Lexical Co-occurrence: The Missing Link

CUCS-486-89

Frank Smadja

Department of Computer Science
Columbia University
New York, NY 10027
Smadja@cs.columbia.edu

February 1989

To appear in the Journal for Literary and Linguistic Computing.

This work was partly carried out at Bar Ilan university, Department of English, and partly at Columbia University supported by DARPA under contract #N00039-84-C-0165.

CUCS-486-89

Lexical Co-occurrence: The Missing Link[*]

Frank A. Smadja[†]
Department of Computer Science
Columbia University
New York, NY 10027

February 2, 1989

1 Introduction

Aside from syntax, linguistic knowledge can be separated into two distinct parts, encyclopedic knowledge and dictionary knowledge. Encyclopedic knowledge describes the world whereas the dictionary describes individual word features, thus capturing lexical knowledge. Among the various types of lexical knowledge, one has generally been overlooked and should bring new results in computational linguistics: co-occurrence knowledge. Co-occurrence knowledge stands for the extent to which an item is specified by its environment independently of syntactic or semantic reasons. The basic concept is that of a *lexical relation* due to Saussure [49]. A lexical relation between two units of language stands for a correlation of common appearance of the two units in the utterances of the language. Consider the following two example sentences:[1]

(1) *"The ambassador of Freedonia delivered a strong protest concerning the violation of his country's sovereignty."*

(2) * *"The ambassador of Freedonia gave a heavy protest concerning the violation of his country's sovereignty."*

In the first sentence, if *deliver, strong* and *protest* co-occur, it is not only due to the fact that they have compatible semantic features and that " ... *delivered a strong protest* ..." is grammatical. This is exemplified by the fact that, in the second sentence, "... *gave a heavy protest* ..." is more than awkward though grammatical. The difference in well-formedness of these sentences is instead dependent on the lexical level[2]. *Deliver, protest* and *strong* are bound by

[*] To appear in The Journal of the Association for Literary and Linguistic Computing.

[†] This work was partly carried out while visiting at Bar-Ilan University, Department of English, and partly at Columbia University supported by NSF and DARPA grants.

[1] The * indicates an awkward sentence.

[2] This has also been termed "lexicalness" by Halliday [66].

lexical relations. Such relations describe lexical co-occurrence knowledge: they embody knowledge necessary for the proper usage of words; and they represent the extent to which an item is specified by its collocational environment. This kind of collocational information is physically reflected in the sentences of a language, making co-occurrence knowledge an observable phenomena.

In this paper, our focus is the automatic acquisition of co-occurrence knowledge. We show that a significant stride can be made in taking lexical co-occurrence knowledge into account in computational linguistic works in general and in language generation in particular. We describe a program, EXTRACT, that retrieves co-occurrence knowledge from the analysis of large textual corpora. EXTRACT can be seen as a co-occurrence compiler, producing lexical relations from texts. We then show how the retrieved knowledge can be used by a language generator.

2 Motivation and Background

Many wording choices in English sentences cannot be accounted for on semantic or syntactic grounds; they can only be expressed in terms of relations between words that usually occur together. That is, given a certain meaning to be conveyed, the choice of one word to express part of the meaning may entail selection of a second word. These two words must co-occur in the same sentence in a given context of meaning and they are part of a lexical relation. Co-occurrence knowledge has often been overlooked in the past, but should be included in computational dictionaries as it is an inherent part of the language. Outside of computational linguistics, the importance of co-occurrence knowledge has been previously recognized in the fields of psycholinguistics, linguistics and lexicography.

