Application of Corpora in Second Language Learning
— The Problem of Collocational Knowledge Acquisition —

Kenji KITA †, *Takashi OMOTO* †, *Yoneo YANO* † *and Yasuhiko KATO* ‡

† Department of Information Science and Intelligent Systems
Faculty of Engineering
Tokushima University
Tokushima 770, JAPAN
e-mail: kita@is.tokushima-u.ac.jp

‡ Section for Dictionary Research
The National Language Research Institute
Kita-ku, Tokyo 115, JAPAN
e-mail: kateaux@tansei.cc.u-tokyo.ac.jp

Abstract

While corpus-based studies are now becoming a new methodology in natural language processing, second language learning offers one interesting potential application. In this paper, we are primarily concerned with the acquisition of collocational knowledge from corpora for use in language learning. First we discuss the importance of collocational knowledge in second language learning, and then take up two measures, mutual information and cost criteria, for automatically identifying or extracting collocations from corpora. Comparative experiments are made between the two measures using both Japanese and English corpora. In our experiments, the cost criteria measure proved more effective in extracting interesting collocations such as fundamental idiomatic expressions and phrases.

1 Introduction

Recent rapid advances in computer technology (particularly the advent of large storage devices and parallel computers) and numerous data collection efforts have caused a shift in natural language applications from a knowledge-based to a corpus-based or data-intensive approach. The knowledge-based approach focused on abstraction of language, describing linguistic phenomena through minimal core knowledge such as parts-of-speech, syntactic and semantic rules. Linguistic phenomena, however, vary so vastly that they cannot be described through core knowledge. In addition, hand-coding knowledge takes a lot of time and hard work. The knowledge-based approach, therefore, has been found wanting in developing large-scale practical NLP systems.

On the other hand, the corpus-based approach makes no claim about the compactness of the knowledge. Rather, the corpus-based approach derives more power from massive quantities of textual data than from hand-coded knowledge, being able to compensate for the weakness of the knowledge-based approach through authentic examples and various statistics of language use. With the availability of large corpora in recent years, many successful results have been derived from corpus-based studies. These include part-of-speech tagging [Kupiec 1992], parsing [Magerman and Marcus 1990], example-based machine translation[Sumita and Iida 1992], statistical machine translation [Brown et al. 1990, Brown et al. 1993], language modeling [Jelinek 1990, Kita 1992] and many other related areas.

One interesting potential use of corpora is for second language learning. Kita et al. [Kita et al. 1993b] discussed various way of using corpora in language learning. The greatest advantage of using corpora in language learning is that the corpora provide a body of evidence for the function and usage of words and expressions. At the same time, deriving lexical knowledge from large-scale corpora via automated procedures, as well as its use in language learning CAI systems, is one of the most important issues.

In this paper, we are primarily concerned with the acquisition of collocational knowledge from corpora. The organization of this paper is as follows. Section 2 gives an overview of corpus-based CALL (Computer-Assisted Language Learning). In Section 3, we describe why collocational knowledge is important in second language learning. In Section 4, we discuss the automatic extraction of collocations, taking up two measures, mutual information and cost criteria, for identifying or extracting collocations from corpora. In Section 5, we describe comparative experiments in extracting collocations and discuss the two measures.

2 The Use of Corpora in Second Language Learning

There have been many language learning systems developed so far. Of course, the goal of creating language learning systems is to have learners master practical language skills. In spite of efforts by many researchers, we are still quite far from this goal although we admit that partial success has been achieved. Why? First, language learning systems developed so far are too domain-limited, i.e. they operate only on a restricted subject matter or purpose and accept sentences only of a limited or restricted nature. Second, researchers paid attention to knowledge representation models themselves rather than to the knowledge to be entered. In consequence, systems lack wide coverage and robustness, being often called "toy systems".

Corpus-based CALL offers great possibilities in building practical language learning systems. Some topics of corpus-based CALL includes:

- **Linguistic knowledge acquisition from corpora.**

  Language learning systems must incorporate many kinds of linguistic knowledge. Usually, the linguistic knowledge is hand-coded by humans. The resulting knowledge, however, sometimes does not match real-world usage. Also, hand-coding knowledge takes a lot of time and hard work. The current availability of large computer-readable corpora presents the possibility of deriving knowledge via automated procedures. Incorporating the derived knowledge makes language learning systems quite useful for dealing with unrestricted texts, making systems eminently robust.

