Dagmar Gromann

*CogALex-VI Shared Task: Bidirectional Transformer based Identification of Semantic Relations*
Saurav Karmakar and John P. McCrae

*Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions*
REYHANEH HASHEMPOUR and Aline Villavicencio

*Definition Extraction Feature Analysis: From Canonical to Naturally-Occurring Definitions*
Mireia Roig Mirapeix, Luis Espinosa Anke and Jose Camacho-Collados

*Speech Disfluencies occur at Higher Perplexities*
Priyanka Sen

*Bilingual Lexical Access and Cognate Idiom Comprehension*
Eve Fleisig

*Schwa-deletion in German noun-noun compounds*
Tom S Juzek and Jana Haeussler

*Translating Collocations: The Need for Task-driven Word Associations*
Oi Yee Kwong
No Day Set (continued)

17:50–18:00  Break

18:00–18:45  Session 2: Oral Presentations

Characterizing Dynamic Word Meaning Representations in the Brain
Nora Aguirre-Celis and Risto Miikkulainen

Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge
Sathvik Nair, Mahesh Srinivasan and Stephan Meylan

Automatic Word Association Norms (AWAN)
Jorge Reyes-Magaña, Gerardo Sierra Martínez, Gemma Bel-Enguix and Helena Gomez-Adorno

18:45–19:00  Closing Remarks