Introduction to the Special Issue

Verena Henrich and Erhard Hinrichs (Guest Editors)

University of Tübingen, Germany
{vhenrich,eh}@sfs.uni-tuebingen.de

This special issue of the Journal of Cognitive Science (JCS) is on the topic of *Computational, Cognitive, and Linguistic Approaches to the Analysis of Compounds and Collocations.*

The papers contained in this special issue share as a common theme the semantic relations that hold between the constituent members of compounds and collocations. The contributions by Bell and by Sorokin et al. focus on the semantics of compounds. Taking Fanselow's distinction between basic and stereotypical relations (Fanselow, 1981) as a starting point, Bell investigates to what extent this distinction can be predicted in a logistic regression model by the degree of productivity of the head and modifier constituent of a compound, by the degree to which a head or modifier denotes a concrete versus an abstract entity, and by the degree of lexical ambiguity of a head or a modifier. Her findings suggest that the type of compound-internal semantic relation cannot be predicted by the semantics of the modifier and head constituents alone, but also involves distributional properties of these constituents. Moreover, the semantic relations as such seem to involve more general ontological categories such as material and location and, thus, seem to generalize across the semantics of individual lexical items.

Bell takes Fanselow's binary distinction of basic and stereotypical relations as a starting point for classifying compound-internal relations.

Journal of Cognitive Science 16-3: 195-199, 2015
Date submitted: 8/30/15 Date reviewed: 9/17/15
Date confirmed for publication: 9/18/15

By contrast, Sorokin et al. propose a hybrid annotation scheme that characterizes these compound-internal semantic relations by prepositional paraphrases and semantic properties. The authors motivate this hybrid approach by looking at compounds from a multilingual perspective. They observe that while prepositions are essential for translating compounds e.g. from Germanic to Romance languages, they are often too coarse-grained and ambiguous for the task. Thus, the motivation for including semantic properties in addition to the prepositions is to further enhance the annotation scheme, since semantic properties offer a more fine-grained semantic resolution compared to prepositions. Further support for the proposed hybrid annotation scheme is offered by a series of machine learning experiments which show that the automatic classification in the multi-label setup clearly outperforms the single-label classification (i.e. predicting the correct preposition or the correct semantic property in isolation).

While compound-internal semantic relations have been extensively studied in theoretical linguistics, computational linguistics, and cognitive psychology, semantic relations of collocations have by comparison received considerably less attention. In their contribution to this special issue, Lothar Lemnitzer and Alexander Geyken discuss the encoding and the semantic grouping of collocations in a semasiological German dictionary (Klein and Geyken, 2010; Geyken, 2013) and address two related research questions: Can the collocates of a given headword be grouped into cohesive lexical-semantic classes? Do semantically related headwords share a significant number of collocates and, if so, does the sharing of collocates extend to hyponyms of the headwords under consideration? In order to answer these questions, Lemnitzer and Geyken make use of a word profile generator (Geyken et al., 2009; Didakowski and Geyken, 2013) that automatically extracts for a given headword a list of statistically relevant word co-occurrences. Lemnitzer and Geyken show that Lexical Functions (in the sense of Mel'čuk's Meaning Text Theory (Mel'čuk, 1995; Mel'čuk 2012)) can then be used for systematically grouping the collocations obtained by the word profile generator.

In order to answer their second research question, i.e., do semantically related headwords have a significant number of collocates in common, Lemnitzer and Geyken make use of the German wordnet GermaNet (Hamp

and Feldweg, 1997; Henrich and Hinrichs, 2010) in order to be able to reliably identify co-hyponyms and hyponyms of a given headword. They show empirically that the intersection of collocates identified by the word profile generator for such sets of headwords is indeed non-empty and contains a significant number of shared collocates that can be grouped by Mel'čuk's Lexical Functions.