- **Enhancement of translation skills through bilingual corpora.**

  Bilingual corpora consist of parallel texts which content is essentially equivalent. Thus, they have potential possibilities in enhancing learners' translation skills.

- **Enhancement of oral/aural skills through speech corpora.**

  There are corpora in which actual speech data are encoded. Speech corpora can be used for enhancing listening/speaking abilities. In particular, for full development of aural comprehension ability, speaker-dependent practice (using speech from one speaker) does not suffice; it is necessary to provide learners with extensive listening practice using speeches from many speakers. For that purpose, speech corpora is indispensable for language learning systems.

- **Multimedia language learning through structured corpora.**

  Learning environment with multimedia aids is one of the recent increasing interests. As stated above, corpus-based language learning enables us to use not only texts but also speech material. Moreover, a corpus encoded with SGML can be used to link words of a text to images of its objects. Thus, multimedia language learning which integrates texts, speech and images would be possible.

- **Retrieving examples from corpora.**

  Learners often want to know how a word is used within a sentence. Although dictionaries are fine for that purpose, they include only typical examples. A corpus includes quantitatively a sufficiently large amount of examples of a language, providing a more extensive usage of words. In addition, various computational tools have been developed for retrieving examples from corpora.

- **Augmenting incomplete knowledge with many examples.**

  Language learning systems are required to accept sentences from learners, where input sentences are judged correct or not through the process of parsing. To do this, a traditional approach uses phrase-structure grammars, sometimes augmented by sematic information. However, a grammatical approach often does not work out because grammars inherently contain problems such as the *overgeneration problem*, the *undergeneration problem*, and the *ambiguity problem*. A corpus includes many examples, being able to compensate the incompleteness of a grammar through actual examples and statistical data.

3 Importance of Collocational Knowledge in Language Learning

There has been much theoretical and applied research on collocations, both from a linguistic and an engineering point of view. Consequently, the definition of collocation differs according to the researcher's interest and standpoint. This paper adopts the most comprehensive definition: a collocation is a cohesive word cluster, including idioms, frozen expressions and compound words.

The importance of collocations has been stressed in an extensive literature. From a language learning viewpoint, it can be summarized as follows:

- In language learning, learners must pay attention to how words are used rather than to individual words by themselves. Collocational knowledge indicates which words co-occur frequently with other words and how they combine within a sentence. Therefore, collocational knowledge is especially effective in sentence generation [Smadja and McKeown 1990, Smadja 1993].

- Collocational knowledge is very difficult to acquire for second language learners. A typical example is the pair of words "strong" and "powerful" [Church et al. 1991, Smadja 1991]. These two words have similar meanings, but their usage is quite different. For example, native English speakers prefer saying "strong tea" to "powerful tea", and prefer saying "powerful car" to "strong car". For non-natives, however, it is difficult to catch the subtle distinctions between these two words. These lexical preferences were sometimes ignored in the traditional knowledge-based approach; nevertheless they are the most important source for word choice and word ordering.

- It is pointed out that human translation process is based on analogical thinking [Nagao 1984]. First, a human translator properly decomposes a sentence into certain fragmental phrases,

then s/he translates each fragmental phrase by analogy with other examples, and finally composes fragmental translations into one sentence. Collocations are suitable for fragmental translation units.

- From a cognitive point of view, it is said that human language acquisition is governed by the law of maximal efficiency [Wolff 1991]. In other words, data compression, often called chunking, is performed to minimize storage demands in the brain. A chunk is considered to be a pattern which repeatedly appears in a variety of contexts. Collocations are good candidates for chunk units.

4 Extracting Collocations from Corpora

In the past, several approaches have been proposed to extract collocations from corpora. Church et al. [Church and Hanks 1990, Church et al. 1991] introduced the association ratio, which indicates how strongly two words are related, based on the information-theoretic concept of mutual information. Smadja et al. [Smadja and McKeown 1990, Smadja 1991, Smadja 1993] take into account word distance as well as word strength for a measure of word association. Also, Basili et al. [Basili et al. 1992] proposed a syntax-based approach. Particularly, mutual information plays a central role in recent lexical statistical research. To take a few examples, Hindle and Rooth [Hindle and Rooth 1993] applied mutual information to disambiguate prepositional phrase attachments, and Brown et al. [Brown et al. 1992] used it in determining word classes.

In this section, after surveying how mutual information can be used to extract collocational information, we introduce another measure, called *cost criteria* [Kita et al. 1993], to automatically extract interesting collocations from corpora. Comparative experiments and discussions will be described in the next section.