In **psycholinguistics**, the role of co-occurrence knowledge has been shown to be of importance in the framework of language learning [Leed 79]. Language learners often stumble across co-occurrence relations. Having a standard dictionary of English and a good knowledge of English grammar, a language learner whose first language is, for example, Hebrew or French will face the following problems while trying to use the words *dream* or *attention*:

Instead of saying: *I pay attention to* ..., if the first language is French, the learner would say: *I make attention at* ... [“*Je fais attention* à...”]. Similarly, if the first language is Hebrew, instead of *I had a dream* ... s/he would say: *I dreamt a dream* ... [“חלמתי חלום...”]. Such examples of production by language learners are numerous and account for a significant part of second language learners errors.

In **Lexicography**, following the work of Hornby [Hornby 42], more and more dictionaries account for lexical co-occurrence [Cowie 81], [Mel'čuk 81], [Benson 85]. In these dictionary, entries are not limited to the syntactic and semantic definition of lexemes, but also contain collocational knowledge in an orderly fashion. Benson and his colleagues in the BBI combinatory dictionary, [Benson 86] bases their work on a model of lexical relations inspired from Mel'čuk [81]. The BBI combinatory dictionary [Benson 86] focuses mainly on collocational information and it is the most complete account of cooccurrence knowledge in English up to date.

Following the early incentive of Saussure, in linguistics, several attempts have been made to model co-occurrence knowledge. Mel'čuk's model [Mel'čuk 81] is integrated as part of a complete linguistic model and is based on the notion of lexical functions (LFs). We present this model in more detail as we use it in this paper.

2.1 Mel'čuk's model of Co-occurrence Knowledge

At the crossroad of linguistics and lexicography, the notion of LF helps formalize the space of possible lexical relations. Each LF stands for an abstracted lexical relation into which words can enter. We give below three examples of primitive LFs derived from [Mel'čuk 73]:[3]

[Oper] associates verbs to a given noun < *noun* >. The noun and any of its associated verbs are frequently used together and enter into a certain structural relation. An element of *Oper*(< *noun* >) is a verb that takes < *noun* > as its direct object. The verb is taken here as a syntactic device, operating on < *noun* >. Common examples are:
Oper(attention) = [pay ...]
Oper(suicide) = [commit]
Oper(subpoena) = [serve, issue ...]
Oper(hint) = [give ...]
Oper[(protest) = [file, deliver ...]
Oper(bath) = [take ...]

[Labor] associates verbs to a given noun < *noun* >. The noun and any of its associated verbs are frequently used together and enter into a certain structural relation. An element of *Labor*(< *noun* >) is a verb, which takes < *noun* > as its indirect object.
Labor*(esteem) = hold [hold someone into esteem]*
Labor*(consideration) = take [take something into consideration]*
Labor*(arrest) = place [place somebody under arrest]*

[Magn] associates adjectives (or adverbs) to a noun < *noun* >. The role of an element of Magn(< *noun* >), is to emphasize, magnify or stress the meaning of < *noun* > when used in combination with it. Common examples are:
Magn*(escape) = [narrow, ...]*
Magn*(car) = [powerful, ...]*
Magn*(sound) = [loud, ...]*
Magn*(tea) = [strong, ...]*
Magn*(corpus) = [large, ...]*
Magn*(argument) = [tough, ...]*
Magn*(protest) = [strong, vigorous, courageous, ...]*

LFs capture a very important aspect of lexical knowledge. Abstracting and classifying lexical relations into which words can enter is necessary for computational linguistics as well as for lexicography. However, directly using the LFs is not desirable for our purpose. As they stand, lexical relations do not really correspond to well defined semantic predicates or primitives although they are based on semantic criteria. LFs are hard to interpret formally, they lack necessary semantic definitions and a much finer granularity is needed. In our work on natural language generation, we use a simpler model based on the notion of interpreted lexical relation [Smadja 89]. However for purposes of clarity, in this paper, we adopt Mel'čuk terminology and demonstrate how to compile a computational dictionary based upon his model.

[3] The LFs given here are simplified versions of the originals given in [Mel'čuk 81].

Let us first explain how such a computational dictionary would be useful for language generation purposes.

