# A Way with Words: Recent Advances in Lexical Theory and Analysis: A Festschrift for Patrick Hanks

**Gilles-Maurice de Schryver (editor)**
(Ghent University and University of the Western Cape)

Kampala: Menha Publishers, 2010, vii+375 pp; ISBN 978-9970-10-101-6, €59.95

*Reviewed by*
*Paul Cook*
*University of Toronto*



In his introduction to this collection of articles dedicated to Patrick Hanks, de Schryver presents a quote from Atkins referring to Hanks as "the ideal lexicographer's lexicographer." Indeed, Hanks has had a formidable career in lexicography, including playing an editorial role in the production of four major English dictionaries. But Hanks's achievements reach far beyond lexicography; in particular, Hanks has made numerous contributions to computational linguistics. Hanks is co-author of the tremendously influential paper "Word association norms, mutual information, and lexicography" (Church and Hanks 1989) and maintains close ties to our field. Furthermore, Hanks has advanced the understanding of the relationship between words and their meanings in text through his theory of norms and exploitations (Hanks forthcoming).

The range of Hanks's interests is reflected in the authors and topics of the articles in this Festschrift; this review concentrates more on those articles that are likely to be of most interest to computational linguists, and does not discuss some articles that appeal primarily to a lexicographical audience. Following the introduction to Hanks and his career by de Schryver, the collection is split into three parts: Theoretical Aspects and Background, Computing Lexical Relations, and Lexical Analysis and Dictionary Writing.

## 1. Part I: Theoretical Aspects and Background

Part I begins with an unfinished article by the late John Sinclair, in which he begins to put forward an argument that multi-word units of meaning should be given the same status in dictionaries as ordinary headwords.

In Chapter 3 Wilks presents a previously published paper (Wilks 1977) in which he discusses how large lexical entries—much larger than typical dictionary entries—could be used to assign interpretations to preference-violating usages.

Pustejovsky and Rumshisky consider extended senses of verbs within the framework of the generative lexicon (Pustejovsky 2006) in Chapter 4. They argue that some extended senses can in fact be viewed as non-metaphorical usages, and offer a further classification of metaphorical usages into strong and weak metaphors depending on whether the core meaning of the predicate is generalized.

## 2. Part II: Computing Lexical Relations

Church kicks off Part II by responding to an earlier position piece by Kilgarriff (2007). Church argues that we should not abandon the use of large, but noisy and unbalanced, corpora, and discusses tasks to which such corpora are better suited than cleaner and more balanced, but smaller, corpora.

In Chapter 8 Grefenstette builds on two ideas previously seen in this collection—the use of massive amounts of data, and the importance of multiword expressions—to estimate the number of concepts. Grefenstette uses the number of frequently occurring noun–noun and adjective–noun sequences on the Web to arrive at an estimate of roughly two hundred million concepts. He acknowledges some of the limitations of his estimate, but nevertheless, this estimate gives insight into the potential number of entries in future lexical resources.

Patrick Hanks has developed corpus pattern analysis, a manual technique for identifying the typical patterns in which a verb is used.[1] Patterns are often described in terms of the classes (i.e., coarse semantic categories) of nouns occurring with a verb. In Chapter 9 Guthrie and Guthrie present an unsupervised statistical method for finding adjectives that are predictive of these noun classes, and present experimental results for the task of automatically determining an ambiguous noun's class from only its modifying adjective.

In Chapter 10 Geyken considers the impact of corpus size on the ability of pointwise mutual information to identify German verb-nominalization constructions. Geyken finds that when using a one-billion-word opportunistic corpus, most verb-nominalizations in a German dictionary are found to have positive pointwise mutual information, but that this is not the case when using a one-hundred-million word balanced corpus.

Word Sketches (Kilgarriff and Tugwell 2002) are automatically derived statistical summaries of the grammatical and collocational behavior of words that have proven to be a useful lexicographic tool. In Chapter 11 Pala and Rychlý examine some of the errors found in a word sketch for the Czech verb *vidět* ('see'), and conclude that the quality of the word sketch is relatively low, but could be improved through, primarily, better part-of-speech tagging and lemmatization.

