De Schryver, G.-M. (Ed.) 2010. *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Kampala: Menha Publishers. (vii + 75 pp.)

Linguistic field(s): lexicography, linguistics, corpus linguistics, computational linguistics

1. Introduction

A Way with Words is a book in honour of Patrick Hanks, a prominent lexicographer, corpus linguist and linguistic theorist. It brings together a collection of papers illustrative of recent advances in lexical theory and analysis.

As observed by Michael Rundell in the last paper of the volume, the field of lexicography is currently undergoing revolutionary changes. The arrival of internet dictionaries, ever-growing corpora and more intelligent software, require lexicographers to rethink their craft as we see particularly in the third part of the book. Furthermore, in linguistics there has been an increase in interest in the lexicon. The lexicon is assigned a more central role and different theories of the lexicon have emerged in various linguistic schools of thought, some of which are discussed in the first part of the book.

Thus, after years of splendid isolation, linguistics and lexicography are finally beginning to recognise the benefits of interaction. Bridging the insights of both fields will most certainly lead to a deeper understanding of how language works, which is a view that Patrick Hanks has been promoting throughout his career.

2. Contents

The book contains twenty papers in total, which are divided over three parts: a theoretical part, a computational part and a lexicographic part. In each part, the contributions have been placed in an order paralleling Hanks' career. We will not necessarily follow this chronological order when discussing them below, but we will tend to group the papers in each part by topic.

International Journal of Corpus Linguistics 17:1 (2012), **131–138**. DOI 10.1075/ijcl.17.1.06tib
ISSN 1384–6655 / E-ISSN 1569–9811

2.1 Part I: Theoretical aspects and backgrounds

The theoretical part contains five papers on various aspects of lexical meaning. It starts with a paper by the late John Sinclair, a first-generation modern corpus linguist, for whom Patrick Hanks worked as a project manager in the 1980s on the intellectually innovative COBUILD project. In this unfinished paper, Sinclair outlines his most radical approach to collocational analysis. He observes that the word is no longer the principal unit of meaning in a language, but that any configuration of text (single or multiple words) which has a distinct sense should be considered as a headword in a dictionary.

Yorick Wilks addresses the topic of large lexical entries and their senses from a more computational perspective and sketches how, within the Preference Semantics system, we can computationally deal with sense extensions that deviate from the norm using so-called pseudo-texts. Although the paper is actually a re-print of a text which appeared in 1977 "in obscure conference proceedings that no-one would now read" (p. 50), the topic is still very relevant and fits nicely within the context of Hanks' theory of Norms and Exploitations (In press).

James Pustejovsky and Anna Rumshisky move on to verbs and examine how to analyse different but related senses of a predicate within the framework of the Generative Lexicon (Pustejovsky 1995). They follow Hanks' notion that metaphoricity is gradable and claim that the different degrees of meaning extension (metaphoricity) can be modelled by a number of formal processes operating on the predicate.

The paper by Mel′čuk presupposes a great deal of background knowledge in both Meaning-Text Theory (MTT) and Explanatory Combinatorial Dictionary (ECD), especially with regard to the formal notations. Mel′čuk develops a demonstration, within the framework of Meaning-Text Theory, of how lexical government patterns function as a device to map different levels of linguistic representation onto each other, and of how they should be described in the dictionary of meaning-text models (the Explanatory Combinatorial Dictionary). After a brief introduction to the theoretical framework, to the notions of government and government patterns themselves as well as to the notion of constraints on government patterns, Mel′čuk successively discusses the correspondence between the semantic and the deep syntactic actant slots of lexical units, and the way in which the latter are translated into surface syntax patterns.

In the final paper of the theoretical part, David Wiggins discusses a well-known problem with analytical sentences which he, following C. H. Langford, calls "the paradox of analysis". In an analytical sentence having the form *X is Y*, both the analysandum *X* and the analysans *Y* have the same meaning, which makes the formula a tautology and thus something trivial. If on the other hand they did not

have the same meaning, the sentence would not be analytical. The same paradox manifests itself in sentences where *X* and *Y* are synonyms: either the sentence is tautological or *X* and *Y* are not synonymous. But if *X is Y* is a tautology, it should have the same semantics as *X is X* or *Y is Y*, which is not the case. The Fregeian distinction between sense and reference cannot help to resolve the paradox, since in the case of synonyms X and Y have the same sense and the same reference. One way to avoid the paradox, as proposed in this paper, is to make the word-sense-reference correlations dependent on contextual discriminations, such that, for instance, a distinction can be made between a purely tautological interpretation of *X is Y* and a nontrivial interpretation in which one speaker who is familiar with *X* and *Y* instructs another speaker who only knows *X* about what *Y* is.