3 Co-occurrence knowledge and Natural Language Generation

The output of a natural language understander is a conceptual structure that is intended to represent the meaning of the read text whereas the output of a natural language generator is text in natural language. This implies that no linguistic element can be omitted when generating. Sentences like sentence (1) in which two words share a lexical relation are more of a problem for language generators than for parsers, simply because parsers are provided with the correct words whereas a generator has to find them. We call such sentences collocotionally restricted sentences. Language generators generally ignore co-occurrence constraints and are thus unable to correctly generate such sentences. Consider the simple task of generating sentence (1), for example. Computational dictionaries used by language generators generally do not contain the lexical relation *protest-deliver* which forces language generators to either consider the phrase *"to deliver a protest"* as an idiom or require that all lexical items (*"protest"* as well as *"deliver"*) be present in the input for the surface generator. Both cases are not desirable in that none takes advantage of the special relation linking *"deliver"* and *"protest"*. In contrast, we argue for the use of co-occurrence knowledge in natural language generation. We claim that collocationally restricted sentences can be correctly handled, and that less specified input structures are needed. For example, for a simple task such as generating sentence (1), the input to the surface generator could have the following form:

Oper(Magn(protest))[ambassador of Freedonia, violation of Freedonia sovereignty].[4]

The LFs act as semantic predicates triggering at the surface level the lexical relations involved. In this example, *Oper[protest]* triggers the lexical relation *protest-deliver* and *Magn* triggers the lexical relation *protest-strong*. In reality, knowing that each LF has several values for each word, the following sentences could have been produced as well:

(3) *"The ambassador of Freedonia exploded in a vigorous protest concerning the violation of his country's sovereignty."*

(4) *"The ambassador of Freedonia rose in an energetic protest concerning the violation of his country's sovereignty."*

Distinguishing among those four sentences when generating requires consideration of other constraints such as pragmatics, politeness, discursive, etc. Incorporating lexical co-occurrence knowledge in the process of generating adds complexity to the generation process. The main problem introduced by co-occurrence knowledge is the problem of interaction between co-occurrence knowledge and other kinds of knowledge in the process of generating. We are currently in the process of developing a generating scheme that will handle these interactions.

[4] We have not developed the noun phrases for clarity purposes.

To another extent, some attempts are currently being made at producing sentences using computational dictionaries with co-occurrence knowledge. Both are based on Mel'čuk 's full fledged linguistic model. Iordanskaja *et. al.*, [Iordanskaja 88], are implementing a computational model of the complete theory. The intent is to validate Mel'čuk's model. Nirenburg *et. al.* [Nirenburg 88] also use LFs in the framework of their work on machine translation. In their work, LFs are used to help in the process of lexical selection. Mel'čuk based approaches present two major drawbacks. First, it is difficult to use a part of Mel'čuk's model in isolation to the full fledged linguistic model, yet that is just what these researchers do, as a full computational model at this point would be quite difficult to implement. Second, these computational researchers incorporate LFs by hand into the computational dictionary. Such a process is both tedious and incomplete. The work that we present here attempts to solve these problems.

4 The Acquisition Method

Acquiring co-occurrence knowledge requires the study of numerous example sentences. Language learners need long hours of exposure to the language before mastering such features. Similarly, lexicographers willing to compile dictionaries accounting for co-occurrence knowledge spend a lot of time studying example sentences. Lexicographers now have to study large samples of English texts of all natures in order to extract and compile co-occurrence knowledge. This particularly overwhelming task is often carried out through the careful study of numerous texts, other dictionaries, the linguistic competence of the lexicographer and other persons' linguistic judgments, [Cowie 81]. In an attempt to relieve lexicographers from the burden of collecting and classifying occurrences, Choueka in [Choueka 83], has proposed algorithms that allow retrieving frequent idiomatic and collocational expressions from the scanning of large textual corpora (millions of words). Although more interested in the retrieval of commonly used expressions such as: *United Nations, Middle East, home run, President Reagan, etc.*, his work describes an interesting methodology for handling large corpora and can be considered as a first step toward automated lexicography. Our work also takes this approach. We also need to investigate large corpora of English texts in order to retrieve co-occurrence knowledge. However, we are more interested in computational dictionaries than dictionaries in general.

