Perspectives on the Lexicon

(Summary)

Written by
Ágoston Tóth

Thesis supervisor: Dr. Béla Hollósy

University of Debrecen
2005

1. Objectives

While specifying a lexicon is sometimes treated as a follow-up to developing a new model of grammar or an application, lexicon design, that is the description of how we store and handle idiosyncratic building blocks (usually morphemes and/or words) of language, is one of the most complex problems of linguistics and by no means secondary in importance. We should consider a wide variety of theoretical questions while keeping an eye on the implementational consequences.

Chapter 2 of my thesis is devoted to the analysis of the position of the lexical component in selected alternative models of Chomskyan Generative Grammar. We study the intricate relationship between syntax and morphology (concentrating on the placement of inflection and derivation) and its effects on how we think about the lexicon. This chapter also gives some basic insight into how morphological processes are approached in Natural Language Processing. *Chapter 3* discusses the problem of homonymy, polysemy and other aspects of lexical semantics to show that meaning is not easy to grasp. Since a lexicon should store (and perhaps work with) meanings, these are relevant considerations. While chapter 3 and the greater part of chapter 2 explore theoretical perspectives, *chapter 4* takes the more practice-oriented point of view of computational linguistics. The main considerations of chapter 4 are not entirely different from those of the previous chapter, however, since we inquire into the question of *representing lexical knowledge* in real-life databases, most importantly, in WordNet and FrameNet. WordNet's sense-relations and the frame-relations in FrameNet are exciting new approaches to representing meaning in a database that is designed to contain lexical information.

While the literature of Generative Grammar is abundant in references to lexicon-related considerations, *chapters 5 and 6* are devoted to a much less researched topic: representing linguistic input and the emergence of a special type of lexis in connectionist models. Chapter 6 presents my own artificial neural network-based experiment in which I trained a network to recognize FrameNet-like frame and frame element labels in a fully or partially pre-trained corpus of generated text. The following design goals have been implemented:

- The system should be able to handle an infinite number of input elements.
- The system should be able to take complex expressions (idioms, compounds or phrases) as input elements.

– No explicit morphological or syntactic information is inputted, but input elements should not be atomic, and the system should have access to the internal structure of input elements.
– The system should not be restricted to processing isolated sentences.

I examined the performance of the network with input that was slightly different from the training data: section 3 and section 4 of this summary overview the testing methodology and the results of the experiments.

I would like to emphasize that the present thesis does not try to prove that connectionism offers superior solutions to those aspects of lexicon design that are presented here as problem issues; in fact, neural network modeling has its own pitfalls and problems to solve. I only aim to add a noteworthy, highly neglected perspective to lexicon design: a non-traditional, connectionist approach to representing certain aspects of linguistic knowledge.

2. Theoretical background

2.1 The lexicon of Generative Grammars

The chapter on Generative Grammar is organized around the following questions:
– Is morphology an independent component in the selected models of Generative Grammar?
– Where is concatenative morphology (more precisely: derivational and inflectional affixation as well as compounding) implemented in various versions of Generative Grammar?

Our starting point is Chomsky (1957): he omits morphology as a separate component from his model. Then we see the rise of morphology as an independent component of grammar, only to be rejected in Lieber's (1992) GB-based unified morphosyntactic model. Based on the discussion of the treatment of morphological processes, the sections in this chapter try to show how these models delineate the notion of the *lexicon*. The chapter concludes that the role of the lexicon and the function of syntax have been treated as related questions in the generative tradition. Those who accepting the *strong lexicalist* hypothesis, which is represented by Halle (1973) and Lieber (1981) in the thesis, see a morphology which is completely detached from the syntactic component. The resulting lexicon is not simply a storage space for idiosyncrasies, but it has morpholexical functions and processing responsibilities. If the *weak lexicalist hypothesis* is accepted, a division of labor must be postulated between the lexicon and syntax from the point of view of morphological processes: inflection is treated by syntax, while less systematic (less productive or "quasi-productive")

processes end up in the lexicon. Finally, accepting Lieber's (1992) approach means that morphology is non-existent in the lexicalist sense, and the role of the lexicon is demoted accordingly.