From the theoretical discussions we move onto the computational part of the book.

2.2 Part II: Computing Lexical Relations

Part II starts with a contribution by Kenneth Church with whom Patrick Hanks wrote the influential paper "Word association norms, Mutual Information, and lexicography", which reintroduced statistical methods of lexical analysis in linguistics back in 1989. The contribution by Kenneth Church can be seen as a response to Adam Kilgarriff's (2007) opinion piece in *Computational Linguistics* "Googleology is a bad science". If we have to choose between a large corpus and a representative corpus, the question is which is more important. Church's conclusion is that different applications require different corpora and thus there is no easy answer to whether quantity should be preferred over quality.

Corpus size is also discussed in the paper by Alexander Geyken, who investigates two particular research questions, namely to what extent growing corpora provide us with more information, and related to that, how large corpora need to be in order to contain at least all constructions listed in a large monolingual dictionary. Geyken's study focuses on support verb constructions in German and is based on a corpus study of two large corpora and a large monolingual dictionary. He compares the data in the dictionary with the corpus data retrieving the statistically salient co-occurrences. Geyken's study confirms Hanks' claim that "in a corpus of 100 million words, a simple right- or left-sorted concordance shows clearly most of the normal patterns of usage for all words except the very rare" (Hanks 2002: 157). However, in corpora of this size rare patterns cannot be found on the basis of statistically salient occurrences alone. Geyken also observes that with increasing corpus size, new constructions can be found in the interval between the size of 100 million and 500 million tokens, but that this is much less so in the interval between 500 million and 1 billion tokens. Finally, he concludes that

both corpora which were used in this study contained lexicographically relevant material which was missing in the dictionary, highlighting again a point made by Hanks that words only have meaning in context and that lexicography should thus be corpus-based.

Corpus size also comes into play in the paper by Karel Pala and Pavel Rychlý who evaluate the output of the Czech word sketches. Word sketches are one-page, corpus-derived summaries of the grammatical and collocational behaviour of words. They are automatically produced by the Sketch Engine, a sophisticated corpus query system (Kilgarriff et al. 2004) and the best results are obtained when they are based on large amounts of data. Pala and Rychlý observe that the quality of the output of the word sketches is still relatively low for Czech, mainly due to tagging errors and imperfections in the sketch grammar, and they provide some suggestions on how the results could be improved.

Gregory Grefenstette used corpus data from the web to estimate the number of multiword concepts that are currently used in English. Starting from a list of nouns and adjectives from the DELA dictionaries, he checked which adjective-noun and which noun-noun pairs occurred on more than 5 different web pages, a threshold set by Grefenstette for being in common use.[1] In total around 200 million two-word combinations occurred on more than 5 different pages which led Grefenstette to conclude that there are roughly 200 million concepts that are useful for the lexicographers of the future.

The paper by David Guthrie and Louise Guthrie deals with automatic disambiguation. They describe and evaluate an automatic method for disambiguating nouns using adjectives. Their results show that adjectives do indeed provide a great deal of information about the semantic class of the nouns they modify. They show that completely unsupervised data gathered by identifying nouns which are unambiguous with respect to semantic category, and collecting their modifying adjectives, does provide substantial information that can be used to tag an unseen set of noun phrases reliably, where those same adjectives are used with head nouns that are both new and ambiguous.

The last two papers in part II revolve around Hanks' Corpus Pattern Analysis (CPA), a new technique for identifying the typical patterns in which a verb is used in corpus context.[2] CPA is currently being used to build a *Pattern Dictionary of English Verbs* (PDEV) (cf. Hanks 2007), which will be a fundamental resource for use in computational linguistics, language teaching, and cognitive science. The contribution by Silvie Cinková, Martin Holub and Lenka Smejkalová reports on an ongoing analysis of PDEV with respect to both its consistency and reproducibility of use by different users. It focuses on the assignment of Semantic Type labels to noun collocates of verbs. It also discusses a pilot study of automating

the lexical population of Semantic Types intended to facilitate a future large-scale expansion of PDEV.