Acquiring co-occurrence knowledge from a large corpus actually encompasses two different activities: the retrieval of the raw information, and the compilation in the computational dictionary. The raw information represents simple co-occurrences and the interpreted information will entered in the computational dictionary. In this paper, we focus on the extraction activity and show how co-occurrence knowledge can be identified from the analysis of a large corpus. To illustrate our approach, we sketch hereafter the acquisition of co-occurrence knowledge for the word "*protest*".

4.1 The retrieval algorithm, EXTRACT.

Ideally, lexical relations are extracted from a text by parsing it. Two words are involved in a lexical relation, if they belong to the same syntactic constituent, *e.g.*, noun-phrase, verb-phrase, etc. However, in real life, free-style texts contain many non-standard features over which automatic parsers would stumble. Since it has been shown that 98% of lexical relations relate words separated by at most five words, [Martin 83], we use this fact to avoid parsing. In other terms, most of the lexical relations involving a word w can be retrieved by examining the neighborhood of w, within a span of five. Figure 1 illustrates a snapshot of the scanning process for sentence (1). The current word under consideration is *"delivered"* and the selected words are *"strong"*, *"protest"* and *"concerning"*. Let us note that for each word, we scan only the five following words and not the preceding ones in order to avoid counting the same lexical relation twice.

The extracting algorithm used by our program takes as input a corpus, a span parameter (five) and a dictionary specifying closed-class words.[5] It produces a list of tuples (w_1, w_2, f), where (w_1, w_2) is a lexical relation between two open-class words identified in the corpus, and f is the frequency of appearance observed. The algorithm consists of the following four steps for each lexical entry, w:

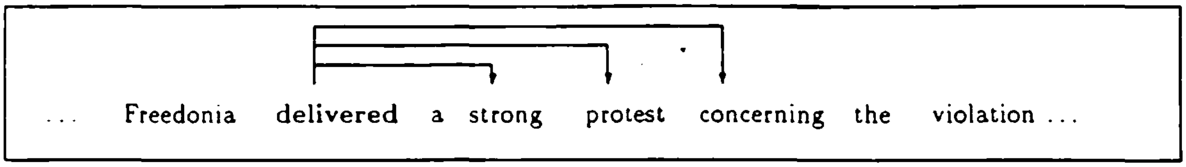
... Freedonia delivered a strong protest concerning the violation ...

Figure 1: Scanning sentence (1) with the word *delivered*

Scan: Scan the whole text for each appearance of w.

Compile: For each sentence containing w, make a note of its collocates[6]. Most of w collocates are retrieved by examining its environment five words after. All collocates are stored along with their syntactic category and their frequency of appearance.

Lemmatize: A basic morphological analysis of every word involved in a lexical relation is performed. The morphological analysis uses the UNIX *spell* program. Each word is mapped into its morphological root[7] using simple inflectional transformations.

Filter: The statistical distribution of the collocates of w is analyzed and the peaks are automatically selected. A peak is defined as a collocate of w whose frequency of appearance is above $\bar{f} + k\sigma$; where $\bar{f}$ is the average frequency of appearance, σ the standard deviation of

[5] Closed class words refer to small syntactic categories, such as articles, preposition etc. In contrast, open class words are nouns, adjectives and adverbs and are therefore much more numerous. Closed class words are somehow reachable by grammar rules whereas open class words are dealt with in the lexicon [Huddleston 85].

[6] By collocate, we mean the nearby open-class lexical items.

[7] Let us note that we do not extract the absolute morphological stem of any word but that we are only interested in the inflectional stems.