4.1 Mutual Information

The mutual information between two words x and y is defined as follows [Church and Hanks 1990, Church et al. 1991]:

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Here, $P(x)$ and $P(y)$ are word occurrence probabilities, and can be estimated from the number of occurrences of the words, $f(x)$ and $f(y)$, and the number of words in the corpus, N.

$$P(x) = \frac{f(x)}{N} \quad \text{and} \quad P(y) = \frac{f(y)}{N} \tag{2}$$

$P(x,y)$, the joint probability of x and y, is estimated in a similar way.

$$P(x,y) = \frac{f(x,y)}{N} \tag{3}$$

where $f(x,y)$ is the number of occurrences of x followed by y.

The mutual information $I(x,y)$ compares the probability of observing x and y together with the probabilities of observing x and y simply by chance. Thus, a large value indicates that the two words x and y have a strong relationship. By extracting word pairs with large mutual information values, we can obtain common collocations.

Because mutual information values are defined for two words, this simple method can only extract collocations of length two. However, a generalization is suggested in [Jelinek 1990] as follows:

1. Start out from the basic vocabulary V_0. Set $n = 0$.
2. Augment the vocabulary V_n by all word sequences "$x\ y$" for which $I(x,y) > Thr$, where Thr is a predetermined threshold.
3. From Step 2, a new vocabulary V_{n+1} is established.
4. Adjust the counts to reflect the new vocabulary V_{n+1}.
5. Resume from Step 1 with V_{n+1} as its basis.

With this iterative procedure, the final vocabulary includes collocations of arbitrary length.

4.2 Cost Criteria

The cost criteria measure is based on the assumptions that (1) collocations are recurrent word sequences, and (2) the recurrent property is captured by the absolute frequency of a word sequence. However, a simple absolute frequency approach does not work, because the frequency of a sub-sequence is always higher than that of the original word sequence. For example, because "in spite" is a sub-sequence of "in spite of", "in spite" appears more frequently than "in spite of". However, given the context "in spite", it is highly probable that "of" follows "in spite". Consequently, we must consider that "in spite of" is a collocation but "in spite" is not. The idea of cost criteria formalizes this, and it can quantitatively estimate the extent to which processing is reduced by considering a word sequence as one unit.

Before the presentation of a formal definition, we introduce the following notation:

$$\alpha \quad \cdots \quad \text{a word sequence.} \tag{4}$$

$$|\alpha| \quad \cdots \quad \text{the length of } \alpha. \tag{5}$$

(the number of words in α)

$$f(\alpha) \quad \cdots \quad \text{number of occurrences of } \alpha \text{ in a corpus.} \tag{6}$$

We define $K(\alpha)$, the cost reduction incurred by handling α as a unit:

$$K(\alpha) = (|\alpha| - 1) \times f(\alpha) \tag{7}$$

$K(\alpha)$ is interpreted as follows. Assume here that, in the corpus, there exists a word sequence α, which is composed of $|\alpha|$ words and occurs $f(\alpha)$ times. Also assume that the cost of processing one word is 1. Similarly, when processing α as a single unit, its processing cost is 1. If a word sequence is processed one word at a time, it is reasonable to assume that the processing cost is proportional to the length of the word sequence. That is, the processing cost for α is $|\alpha|$. By considering α as one unit, the processing cost is reduced by $|\alpha| - 1$. Since α appears $f(\alpha)$ times, we can conclude that the total cost reduction becomes $(|\alpha| - 1) \times f(\alpha)$, which is the definition of $K(\alpha)$.

In reality, however, the problem is not so simple, because word sequences are not mutually disjoint. Consider the case where a word sequence α is a sub-sequence of β (for example, $\alpha =$ "in spite", $\beta =$ "in spite of"). Then, we have:

$$f(\alpha) \geq f(\beta) \tag{8}$$

Further, the word sequence α, $f(\beta)$ times out of $f(\alpha)$ times, will be identified as β. Thus, the actual cost reduction for α is defined as:

$$K(\alpha) = (|\alpha| - 1) \times (f(\alpha) - f(\beta)) \tag{9}$$

Finally, we can extract collocations from a corpus by the following steps:

1. Calculate $K(\alpha)$ for each word sequence α in a corpus.
2. Rank a word sequence α by using the value $K(\alpha)$.
3. Extract higher rank word sequences as collocation candidates.
4. Re-calculate $K(\alpha)$ for each α in the collocation candidates.