A section of chapter 2 deals with the Natural Language Processing implementations of morphological phenomena. Grammars developed for parsing are usually quite different from the transformational grammars that underlie the models presented in the first part of this chapter, because computational linguists are seeking devices that are *efficient,* too, while this is not an issue in theoretical linguistics. Prószéky (1989:53) points out that these two paths of seeking appropriate grammars remained completely isolated through the end of the 1970s. Unfortunately, the literature of morphosyntactic parsing does not seem to have been proliferating. Kashket (1986) is one of the few exceptions. He outlines a GB parser for Warlpiri, a language with free word order. His system uses morphological case information to identify syntactic arguments in the "lexical parsing" stage (working with a lexicon that lists morphemes), then the same parser engine carries out syntactic parsing to assemble the final phase marker from the output of the first stage. His system has its own problems, however, and has not attracted significant attention in the linguistic literature. Finite-state solutions seem to be preferred in the morphology-related literature of NLP. The most successful model that uses finite-state networks is probably Kimmo Koskenniemi's *two-level morphology* (cf. Koskenniemi 1983), which is discussed briefly in the thesis, together with Gábor Prószéky's solution to morphological analysis, called Humor (*H*igh-speed *U*nification *Mor*phology, cf. Prószéky 1994). Its derivative, the HumorESK system, can be used for morphosyntactic analysis, too (cf. Prószéky 1996).

2.2 Sense delineation and the Polysemy - Homonymy distinction

Modeling the morphosyntactic behavior of languages is but one issue that is relevant to lexicon design. Any kind of linguistic project that involves the enumeration and delineation of word senses (including the compilation of dictionaries and the development of lexicons for NLP applications) must account for many issues related to word meaning, including *polysemy* and *homonymy*. The polysemy-monosemy and polysemy-homonymy 'dichotomies' are introduced as continua (cf. Cruse 1986, 2000), which foreshadows the problem of representing meaning variations and various readings of a word in a dictionary or a lexical database. Let me also highlight the section on Verspoor's (1997). She argues that traditional dictionary design hinges on "the discreteness of the meaning expressed in a usage of a word – where there is ambiguity, only one sense of a word can be active at any one time" (p. 219),

and these distinct senses correspond to distinct entries in the lexicon. She goes on to explain that this view is incompatible with the idea of underspecified representations and emphasizes a potential clash with Cruse's (1986) sense-modulation theory. Verspoor points out that the use of a finite set of pre-fabricated choices "from which the NLP system can choose the most appropriate" has remained a practice nevertheless (p. 220).

2.3 WordNet, FrameNet and representing lexical knowledge in a lexical database

Chapter 4 includes a report on two major relational lexical databases, WordNet and FrameNet. They are likely to influence NLP system design for a utilitarian reason: building a database that represents the idiosyncrasies of a language is a labor-intensive task, so existing major data sources that are freely available are being integrated into a large number of NLP systems. WordNet's sense-relations and the frame-relations in FrameNet are exciting new approaches to representing meaning in a database that is designed to contain lexical information. MindNet, which is also discussed in this chapter, illustrates that it is possible to store lexical information that seems compatible with the important advancements of lexical semantics (more specifically, Cruse's conception of word meaning). The discussion of a spreading activation approach to storing word meaning is also included.

WordNet distinguishes between minuscule meaning variations assigning them to synonym sets that are in some cases difficult to tell apart even for the human user. The 'highly homonymous' lexical content of the database causes problem for word-sense disambiguation, too. The thesis surveys various solutions for 'compacting' WordNet senses, including Mihalcea és Moldovan (2001), Chen and Chang (1998) and Seagull (2000). This chapter also contains my analysis of how derivational morphology is treated in WordNet. The results show that the CELEX database (that is its 18-million-word corpus source) shows similar proportions of derived nouns as the noun word-stock of WordNet.

2.4 Artificial Neural Networks in Linguistics

Pléh (1998) argues that symbol manipulation and connectionism are two approaches to cognitive science; while chapter 2 deals with a prominent linguistic manifestation of the symbol-manipulation approach, the final chapters take us to the connectionist side. **Chapter 5** outlines the general features of connectionism and surveys selected neural network models. Via an overview of the experiments reported in Elman (1990), Kohonen (1984), James and Miikkulainen (1995) and Mayberry and Miikkulainen (1999), the chapter illustrates possible approaches to representing linguistic input in an Artificial Neural Network.

A common feature of the models developed in the connectionist framework is an ability to take and coordinate input from a plethora of sources while finding regularities and particularities automatically during the training process (whether it is supervised or unsupervised). As a down-to-earth problem, however, all connectionist models face the difficulty of representing temporal sequences in general and *linguistic input* in particular.

McClelland and Rumelhart's (1981) Interactive Activation (IA) model is constructed to model human letter perception in a word recognition task. The system takes orthographic features of four alphabetical characters comprising a single word as input and ouputs the corresponding word. The model uses a strictly localist representation of linguistic features.