We return to corpus pattern analysis in Chapter 12. Cinková et al. conduct a study to determine whether humans can reliably reproduce corpus pattern analysis, and further examine the relationship between nouns and semantic types in the existing Pattern Dictionary of English Verbs.[2]

In Chapter 13 Jezek and Frontini discuss an extension of corpus pattern analysis to produce a pattern bank—a resource of corpus instances annotated more richly than in corpus pattern analysis that could potentially benefit many natural language processing tasks—for Italian.

## 3. Part III: Lexical Analysis and Dictionary Writing

In Chapter 15 Atkins presents DANTE (a new English lexical database manually constructed by analyzing a large corpus of English) and contrasts it with FrameNet. Atkins discusses the possibility of (semi-automatically) linking the lexical units in DANTE and FrameNet. This is an interesting research problem, and moreover, the results would be a very rich lexical resource which would have many potential applications in computational linguistics.

In Chapter 16 Kilgarriff and Rychlý describe a system for semi-automatically deriving a draft of a dictionary entry, in particular, determining a word's senses. Beginning with an automatically produced clustering of a word's collocates, their method uses

1 http://nlp.fi.muni.cz/projects/cpa/.
2 http://nlp.fi.muni.cz/projects/cpa/.

an iterative process in which a lexicographer first provides sense annotations. These annotations are then used as training data for a word sense disambiguation system which is in turn applied to unannotated items. The lexicographer can then provide more annotations and the process continues. Although this could potentially be a very powerful tool for dictionary writing, the authors do note some problems with this prototype system. For example, they find that the "one sense per collocation" hypothesis on which much of the annotation process is based does not always hold. They also discuss the dream of a disambiguating dictionary—a dictionary that determines the sense of a usage in text, and returns the appropriate entry.

Rundell concludes this collection with a discussion of elegance and its importance in dictionaries, particularly for writing definitions. Rundell further comments on the continuing need for elegance in lexicography even as the space restrictions of dictionaries are reduced due to their increasingly electronic format.



## 4. Conclusion

Overall this Festschrift should be commended for presenting a range of research related to Hanks's career. Those interested in Hanks's work, or new ideas related to Hanks's work, would gain from taking a look at this collection. Also of potential further interest—especially for those interested in issues related to word senses and the relationship between words and their meanings in text—may be Hanks's forthcoming book detailing his theory of norms and exploitations (Hanks forthcoming).

This Festschrift also succeeds in drawing attention to several interesting and open computational problems related to lexicography and lexical acquisition.[3] Graduate students or other researchers wanting to learn more about problems in these areas may therefore particularly enjoy this collection. It should be mentioned, however, that those accustomed to the high standards for empirical research in this journal and at ACL conferences may take issue with some technical points in some of the contributions. Nevertheless, such issues do not detract from the presentation of interesting problems, and this collection may very well stimulate more computational research related to lexicography.

## References

Church, Kenneth W. and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver.

Hanks, Patrick. Forthcoming. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Johnson, Samuel. 1755. Preface to a Dictionary of the English Language.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.

Kilgarriff, Adam and David Tugwell. 2002. Sketching words. In Marie-Hélène Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Euralex, pages 125–137.

Pustejovsky, James. 2006. Type theory and lexical decomposition. *Journal of Cognitive Science*, 7(1):39–76.

Wilks, Yorick. 1977. Knowledge structure and language boundaries. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, volume 1, pages 151–157, San Francisco, CA.

3 Atkins concludes Chapter 15 with a quote that echoes Johnson (1755) and refers to the possibility of automatically mapping entries in DANTE and FrameNet: "These are the dreams of a lexicographer, doomed at last to wake in a universe of geeks."

*Paul Cook* is a MITACS Elevate Postdoctoral fellow in the Department of Computer Science at the University of Toronto. His research focuses on problems related to lexical acquisition and multiword expressions. He is a member of the Dictionary Society of North America, and was the recipient of the 2009 Dictionary Society of North America Award for Research in Lexicography. Cook's address is Department of Computer Science, 10 King's College Road, Rm. 3302, Toronto, Ontario M5S 3G4, Canada; e-mail: pcook@cs.toronto.edu.