Elisabetta Jezek and Francesca Frontini extend PDEV to cover Italian. Their paper focuses on the tension between the semantic types (ST) associated with verb arguments and their extensional definition, i.e. the lexical sets (LS) that may fill the different argument positions. They propose that the mismatch between STs and LSs can be partly remedied by extending the Corpus Pattern Analysis technique used in PDEV, so that it includes the annotation of verb patterns onto the corpus instances that instantiate them, i.e. by creating what they call a 'Patternbank'. In their paper they report on the first steps taken in the planning of a 'Patternbank' for Italian.

2.3 Part III: Lexical analysis and dictionary writing

The third part of the book links through to lexicography proper, dealing with the whole spectrum of present-day lexicography from manual analysis to automatic dictionary compilation.

Rosamund Moon kicks off with a detailed corpus study of the phraseology of *spring to mind*. Such patient studies at word-face are necessary to show the systems that are at work in language and to allow explanations to be formulated.

Another detailed lexical study is offered by Jonathon Green, an expert on slang dictionaries, who gives a detailed definition of and provides the etymology of the lexeme *argot*.

Dictionaries, generally, do not have the luxury of such extended discussions of individual items due to space and time restrictions, unless, of course, part of the process is automated. Adam Kilgarriff and Pavel Rychlý, present a prototype of such a piece of software which they call SADD, Semi-Automatic Dictionary Drafting. Their program allows the computer to semi-automatically perform the first step of word sense disambiguation doing much of the footwork for Corpus Pattern Analysis. Although the prospects are promising, Kilgarriff and Rychlý note rightly that to turn the prototype in a viable system for production-mode lexicography, a lot of work remains to be done.

The DANTE lexical database for English, discussed by Sue Atkins is also a good illustration of a contemporary lexical resource providing a systematic corpus-based description of the meanings, grammatical and collocational behaviour, and text-type characteristics of over 42,000 headwords, 23,000 multiword expressions, and over 27,000 idioms and phrases. In her paper, she outlines an approach for enriching the FrameNet database with additional information present in DANTE, which would result in a very rich lexical resource with potential for a range of applications in computational linguistics and lexicography.

These first three papers in the third part of the book illustrate nicely that lexicography is currently a rapidly changing field. Consequently, lexicographers have begun to see the need for better theoretical foundations, a topic discussed by Paul Bogaards. In "Lexicography: Science without theory?" Bogaards wonders whether there is currently a theory in lexicography. Based on an analysis of papers published in the *International Journal of Lexicography* and recent books on lexicography, Bogaards observes that the general trend seems to be that lexicographers long for a good theory in order to improve the practice of dictionary writing, but that lexicography is, as most practices, without one single theory. Lexicographic practice is far too varied and multi-faceted to allow for one independent theory covering all aspects of the subject field.

Innovation and theory are also key in the paper by Mirosław Bańko. In this autobiographical note, Bańko describes how he came to the idea of compiling a COBUILD-like dictionary of Polish, the *Inny słownik języka polskiego* and how the concept of the COBUILD dictionaries indirectly influenced Polish lexicography. Although the effects of the Polish COBUILD project fell short of the expectations, the project is a good illustration of how lexicographic innovations, when first introduced, may not catch on or may be absorbed in a different form than originally intended, and that even small effects are important, especially in lexicography.

Michael Rundell concludes the book with a paper on "Defining elegance" observing that lexicography needs elegance now more than ever. Lexicographers have more and more data at their disposal and they can make more detailed analyses of words in context using more and more sophisticated software to help them. In this new scenario, the crux is not to list everything but "knowing what not to say".

