Breadth and Depth of Semantic Lexicons

Evelyne Viegas (editor)
(New Mexico State University)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 10), 1999,
xix+328 pp; hardbound, ISBN
0-7923-6039-7, $144.00, £89.00,
Dfl 270.00

Reviewed by
John S. White
Litton PRC

1. Introduction

Breadth and Depth of Semantic Lexicons is based on a 1996 workshop. It represents several fundamental lines of thinking that were, and remain, a useful organization of the issues in lexical semantics, and specifically in the organization and development of new generations of lexical corpora. The volume contains very useful background discussions of classic issues in lexical semantics (e.g., English count/mass nouns), and some rather striking claims about what should and should not be included in a lexicon.

Several of the papers show progress along contemporary models of lexical semantics, in particular the various current theories of lexical rules. The book represents a period, ongoing today, that welcomes the convergence of linguistic semantic thought and its computational implications. The conclusion seems to be, as asserted both by Nirenburg and Raskin and by Helmreich and Farwell, that certain of the theoretical or heuristic linguistic models of lexical semantics are useful for formulating computational implementations, but remain profoundly problematic as linguistic theories.

The seminal positions from which the papers generally proceed are not directly represented in the volume, primarily Levin's work on English verb classes (Levin 1993), and Pustejovsky's generative lexicon (Pustejovsky 1995). Nevertheless, the currency of these constructs forms the context for the organization of the workshop and the volume. These positions differ from each other, naturally, but each has the effect of claiming that there are generalizations that may be captured between lexical items and either forms or functions, and thus that the lexical entry itself may be simpler, underspecified for elements of reference, or resident in a smaller, nonpolysemous lexicon.

Such theoretical elegance appears at first examination to benefit efficiency at implementation time, which is what attracts computational linguistics to these approaches. However, the difficulties encountered in trying to extend their coverage at the same time delimit the level of their usefulness for system development, while helping the theorists further the instantiation of the paradigms implied by their theories. By the juxtaposition of theory and implementation we become aware of the problems at the poles of the current models: the claims they make are so general that numerous exceptions arise in implementation (like the Levin classes); or they require such specificity that their limited coverage (e.g., the conversion of a count noun to a mass noun in the case that the referent is an animal fur) makes implementation unproductive.

The volume is organized into four sections: "Lexical rules and underspecification," "Breadth of semantic lexicons," "Depth of semantic lexicons," and "Lexical semantics and pragmatics." The intent of this organization is to cover the considerations in building semantic lexicons, and the goal is achieved both in the characterization of the issues of the time and the general concerns that will have to be addressed for some time to come.

2. Lexical Rules and Underspecification

This section concerns the assertion and application of lexical rules as a means of reducing the static size of the lexicon—that is, to distill out all predictable properties and create lexicons that contain only idiosyncratic information. The boundaries of such rules—for example, whether they should also cover phenomena hitherto considered morphological—form the basic body of issues in this section.

"Categorization of types and application of lexical rules" by Boyan Onyshkevych describes the implementation issues associated with the coverage of lexical rules. This is a good organization and reference paper describing the implications for lexicon and processing efficiency of covering particular phenomena in lexical rules, and also of applying these rules (thereby extending the lexical corpus) at particular points in natural language processing systems. This paper actually reads like two separate papers, the latter half reviewing several lexical rule implementations.

Also valuable as a reference work in this section is "The lexical semantics of English count and mass nouns" by Brendan Gillon. This paper provides a synopsis of the empirical contrasts of mass and count nouns, at both the phrase and clause levels in English. This paper is worth reading for the concise presentation of the phenomena and the pertinent syntactic tests. The same bent toward empiricism compels Gillon to be very careful about characterizing the lexical rules associated with the conversion from count to mass and vice versa, and as a result he seems hesitant to commit to more general claims that would combine the animal–meat and animal–fur phenomena (where the count noun animal term is converted to a mass noun).