The thesis also surveys Elman's (1990) experiments with his novel network architecture, the Simple Recurrent Network (SRN). This model supports dynamic allocation of network resources to implement *memory*, which is not a commodity in neural network systems. Elman's experiments also have a strong linguistic bias (please note that neural networks are not only used for linguistic purposes, although linguistics is one of the major areas of their application). One of Elman's experiments investigates if and how SRNs can identify subsequences in a sequence of inputs. The network is constructed to accept three consonants and three vowels on its input; the representation is based on six phonological features. In another experiment, an SRN carries out sentence segmentation. Elman (1990) also presents an SRN experiment that aims to show that an SRN can discover word classes from word order information.

At the end of the chapter, I describe Kohonen's Self Organizing Maps and one of its derivatives that can process temporal sequences, too. Kohonen maps are used in my own experiment, too.

3. Form and meaning in an Artificial Neural Network: A twinmap-driven frame-relative frame-element recognizer network

Chapter 6 of the thesis reports on an Artificial Neural Network (ANN) experiment that features a unique design that marries up a non-localist, self-organizing input interface taking syllabified phonetic transcriptions as input with a recurrent structure that carries out a *semantic parsing task*: FrameNet-style frame and frame element recognition. The system is trained on a large and highly redundant set of generated sentences, which is reflected in the extremely high recall and precision figures that are achieved by the network. The performance

of the system is also assessed when exposed to input which is not present in the training corpus.

The network consists of an input interface, which I call the "twinmap", because it is a combination of two Kohonen maps working in parallel, and a recurrent part above it, containing two SRN-like constructions.

In each step, the twinmap takes a word that is transcribed phonetically on a per-syllable basis as input, and produces an activation pattern that is representative of the input word. A pre-training phase is used to develop these representations. The syllabic structure groups use approximately 400 neurons, each representing one syllable type. The two groups accept different feature vectors as input. The unit in "Syllabic input group 1" that corresponds to the first syllable of the incoming word form is set to an activation level of 1. Units representing the second, third etc. syllables are set to decreasing, non-zero values. This weighting procedure is introduced as a way of representing the temporal sequence of syllables: in this way, the map is able to tell apart word forms containing permuted (but otherwise identical) syllable sequences or subsequences. Due to the weighting process, trailing syllables in long word forms can hardly influence the resulting map representation, which prevents the system from using information present in word endings. To overcome this limitation, a second map is introduced (driven by "Syllabic input group 2") with a weighting function favoring trailing syllables and ignoring leading ones.

When tested on all examples that the ANN had been exposed to in the training phase, FR and FE recall reached 98%, and the precision was around 98%, too. These values show that the network makes hardly any mistakes when exposed to familiar input.

In the next step, *I kept only 5% of the training examples* (1068 examples that were randomly selected; they contained 235 word types and 17173 word tokens, type/token=1%), and the remaining examples (20542 sentence sequences) were set aside for testing. The result was a hardly noticeable decrease in the recall figure, but precision remained very high (above 99%).

In the next experiment, I examined whether the *homophones* $crack_1$ ("to get into a computer system illegally"), $crack_2$ ("excellent", "fantastic") and $crack_3$ ("illegal drug") cause problems in the frame/frame element recognition process. Both the statistical data and a visual assessment of activation levels of the output show that homophony did not cause problems in the context of this experiment.

In the remaining three experiments, my aim was to test how the system handled never-before-seen sequences. In the *first experiment*, I selected two text templates (each containing

two sentences) and checked whether the sentences were annotated independently of one another or were handled as sentence sequences.

In the *second experiment*, I removed a non-terminal symbol from a sentence sequence template. Change-of-performance data were collected.

In the *third experiment*, I was interested in the change of performance caused by the *replacement* of words or expressions by other forms that were originally associated with different FR/FE targets or no target at all. I worked with a single template belonging to the *computer crime* situation type. I replaced a non-terminal symbol (corresponding to a NP containing a noun head only) in the template by non-terminal symbols that were rewritten to:

– unknown words: twinmaps that were not used at all during training the FR/FE recognizer network, e.g. *ablaut, ablauts, wash, washing, xerox, xeroxing, ogled,*
– known and "situation-friendly" words and expressions, e.g. *protect, fixing, secured, spoof address, patching, protected,*
– known but "situation-external" words and expressions, e.g. *cocaine, heroin, mescaline, trading, produce, obtain, buy opium,*
– the plurals of the original forms (i.e. *computer* became *computers*, *program* became *programs*, and so on). It is important to point out that two templates (different from the template being manipulated here) had been used to connect these singular and plural forms during training; they generated parallel texts differing only in a noun (singular vs. plural) and since the target labels were the same, I expected the network to establish some kind of connection between the two sets.