5 Experiments and Discussions

5.1 The ADD Corpus

In our experiments, the ADD (ATR Dialogue Database) Corpus [Ehara et al. 1990] created by ATR Interpreting Telephony Research Laboratories in Japan was used. The ADD Corpus is a large structured database of dialogues collected from simulated telephone or keyboard conversations which are spontaneously spoken or typed in Japanese or English. This corpus consists of parallel texts of Japanese and English, aligned by utterance. Also, sentences in ADD are morphologically analyzed and annotated with various kinds of syntactic, semantic, and phonological information.

Currently, the ADD Corpus contains textual data from two tasks (text categories); one consists of simulated dialogues between a secretary and participants at international conferences (Conference Task), and the other of simulated dialogues between travel agents and customers (Travel Task).

In our experiments, we used the keyboard dialogues from the Travel Task, which include approximately 120,000 Japanese words and 100,000 English words. The telephone dialogue include linguistic phenomena, such as filled pauses ("ah", "uh", etc.), restarts (repeating a word or phrase) and interjections, so we did not use them for our experiments.

5.2 Results and Discussions

Figure 1 shows some interesting Japanese collocations extracted using respectively mutual information and cost criteria. Figure 2 shows some English ones.

Before discussing the results, we first overview the characteristics of Japanese phrases. In general, the order of major constituents in a Japanese sentence is rather free. However, predicate phrase positioning is dominated by the so-called predicate-phrase ending constraint: a predicate phrase appears at the end of its clause. Furthermore, a predicate phrase often has a complex form, consisting of a main predicate such as a verbal noun, verb or adverb, combinations of auxiliary predicates, and a sentence-final particle. These auxiliary predicates and sentence-final particles add various complementary meanings to a sentence, such as honorific, causative, and prohibitive meanings, etc.

As can be seen from the experimental results (Figure 1), the method based on mutual information tends to extract compound noun phrases, while cost criteria tends to extract complex predicate phrase patterns. Almost all the collocations extracted are in this category. For example,

Mutual Information	Cost Criteria
ichi ryuu no orchestra ni yoru ensou	desho u ka
Jouzankei-onsen to set ni nat ta golf-pack	desu ka
Kunitachi-shi Ishida	desho u
chijou e	mashi ta
night-tour ya dinner-show	sou desu
kaihatsu ga sakan	sou desu ka
buchou ya kachou	to iu koto
6 mai tsuzuri	sou desu ne
hizuke henkou sen wo koe	masu ka
moushikomi kin toshite o azukari	desu ne
Shinjuku-ku Naitou-chou 1 banchi	o negai shi masu
kokunai sen no daiya	itashi masu
hakkou kaisha ni teishutsu	o negai itashi masu
Kenya Tanzania Safari	to omoi masu
yuuran sen no senchou	te ori masu
kaisui yoku	tai no desu ga
yuukyuu kyuuka	wakari mashi ta
Matsushima-wan meguri	kashikomari mashi ta
resort kaihatsu	ni nari masu
umi to yama	to iu no ha
yuujin no hahaoya	shi tai no desu ga
Hachiman-daira Towada Hiraizumi	to iu koto de
danjo betsu no uchiwake	na n desu ga
hakubutsu kan	shi tai no desu
dou nenpai	shouchi itashi mashi ta
senmon yougo	to iu koto desu
yuukou kigen	sou na n desu
genkin kakitome	arigatou gozaimashi ta
Shanghai Sian	sa se te itadaki masu
Setagaya-ku Kyoudou	o mata se itashi mashi ta
seinen gappi	sou na n desu ka
moyori no eki	shitsurei itashi masu
choushoku to chuushoku	yoroshii desho u ka
Fuji-ginkou honten	ka mo shire mase n
gouka kyakusen	irasshai masu ka

Figure 1: Some examples of extracted collocations. (Japanese)

Mutual Information	Cost Criteria
yacht harbor	is that so
Echigo Yuzawa	thank you very much
Fifth Avenue	I would like to
General Affairs	I see
Mitsuboshi trading	my name is
slide projector	sorry to have kept you waiting
strong background	is that right
cross the International Date	in that case
the F1 Grand Prix	I understand
Shiretoko Sightseeing Boat Inc.	thank you for
it's my pleasure	do you have any
at the Hotel New Tanda	good bye
give a speech	would you like to
head Mr. Kuwata	I am very sorry
wine production	a little
Wall Street	be able to
jazz dance	I got it
my mother in law	I'll be waiting for your call
to the historic sites	may I have your name and address
I am not that familiar	how much
Keirin and Peking	all right
cause the inconvenience	as soon as possible
holding a paper	then would you give me your
baths and toilets	the other day
Las Vegas	make the reservations
Queen Elizabeth	a lot of
Main Branch	I will call you
Sales Department	that's right
self introduction	how about
zip code	at that time
international cards	the application fee
to the Grand Canyon	is that okay
The Hyatt Regency	I appreciate your
flight number JS	of course
Canadian Rockies and Vancouver	so please hold the line