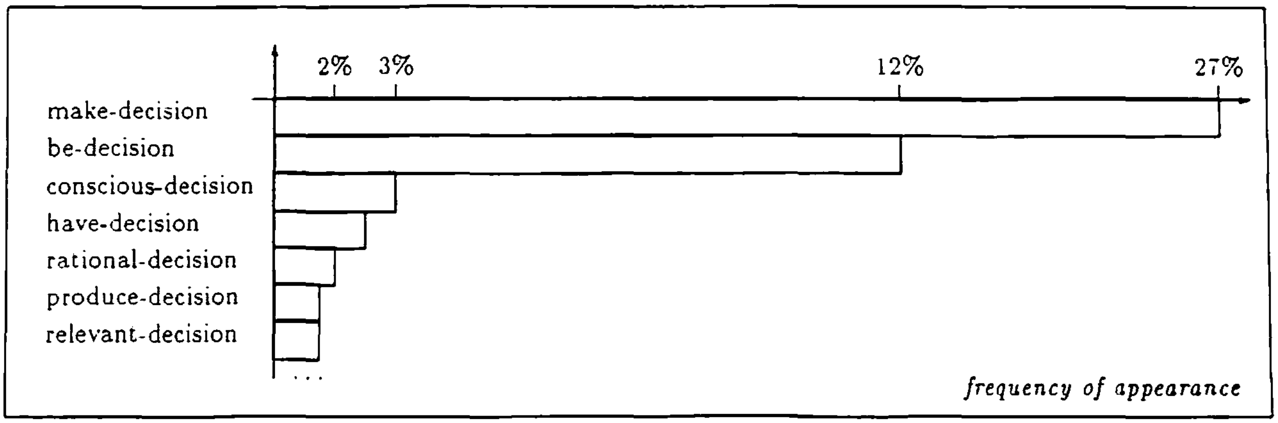

Figure 2: An example histogram for the word *decision*

the distribution, and k a factor that has to be empirically determined according to the size and nature of the corpus. Let us call k, the co-occurrence factor.

At this point, insignificant information, *i.e.* atypical word juxtapositions, have been filtered out, and what remains is an ordered set of words each sharing a lexical relation with w. Each word appearing on the collocate list of w is bound to it by a lexical relation. For example, if EXTRACT is run on the word *"protest"* on a sufficiently rich corpus, the high-frequency adjectival collocates retrieved are: *indignant, feeble, ineffective, earnest, passionate, respectful, strong.* And similarly the high frequency verbal collocates are: *issue, growl in, anticipate, dismiss, give rise to, explode in, deliver, maintain.*

The algorithm has been fully implemented and tested on a 300,000 word corpus taken from the UNIX Usenet, see next section. It is currently being tested on a more than 2,500,000 words corpus taken from the archives of the daily Israeli newspaper, **The Jerusalem Post**.

This retrieved information could already be used for lexicographic purposes, as it provides a combinatorial description of English. However, as we will see in section 4.3, before being compiled in computational dictionaries, this knowledge has to be further refined. We first give some results obtained with **EXTRACT**.

4.2 Some Results

For the retrieved knowledge to be valuable, several hundred occurrences of the same word must be examined, which implies that a very large corpus must be used. In spite of the small size of the corpus, we have been able to make useful lexicographic observations from the 300,000 word corpus taken from the UNIX news net. The program has been tested on 50 nouns that appear more than two hundred times in the corpus. The results can be represented by histograms, where each line stands for a given association between two words and the length of the line represents the frequency of common appearance of the two words. Figure 2 represents the histogram for the word *decision. Decision* appears 330 times in the whole corpus, and *to make* co-appears once every four appearances. Some histograms have very marked peaks whereas others are almost flat. Each peak represents a lexical relation and around 200 such peaks have been automatically retrieved.

