

A Way With Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks, edited by Gilles-Maurice de Schryver. Kampala: Menha Publishers, 2010. Pp. vii + 375. €59.00. ISBN 978-9970-10-101-6

It has been very motivating to read a book as thoughtfully assembled as *A Way With Words*. It would have been no failure for the editor to gather a couple dozen suitable papers by esteemed colleagues of Patrick Hanks and publish them in a bundle without creating a deeper thematic unity. But under Gilles-Maurice de Schryver's editorial hand, the push and pull of ideas from paper to paper has created a conversation in which the reader can be an active participant. The *Festschrift* has a remarkable thematic coherence through a tightly woven assemblage of different approaches to the study of the lexicon. Starting from de Schryver's introductory chapter—a thorough, warm account of Hanks's career—the book as a whole makes a fitting tribute to a dedicatee whose influence has ranged widely indeed, substantially advancing both lexical theory and practical lexicography.

Thematic Overview

There is enough material in the *Festschrift* that a reader might take many paths through it and reach many conclusions,¹ but the book also seems to contain some thematic progressions in arguments that carry through the papers. Let me illustrate with two examples of such unifying paths.

Lexical descriptions. The appropriate definable unit in a lexical resource is a multiword expression (John Sinclair, “Defining the Definiendum”). A large, comprehensive standard for lexical entries will surely help computational systems understand natural language more deeply (Yorick Wilks, “Very Large Lexical Entries and the Boundary Between Linguistic and Knowledge Structures”). Some of the information that could influence such a lexical entry include: specific mechanisms of sense extension (James Pustejovsky and Anna Rumshisky, “Mechanisms of Sense Extension in Verbs”); patterns of semantic and syntactic governance (Igor Mel'čuk, “The Government Pattern in the Explanatory Combinatorial Dictionary”); an adjective's propensity to modify nouns

¹ For a complete table of contents, see the publisher's informational web page: <<http://www.menhapublishers.com/products.html#tablecontents>> accessed September 14, 2014.

of particular semantic classes (David Guthrie and Louise Guthrie, “Identifying Adjectives that Predict Noun Classes”); an orderly typology of semantic types (Silvie Cinková, Martin Holub, and Lenka Smejkalová, “The Lexical Population of Semantic Types in Hanks’s PDEV”); disjunction between semantic and syntactic roles in lexical patterns (Elisabetta Ježek and Francesca Frontini, “From Pattern Dictionary to Patternbank”); the noncompositional, pragmatic import of compositional phrases like “spring to mind” (Rosamund Moon, “Words that Spring to Mind: Idiom, Allusion, and Convention”); or valency constructions and collocations explicitly encoded by their pattern and pragmatic effect (Sue Atkins, *The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data*).

Inputs, theories, and outputs. Basing lexicographical judgments on collections of real usage data is fundamental and invaluable (David Wiggins, “The Paradox of Analysis and the Paradox of Synonymy”) and if you can get it in a clean form, you’re lucky (Kenneth W. Church, “More Is More”), but if you crawl the web, your corpus will inevitably be hard, dirty, redundant, and unbounded (Gregory Grefenstette, “Estimating the Number of Concepts”). A large amount of data is crucial to getting comprehensive coverage of the range of lexical and syntactic behavior (Alexander Geyken, “Statistical Variations of German Support Verb Constructions in Very Large Corpora” and also Grefenstette). Better tools can make better lexical analysis automatable, but such automation is trickier than one might hope (Adam Kilgarriff and Pavel Rychlý, “Semi-Automatic Dictionary Drafting”). Many theories make lexicography possible, but lexicography as a whole is too practice-driven to be unified by a single theory (Paul Bogaards, “Lexicography: Science without Theory?”). A dictionary’s prominence or obscurity is a function of the needs of the user, the data available, and especially the market where it is released; even obscure dictionaries may have impact beyond their sales numbers (Miroslaw Bańko, “The Polish COBUILD and its Influence on Polish Lexicography”). The history and historians of French argot are as fascinating and intricately woven as the underworld from which it sprang (Jonathon Green, “ARGOT: The Flesh Made Word”). An elegant definition is brief, concrete, and direct without aiming at exhaustiveness (Michael Rundell, “Defining Elegance”).

Discussion of Selected Papers

Hanks’s work has explicitly drawn from diverse fields for its theoretical background, including the philosophy of language and cognitive and

computer sciences, and the *Festschrift* reflects this diversity. Consider, for example, the juxtaposition of the papers by Wiggins and Church. Wiggins's paper is the last in the book's *Part I: Theoretical Aspects and Background*, and Church's is the first in *Part II: Computing Lexical Relations*, but both papers could have fallen on either side of the theory/computing boundary. Wiggins is an Oxonian philosopher, Church a computer scientist from Bell Labs and IBM; but the two authors' arguments for data-driven empiricism are informed, persuasive, and complementary to a degree just short of redundancy.