The remaining two papers in this section show some of the variety of thinking about the extent of lexical rules. Sehitoglu and Bozahin ("Lexical rules and lexical organization") describe a model of lexical rules for highly agglutinating languages such as Turkish, in which lexical rules handle inflectional and derivational morphology as well as classic lexical-semantic phenomena. Sanfilippo ("Word disambiguation by lexical underspecification"), using the Pustejovsky notion of "qualia" to decompose lexical reference, argues that lexical underspecification coupled with mechanisms for capturing local grammatical context can account for disambiguation without lexical rules. These two papers express near-extremes with respect to the articulation of lexical rules, but share the goal of reducing overgeneration of forms and interpretations.

3. Breadth of Semantic Lexicons

This section of the volume is for the most part a contrast in views of Levin's verb classes. Both Dorr and Jones ("Acquisition of semantic lexicons") and Burns and Davis ("Building and maintaining a semantically adequate lexicon using CYC") attempt to take advantage of the encyclopedic work of Beth Levin (1993) in categorizing verbs according to an association of meaning and subcategorization frames. Levin's book does not claim to prove that there is a hard-and-fast correlation between the form of verb arguments and the meaning of the verb; rather, Levin "assumes" the association. Since its publication, many researchers have been torn between attempting to employ the categorizations as a handy organization of English verbs, and trying to test the hypothesis that Levin appears to assert as axiomatic. Both the papers in this section that deal with Levin classes face this conflict. Dorr and Jones attempt to match Levin's classes with the subject-area codes in the *Longman Dictionary of Contemporary English* to automate lexical acquisition, using syntactic cues. Burns and Davis attempt to use Levin's classes to make generalizations about large groups of verbs for the natural-language linkage to the CYC knowledge base. In both cases, the authors find that Levin's work is useful as a comprehensive description of subcategorization patterns for lexical implementation, but much less reliable as a predictor of semantic behavior.

Also in the section on breadth is the first of two papers by Nirenburg and Raskin, "Lexical rules for deverbal adjectives," in which the authors present an ontology-based method of characterizing lexical rules for adjectives. They introduce lexical rules intended to capture the semantic side of the derivation from verb to adjective, but find numerous exceptions, gaps, and departures from an already stretched assumption about what is deverbal (e.g., *international*). They add to the other voices in the volume who recognize the value of lexical rules only when they are practical for implementation.

4. Depth of Semantic Lexicons

As Levin was the absent guest in the section on breadth, Pustejovsky's generative lexicon is a theme in the papers on depth. Two papers use the qualia structure to explain coverage phenomena in noun phrases. In "The adjective *vieux*: The point of view of 'Generative Lexicon'," Pierrette Boullion makes a strong claim for the value of generative lexical processes to account for the otherwise underivable range of meanings of adjectives like the French *vieux* 'old'. Here, both the scalar and property-modifying senses of *vieux* can be distilled down to a single meaning by showing that the scope of modification is not the reference of the modificand but rather parts of the semantic decomposition of it, and further that the syntactic distribution of the adjective is consistent with this explanation. Michael Johnston and Federica Busa approach noun compounds in terms of qualia structures in "Qualia structure and the compositional interpretation of compounds." Clues about noun compounding in both English and Italian suggest the means by which aspects of noun modifiers instantiate variables in the modificand. While this approach should appear promising for machine translation and multilingual information retrieval, other issues remain, such as ambiguity or overgeneration in noun compounds of three or more nouns.

The holy grail of lexical semantics (in computational linguistics, at least), has been automatic accumulation of word senses in the right granularity, and ultimately the automatic disambiguation of words in text. Machine-readable dictionaries provide the well-known relations of hyponymy and meronymy as hierarchical relations, but also can provide clues about collocational facts that will enable automatic disambiguation. Jen Chen and Jason Chang ("Integrating machine readable dictionary and thesaurus ...") investigate these principles as discoverable from the *Longman Dictionary of Contemporary English* merged with other collections such as *Roget's Thesaurus*, the *Longman Lexicon of Contemporary English*, and WordNet. Applying information retrieval-techniques, they show the potential of forming word-sense clusters across multiple collections to achieve senses for disambiguation that are neither too coarse-grained nor too fine-grained to be useful for text handling.

Burstein, Wolff, and Lu ("Using lexical semantic techniques to classify free-responses") present a very engaging application of lexical semantics to educational testing in the scoring of free-word test responses. A comprehensive review of lexical approaches finds no one perfect model for representing paraphrases, but relying on a sublanguage-dependent model demonstrates that concept-based lexicons for scoring written responses show promise.