4. Results and conclusions

When I used only 5% of the otherwise very dense training/testing corpus for training, the remaining 95% of the corpus was annotated reliably by the network, which means that it was able to develop the right generalizations and to store (memorize) them.

I found that the network had developed inter-sentential connections, too.

My neural network construction introduced in the thesis makes hardly any mistakes in its native task when exposed to familiar input, and is also able to compensate for missing or distorted (i.e. noisy) input. In an experiment, it perfectly compensated for missing sentences, while missing constituents seem to have caused more trouble, but precision remained above 93%. The explanation is that the network has enough contextual clues to make up for the loss of one constituent (at least when it is stored as one event, as in the above examples). These clues include function and content words, as well as multi-word expressions. All of them are

treated as equally important sources of information: it is in fact up to the network to find those pieces of information that are useful in the process.

The constituent replacement experiment was also handled successfully, and various levels of unexpectedness were reflected well in the recall and precision figures. Replacing *computer* by *computers,* or *server* by *servers*, etc. had a negligible effect only: the ANN had indeed established some sort of connection between the two sets of words. The use of completely unknown input caused the most trouble.

Finally, let me highlight some major features of the model. The network used for the simulations works with unannotated input consisting of actual (transcribed) word forms fetched from the training corpus. No morphological or syntactic information is included to help the learning process. Ambiguous entries have also been used (e.g. the word “crack” in two meanings: to *crack* a computer and the *drug*), and no noticeable performance degradation has been found. It is also important that the system processes sentence sequences rather than individual sentences, which is implemented by an unusual combination of recurrent structures. The syllable-based “twinmap” input interface is another unique feature of the system, which seems to be able to accommodate a large number of input words and multi word expressions; and to my best knowledge, the task itself, FrameNet-style semantic parsing, is also a novelty in the connectionist literature.

References

Koskenniemi, K. (1983). *Two-level Morphology: A general computational model for word form recognition and production.* Helsinki: University of Helsinki, Department of General Linguistics.

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus based study. *Linguistics, 29,* 801-843.

Chen, J. N., & Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics, 24(1).*

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, D. A. (2000). *Meaning in language.* Oxford: Oxford University Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Halle, M. (1973). Prelogomena to a theory of word formation. *Lingusitic Inquiry, 4(1).*

Lieber, R. (1981). *On the organization of the lexicon*. Bloomington, IN: Indiana University Linguistics Club.

Lieber, R. (1992). *Deconstructing Morphology*. Chicago and London: The University of Chicago Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception. Part 1: An Account of Basic Findings. *Psychological Review, 88*, 375-407.

Mihalcea, R., & Moldovan, D. (2001). EZ.WordNet: Principles for automatic generation of a coarse grained WordNet. *Proceedings of FLAIRS 2001*, 454-459, Key West, FL.

Prószéky, G. (1994). Industrial Applications of Unification Morphology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 157-159, Stuttgart, Germany.

Prószéky, G. (1996). Morphological Analyzer as Syntactic Parser. In *Proceedings of COLING 1996, 16th International Conference on Computational Linguistics*, 1123-1126, Center for Sprogteknologi, Copenhagen, Denmark.

Seagull, A. (2000). *A Compaction of WordNet Senses for Evaluation of Word Sense Disambiguators. TR726.* Computer Science Department, University of Rochester.

Verspoor, C. M. (1997). *Contextually-Dependent Lexical Semantics.* PhD thesis. Edinburgh: The University of Edinburgh.

List of publications written by the author, conference presentations

Published articles

(2000). Szóhálózat. *Magyar Tudomány, 10,* 1235-1237.

(2002). Derived nouns in WordNet and the Question of Productivity. *Studies in Linguistics, 6,* 433-449.

(2003). Derived Nouns and Gerunds in WordNet. In M. Fabian (Ed.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 1.* 149-154. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

(2004). Polysemy and Homonymy in WordNet. In M. Fabian (Ed.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 2.* 28-36. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

(2005). Form and Meaning in a Recurrent Neural Network. In M. Fabian (Ed.), *Sučasni doslidženńa z inozemnoï filologiï. Volume 3.* 22-31. Uzhhorod: Department of Foreign languages, Uzhhorod National University.

Review

(2002). Lawler, J., & Dry, H. (Ed.) (1998). *Using Computers in Linguistics: A practical guide.* Studies in Linguistics, 6, 490-493.

Conference presentations

(2003). HUSSE VI, Debrecen: *WordNet and the Lexicon*

(2005). HUSSE VII, Veszprém: *Form and Meaning in an Artificial Neural Network.*