Among the strongest peaks noticed were: *to make* with *decision*, *"John makes a decision"*, see Figure 2. *To send* with *mail*, *"John sends mail to Mary"*. *To take* with *note*, *"Mary takes notes of ..."*. *To answer* with *question*, *"John answers Mary's questions"*. *Send* with *request*, *"John sends a request"*. And *take* with *approach*, *"Mary takes a new approach"*. All these pairs of words have a co-occurrence factor above 2.5, *i.e.*, the observed frequency of common appearance of the two words is 2.5 σ above the expected one. Some peaks do not stand for cooccurrence knowledge but rather for domain specific or paradigmatic knowledge, some examples are: *mail* with *Arpanet* and *human* with *machine* or with *behavior*. Finally, some words were almost non productive, their histograms very flat, among them: *abstract*, *definition* and *information*. The reason for this is due both to the corpus and to the nature of the words.

The Jerusalem Post corpus consists of several thousand articles that have been published recently in the newspaper. EXTRACT is currently being tested on it, and the results already obtained are of much better quality than for the Usenet corpus. EXTRACT has for the moment only been run on a few hundreds words appearing more than 300 times in **The Jerusalem Post** corpus. The difference in quality of the results is easily explained, first the corpus is six time bigger, and second the articles cover a wide range of topics, from sports to politics and science. We can already predict that more than 1000 words will be productive, and that the peaks will be much sharper than for the smaller corpus. In the future, we plan to investigate the use of EXTRACT on different specialized corpora, and thus retrieve domain dependent information.

4.3 Linking Lexical Relations to Semantics

To illustrate our approach, we present here the task of analyzing previously retrieved lexical relations using the formalism of LFs. As mentioned above, this formalism is not the one we use in our generation work but is simply used here for presentation purposes. In this simplified context, what we mean by *interpreting* previously retrieved lexical relations, is to map them to the values of LFs for the given word. Lexis is more item-bound than grammar in the sense that more abstraction is involved in grammar than is possible in lexis. Each LF requires a different learning strategy, for *Magn* the interpretation method is the following.

Magn maps nouns to adjectives. For the noun *protest*, the set to be considered for analysis is: *indignant*, *feeble*, *ineffective*, *earnest*, *passionate*, *respectful*, *strong*, as determined by EXTRACT (see previous section). *Magn*(*protest*) is a subset of it. The role of *Magn* is to stress, emphasize or intensify the meaning of *protest*. In order to determine *Magn*(*protest*), we need to select from the candidate adjectives the one which bear this semantic trait. Those selected will participate in a *Magn* construction and *Magn*(*protest*) will be exactly this set of adjectives. In the considered case, if EXTRACT is run on a sufficiently rich corpus, *indignant*, *vigorous*, *vociferous*, *energetic* and *strong* would be selected.

Other LFs require similar strategies, for example, for *Oper*, from the the candidate set: *issue, growl in, anticipate, dismiss, give rise to, explode in, deliver, maintain*, the following are selected: *issue, deliver* and *give rise to.* We are currently working on developing a method for interpreting lexical relations in an orderly way in restricted domains. The method should be partly automatable for restricted domains if a thesaurus is used.

5 Conclusion

In this paper, we have shown that co-occurrence knowledge could be brought to bear in computational linguistics. While syntax and semantics are usually taken into account, one level is missing: the lexical level. We demonstrated that the use of computational dictionaries accounting for this level could help language generators correctly handle collocotionally restricted sentences.

To incorporate co-occurrence knowledge into natural language generation there is a need for computational dictionaries containing this knowledge. The work we presented here attempts to solve this problem through EXTRACT, a co-occurrence compiler that produces lexical relations from the statistical analysis of a large corpus. Our extracting algorithm is based on the basic definition of lexical relations and uses a statistical threshold in order to filter out non relevant information. This algorithm is fully implemented and has already produced interesting results.