3. Evaluation

The book forms a valuable contribution to our understanding of the lexicon. It highlights two clear trends in lexicography. First, the recognition of multi-word units as lexical items. For the past twenty years evidence has been growing that the word is no longer the principal unit of meaning in a language, but any configuration of text (single or multiple words) which has a distinct sense (p. 38). Many papers in the book deal with large lexical entries, multiword expressions and idioms, much in the spirit of Patrick Hanks who proposes that lexicographers need to "do away with words" in order to focus on phraseology (p. 4).

The second trend that can be observed is the importance of objective empirical evidence when describing language. It has become clear that words (or better, groups of words) should not be studied in isolation but in context. Patrick Hanks

observes in his 2000 publication "Do word meanings exist" that word meanings do exist but only in context.

As a consequence, it has become impossible to imagine lexicography today without large digital corpora and corpus methods and tools to analyse them (at least for the major languages). Within lexicography, there exists a general consensus that for creating general-purpose dictionaries, corpus size is to be preferred over granularity (Atkins & Rundell 2008). However, the analysis side of the data still gives rise to deep issues. What lexicographers want from a corpus differs from what other kinds of corpus linguists may want, thus requiring different solutions. In a recent publication, for instance, Rundell & Kilgarriff (2011) point out that although corpus methods can straightforwardly find high-frequency single-word items and thereby provide a fair-quality first pass at a headword list for those items, they cannot yet do the same for multiword items. Optimising the corpus-query software which enables the lexicographers to efficiently track down lexicographically relevant facts in corpora will be a main concern in the near future and advances in corpus linguistics can help here. That this may not necessarily involve sophisticated maths is discussed by Kilgarriff (2009). In the years to come, the interaction between lexicography and corpus linguistics will become stronger, and it is this kind of dialogue between two fields which will lead to new insights.

The book also provides an insight into the work and ideas of Patrick Hanks and as such can be considered as background reading to Hanks' forthcoming book –*Analyzing the Lexicon: Norms and Exploitations*–, as the reader will see how (former) colleagues and friends have inspired Hanks and how Hanks has inspired them. Hanks' theory of Norms and Exploitations is a central theme in the book, often referred to by the different authors. It states that a natural language is indeed a rule-governed system of linguistic behaviour, but that there are two systems of rules. One rule system governs normal phraseology and meaning of words in use, while the other allows language users to exploit normal phraseology in all sorts of creative ways.

The editor of the Festschrift, Gilles-Maurice de Schryver, has done a meticulous job. The book has been carefully edited assuring the overall quality of the papers, both linguistically and technically. The book reads pleasantly, as do several amusing anecdotes about Patrick Hanks. The papers can be read individually, but they do form a coherent collection without too much overlap when the book is read from cover to cover. We have found many papers very stimulating and informative and thoroughly recommend the book to those interested in the issues which are currently at stake in lexicography.

Notes

1. For the DELA dictionaries cf. http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Dictionnaires/download.html (accessed December 2011).

2. http://nlp.fi.muni.cz/projekty/cpa/ (accessed December 2011).

References

Atkins, S. B. T. & Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Hanks, P. 2000. "Do word meanings exist?". *Computers and the Humanities*, 34 (1–2), 205–215.

Hanks, P. 2002. "Mapping meaning onto use". In M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* Grenoble, France: Euralex, 156–198.

Hanks, P. 2007: online. *Pattern Dictionary of English Verbs (PDEV) — Project Page.* Available at: http://deb.fi.muni.cz/pdev/ (accessed December 2011).

Hanks, P. In press. *Analyzing the Lexicon: Norms and Exploitations*. Cambridge, MA: MIT Press.

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

Kilgarriff, A. 2009. "Simple maths for keywords". In M. Mahlberg, V. González-Díaz & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference CL2009, University of Liverpool,* 20–23 *July*. Available at: http://ucrel.lancs.ac.uk/publications/cl2009/ (accessed December 2011).

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. 2004. "The Sketch Engine". In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July* 6–10. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne-Sud, 105–116. [See also: http://www.sketchengine.co.uk]

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Rundell, M. & Kilgarriff, A. 2011. "Automating the creation of dictionaries: Where will it all end?". In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora. In Honour of Sylviane Granger*. Amsterdam/Philadelphia: John Benjamins, 257–281.

Reviewed by Carole Tiberius and Frans Heyvaert,
Institute for Dutch Lexicology (INL)