While the three contributions described above focus on compounds and collocations only, the study by Osenova and Simov consider the syntax and semantics of multi-word expressions more generally. Their goal is to link multi-word lexicon entries with attested corpus instances. They use a Bulgarian valency lexicon (Osenova et al., 2012) and a syntactically annotated Bulgarian treebank (Simov et al., 2004) for modeling this interaction. They invoke the notion of catenae (O'Grady, 1998; Gross, 2010) to identify the pieces of dependency structure in a syntactic tree as well as the corresponding information present in the valency lexicon. The modeling potential of catenae is exemplified for a range of syntactic and semantic phenomena in Bulgarian. These examples show that the notion of catenae is able to cope with different kinds of multi-word expressions, including idioms, compounds, and light-verb constructions. Moreover, the encoding scheme is able to distinguish between literal and non-compositional interpretation of idioms and to account for the range of potential syntactic modifiers of an idioms. Osenova and Simov apply their annotation scheme to Bulgarian. However, they argue that their approach is language independent and could also be applied to other languages.

Acknowledgments

We would like to thank the reviewers to contribute their expertise and the authors to submit and revise their papers to this special issue.

Support for both guest editors of this special issue was provided as part of the DFG grant to the Collaborative Research Center Emergence of Meaning (SFB 833). We are grateful to the editors of the Journal of Cognitive Science for giving us the opportunity for compiling this special issue.

References

Didakowski, J., & Geyken, A. 2013. From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network "Internet Lexicography"*, 43-52. Mannheim: Institut für Deutsche Sprache. (OPAL - Online publizierte Arbeiten zur Linguistik X/2012).

Fanselow, G. 1981. Zur Syntax und Semantik der Nominalkomposition. Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen, *Linguistische Arbeiten* 107. Tübingen: Niemeyer.

Geyken, A., Didakowski, J., & Siebert, A. 2009. Generation of word profiles for large German corpora. In. Y. Kawaguchi et al. (eds.), *Corpus Analysis and Variation in Linguistics,* 141-157. Tokyo University of Foreign Studies, Studies in Linguistics 1, John Benjamins Publishing Company.

Geyken, A. 2013. Large-Scale Documentary Dictionaries on the Internet. In: Gouws, R. et al. (eds.). *An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* de Gruyter, 1053-1069.

Gross, T. 2010. Chains in syntax and morphology. In Otoguro, Ishikawa, Umemoto, Yoshimoto, and Harada (editors), *PACLIC,* 143–152. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Hamp, B. & Feldweg H. 1997. "GermaNet - a Lexical-Semantic Net for German". In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid, 1997.

Henrich, V. & Hinrichs, E. 2010. GernEdiT – The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC 2010), Valletta, Malta, pp. 2228-2235.

Klein, W., & Geyken, A. 2010. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In Heid, U. et al. (eds.). *Lexikographica.* Berlin/New York, 79-93.

Mel'čuk, I. 1995. Phrasemes in language and phraseology in linguistics. In Everaert, M.; van der Linden, E. Schenk, A. et al. (eds.): *Idioms: Structural and psychological perspectives* 167–232. Lawrence Erlbaum Associates, Hillsdale.

Mel'čuk, I. 2012. Phraseology in the language, in the dictionary, and in the computer. In *Yearbook of Phraseology* 3(2012)1, 31-56.

O'Grady, W. 1998. The syntax of idioms. *Natural Language and Linguistic Theory,* 16:279–312.

Osenova, P., Simov, K., Laskova, L., & Kancheva, S. 2012. A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In: Nicoletta Calzolari

(Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12), Istanbul, Turkey. ELRA, 2636–2640.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. 2001. Multiword expressions: A pain in the neck for nlp. In *Proc. of the CICLing*, 2002, 1–15.

Simov, K., Osenova, P., Simov, A., & Kouylekov, M. 2004. Design and implementation of the Bulgarian HPSG-based treebank. *Journal of Research on Language and Computation,* 495–522, Kluwer Academic Publishers.