We are currently working on developing a systematic method of semantic interpretation that would allow entering lexical relations into computational dictionaries. In parallel we are developing a generation scheme that would make explicit use of the above computational dictionary. The acquisition of lexical co-occurrence knowledge is an important task for second language learners and also for lexicographers. For a computer program, since currently no computational dictionary accounts for it. Learning co-occurrence knowledge not only constitutes a challenging task but also provides useful and practical results.

Acknowledgments

Doree Seligman, Tom Ellman and Joel Walters read earlier versions of this paper and made useful criticisms and suggestions and the comments of Kathy McKeown were indispensable in the research as well as in the writing. We are grateful to The Jerusalem Post for graciously giving us access to their archives. I would also like to thank Yaacov Choueka and Joel Walters at Bar Ilan University, Ramat Gan, Israel, for providing me with a rich and pleasant research environment during my stay there.

References:

[Benson 86] M. Benson, E. Benson, R. Ilson. *The BBI combinatory dictionary of English: A guide to word combinations.* Amsterdam: John Benjamin, 1986.

[Berwick 85] R.C. Berwick, *The Acquisition of Syntactic Knowledge.* MIT Press, 1985, Cambridge, MA.

[Benson 86] M. Benson, E. Benson and R. Ilson, *Lexicographic Description of English.* John Benjamins Publishing Company, Philadelphia, 1986.

[Choueka 83] Y. Choueka, T. Klein and E. Newitz, *Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus.* ALLC journal, vol. 4, 1983, pp. 34-38.

[Cowie 81] A.P. Cowie, *The Treatment of Collocations and Idioms in Learner's dictionaries.* Applied Linguistics, 2-3 (1981), pp. 223-235.

[Cruse 86] D.A. Cruse, *Lexical Semantics.* Cambridge University Press, 1986.

[Halliday 66] M.A.K. Halliday, *Lexis as a Linguistic Level.* In C,E. Bazell, J.C. Catford, M.A.K Halliday and R.H. Robins (eds.), *In memory of J.R. Firth* London: Longmans Linguistics Library, 1966. p: 148-162.

[Hornby 42] A.S. Hornby, E.V. Gatenby & H. Wakefield, *Idiomatic and Syntactic English Dictionary.* Kaitakusha. Tokyo, Japan, 1942.

[Huddleston 84] R. Huddleston, *Introduction to the Grammar of English.* Cambridge Textbooks in Linguistics, Cambridge University Press, 1984.

[Iordanskaja 88] L. Iordanskaja, R. Kittredge, A. Polguere, *Lexical Selection and Paraphrase in a Meaning-Text Generation Model.* Presented at the fourth International Workshop on Language Generation, Catalina Island, CA, 1988.

[Leed 79] R.L. Leed & A.D. Nakhimovsky, *Lexical Functions and Language Learning.* Slavic and East European Journal, Vol 23-1, 1979.

[Martin 83] W.J.R. Martin, B.P.F. Al and P.J.G van Sterkenburg, *On the processing of a text corpus: from textual data to lexicographical information.* Lexicography: Principles and Practice, Ed. R.R.K Hartmann, Applied Language Studies Series, Academic Press, London, 1983.

[Mel'čuk 73] I.A. Mel'čuk *Lexical Functions in Lexicographic Description.* Proceedings of the Berkeley Linguistics Society, 8, 1973.

[Mel'čuk 81] I.A Mel'čuk, *Meaning-Text Models: a Recent Trend in Soviet Linguistics.* The annual review of anthropology, 1981.

[Nirenburg 88] S. Nirenburg *et. al.*, *Lexicon building in natural language processing.* Fifteenth International Conference of the Association for Literary and Linguistic Computing, Jerusalem, Israel, June 1988.

[Saussure 49] F. De Saussure, *Cours de Linguistique Generale, 4eme edition.* Librairie Payot, 1949, Paris, France.

[Smadja 89] F. Smadja, *Dictionaries for Language Generation Accounting for Co-occurrence knowledge* Submitted to IJCAI, 1989.