Corpus-driven lexicography requires dealing with human linguistic expression in all its diverse forms and with all its divergent motivations. On the challenges of corpus data, Church's paper engages nicely with the paper by Grefenstette, which follows it. Church's main argument, based on his experience with Associated Press newswire feeds, is that a larger amount of data will reliably be more informative: "We have better [. . .] technology for deduping [. . .], but the processes for creating dups have also improved. I don't know if we are catching up or losing ground" (139). Grefenstette responds with the dismaying observation that, in web corpora, we are losing ground. A corpus gathered through unsupervised trawls of the web is rife with bizarrely motivated kinds of text (spam and its analogues) that pollute a corpus with low-frequency noise, making it harder to do the simple-sounding task addressed in Grefenstette's paper: to count the number of two-word noun phrases in common use.

One of the hallmarks of Hanks's recent work has been the closer linkage of lexical judgments to the corpus data that drives it. A good number of the book's papers look for ways to explicitly record the insights that, in previous generations of lexicography, have been at best subliminal suggestions in the phrasing of a definition. These papers are particularly exciting for me from a practical computational standpoint, and they tend to be the stronger papers theoretically as well.

An important development towards making dictionary and corpus interdependent in a way that strengthens both is made in the paper by Cinková, Holub, and Smejkalová. They implement a shallow ontology of high-level concepts that can be adequate to filling all of the semantic-role slots in corpus pattern analysis at a suitable level of specificity. They propose that semantic role labels (such as Human, Road Vehicle, Institution, or Firearm) be manually recorded as an annotation layer on the source corpus, so that future work in that corpus may benefit from knowing the semantic role assigned by the lexical analyst.

Ježek and Frontini's brilliantly integrative paper puts into practice a great many of the book's theoretical and practical insights.

The paper builds on earlier work that studied semantic type coercion (Pustejovsky and Ježek 2008; Hanks and Ježek 2008)—a consequence of the semantic types used in corpus pattern analysis, whereby if you *read* a shampoo bottle it temporarily becomes an instance of the semantic class [[Document]] by virtue of being read. Ježek and Frontini show that a richer data structure, one that records both the semantic and syntactic types of corpus sentence, can allow a more refined understanding of lexical semantic behavior. This data structure also records the semantic activity present in anaphoric patterns that may be altogether erased by displays like word sketches: “The champagne arrived, and everyone drank it” wouldn’t typically count as evidence that champagne gets drunk, only that champagne arrives; Ježek and Frontini’s system would show both that champagne was drunk and that it was drunk anaphorically. They list new applications for the resulting “Patternbank,” which could provide wonderful new methods for both understanding and generating text—in addition to “a whole range of corpus-driven lexicographic products such as combinatorial dictionaries” (234).

Guthrie and Guthrie examine whether an adjective can predict the semantic class of the noun it modifies. They note that “one does not refer to ‘a cowardly building’ or ‘a glossy man’ as these adjectives seem to prefer nouns that are animate and inanimate respectively” (160). The authors propose a means of automatically deriving adjective preferences from a corpus and point to a rich vein for further research.

Moon’s paper shows a master lexical analyst at work. To this reviewer—an NLP-focused lexicographer with ambitions of teaching computers to deeply understand meaning—the paper is almost terrifying in its exposure of the many currents of pragmatics that cross beneath even the most mundane utterances. The paper may not break new theoretical ground, but it gives a glimpse of how deep the waters run beneath that ground.

Conclusions

The papers in the *Festschrift* range across theories and practices; some offer historical reflection, while others describe incremental steps toward the future. The arrangement of papers constitutes an argument of its own: a vision for the future of lexicography. Lexicographers and metalexicographers will find worthy and important papers here; historians will also be interested in the many personal anecdotes sprinkled through the theories. Not all of the papers are as densely informative, or straightforward to act upon, as the best of them, but each

contributes distinctly to a fitting celebration of Hanks's career (missing only a voice from Hanks's work in dictionaries of proper names). The book should remain a useful provocation for computational linguists for the foreseeable future: any time you think you've found a way to extract meaning from text, take another look at, say, Rosamund Moon's paper, and realize how much you're still missing.

The key question for a review to answer is whether the book in question is worth reading. For this book, the stakes are higher: is it worth tracking down? The *Festschrift* is not available on Amazon nor as an eBook. Worldcat reports only one copy in a US library and three in Canada. The book must be purchased directly from its Ugandan publisher, via wire transfer to Belgium. So a North American reader may hold little hope of encountering this book by chance: you will need to ask your library to order a copy or expend considerable effort to buy one for yourself. But the book is very much worth the effort to procure it.

Orion Montoya

References

- Hanks, Patrick, and Elisabetta Ježek. 2008. Shimmering lexical sets. In *Proceedings of the XIII Euralex International Congress (Barcelona, 15–19 July, 2008)* (Sèrie Activitats 20), edited by Elisenda Bernal and Janet DeCesaris, 391–402. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Pustejovsky, James, and Elisabetta Ježek. 2008. Semantic coercion in language: Beyond distributional analysis. In *Italian Journal of Linguistics / Rivista Italiana di Linguistica* 20.1: 215–240.