5. Lexical Semantics and Pragmatics

This section seems more of a workshop complement section than a directed focus on pragmatics, but contains implementation attempts and engaging theoretical claims. Christiane Fellbaum ("Semantics via conceptual and lexical relations") presents a discussion of WordNet, one of the preeminent research corpora for eliciting and examining lexical relations. Of particular interest are the lexical gaps, concepts suggested by the semantic hierarchies that do not have a lexicalization. There is some discussion of the potential for universal lexical gaps, a discourse reminiscent of early cognitive anthropology. The relational structure of a lexical database does reflect some conceptual reality as far as organization, but it may also obscure other salient relations—thus some of the apparent gaps may not actually be conceptualized at all, while there may be other, covert taxonomic levels that remain undiscovered.

Evelyne Viegas, the editor of the volume, in "Opening the world with active words and concept triggers," addresses the generation of new entries and new concepts within the Mikrokosmos environment, as a means of transcending the time-intensive knowledge-base population processes. Lexical rules derive a new lexical entry from a semantic relation and a base form (*buyer* as the agent of *buy*), performing derivational morphology as well as creating a lexical entry and linking it to a knowledge-base concept. The approach may be more difficult to generalize than Viegas currently envisions, given such common forms as *teller, prayer, beer,* and so on.

Raskin and Nirenburg apply their perspective gained from adjectival semantics to a more general discussion of research and theoretical approaches, in "Supply-side and demand-side lexical semantics." Here they distinguish between approaches that proceed by incremental, syntax-like, coverage of small, interesting phenomena, and those that try to cover an entire sublanguage semantically with or without existing tools to do so. Their description of Mikrokosmos, microtheories that converge to cover a discourse, is a mediating position. In both their previous paper and this one, the authors assert certain lexical phenomena as universal: that adjectives and participles are not really distinct in any language; and that every language has a word that covers the same generalized notion of *goodness* as the English adjective *good*. Assertions like this beg for empirical confirmation, and I trust that the authors will refrain from reasserting them until such studies are done.

Another challenging claim comes from Helmreich and Farwell, who argue that the semantics of a lexical entry is very sparse indeed, and that there is no need for lexical rules at all, given the linguistic and pragmatic context of the expression. This is general enough a claim to allow for the mechanism of count-to-mass noun conversions, for example. But it seems also to overgenerate, especially from an implementation perspective. Their model would seem to imply a syntactic principle, say, that allows *in school* and *in jail* in American English, and *in hospital* in British English, but never *in apartment*. Implementation should simply see the former set as idiosyncratic and thus appropriately lexical, but ruling out the latter requires both syntax and semantics.

In summary, the Viegas volume is of value in providing overviews of several descriptive and theoretical aspects of the building of lexical semantics within language systems. The viewpoints express sufficient polarity as to suggest both committed enthusiasm in continued exploration, and also a sort of millennial expectation of some unifying breakthrough. While the volume is at times curious in its organization, and has enough typographical errors to be a bit distracting, it is still a quite useful guide through the theories, claims, and issues in computational lexical semantics.

References

Levin, Beth. 1993. *English Verb Alternations: A Preliminary Investigation*. The University of Chicago Press.

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

John White is Director of Research and Development for Litton PRC. In this capacity he is engaged in evaluation of multilingual text-handling approaches and systems. His research efforts have included machine-readable dictionaries, lexical semantics, natural language interfaces, and machine translation. White's address is: Litton PRC, 1500 PRC Drive, McLean, VA 22102 USA; e-mail: white_john@prc.com.

Most of the books about computational (lexical) semantic lexicons deal with the depth (or content) aspect of lexicons, ignoring the breadth (or coverage) aspect. This book presents a first attempt in the community to address both issues: content and coverage of computational semantic lexicons, in a thorough manner. Moreover, it addresses issues which have not yet been Most of the books about computational (lexical) semantic lexicons deal with the depth (or content) aspect of lexicons, ignoring the breadth (or coverage) aspect. This book presents a first attempt in the community to address both