Figure 2: Some examples of extracted collocations. (English)

the collocations "desho u ka" and "desu ka", which had a high cost reduction, are used very often to make interrogative sentences in Japanese. The collocation "tai no desu ga" is usually used to express a speaker's request, whose meaning is "(I) would like to".

Considering that beginners in the Japanese language are sometimes annoyed by the complex conjugation properties of predicate constituents, it is educationally effective to provide them with typical and frequently used predicate phrase patterns. In that respect, we can say that the cost criteria measure is superior to mutual information.

The comments above are also true of the English data. Mutual information tends to extract compound noun phrases, while cost criteria tends to extract frozen phrase patterns such as "thank you very much" and "I would like to".

Why does mutual information fail to extract these patterns? Here, let us take "I will" as an illustrative example, which has been picked out by cost criteria ("I will" is omitted from Figure 2) but not by mutual information. In our corpus, "I" occurs 2,907 times, "will" occurs 920 times, and "I will" occurs 264 times. Therefore, we have

$$\begin{aligned} I(\mathrm{I}, \mathrm{will}) &= \log \frac{\frac{264}{100{,}000}}{\frac{2907}{100{,}000}\,\frac{920}{100{,}000}} \\ &= 3.3 \end{aligned} \tag{10}$$

This value is not so large, so the two words "I" and "will" cannot be considered to have a significant relationship.

According to the same reasoning, patterns such as "I would like to" and "thank you very much" are excluded as collocation candidates. However, in the ADD Corpus, more than fifty per cent of the sentences that involve the word "would" are subsumed under the pattern "(I) would like to ∼". Therefore, this pattern should be included in the collocation list.

Another drawback using mutual information is the sparseness of data. A corpus cannot provide sufficient data about every word-word relationship. Some word pairs may have high mutual information values in spite of their low frequency in the corpus. For example, the first ranked collocation was "yacht harbor", which occurs only twice in the ADD Corpus. On the contrary, since the cost criteria measure is based on absolute frequency, such phenomena never happens.

Furthermore, because the cost criteria measure estimates the extent to which processing is reduced, it can be considered to be a model of learners' work load. Also, collocations extracted using cost criteria can cover a wide range of human linguistic behavior. To sum up, we can say that the cost criteria measure is more suitable from the viewpoint of language learning.

6 Conclusion

With the growing availability of large textual resources, corpus-based studies are gaining more and more attention among computational linguists and computer scientists. In Particular, automatic acquisition of lexical knowledge from corpora is one of the most important and interesting issues. In this paper, we have taken up the problem of how to acquire collocational knowledge and discussed its importance for language learning. We have also presented an effective measure, called cost criteria, for automatic extraction of collocations from corpora. Comparative experiments with mutual information have shown that the cost criteria measure is more suitable for the purpose of language learning.

Unfortunately, the current implementation can only extract collocations of uninterrupted word sequences. Our next plan is to refine the method to extract collocations of interrupted sequences, and to utilize lexical information such as parts-of-speech in order to prevent an improper word sequence from being recognized as a collocation. Also, we hope to incorporate extracted collocations into a language learning CAI system.

Acknowledgments

The idea of cost criteria was developed while the first author was staying at ATR Interpreting Telephony Research Laboratories. The authors are deeply grateful to co-researchers in ATR, Kentaro Ogura (currently with NTT Network Information Systems Laboratories) and Tsuyoshi Morimoto, for their fruitful discussions and comments. The authors are also grateful to the members of our laboratory in Tokushima University for their help and encouragement. Special thanks to Gerardo Ayala, Ingrid Kirschning and John Phillips for reading the manuscript.

References

[Basili et al. 1992] Basili, R., Pazienza, M. T. and Velardi, P.: "A shallow syntactic analyzer to extract word associations from corpora", *Literary and Linguistic Computing*, Vol. 7, No. 2, pp. 113-123, 1992.

[Brown et al. 1990] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S.: "A statistical approach to machine translation", *Computational Linguistics*, Vol. 16, No. 2, pp. 79-85, 1990.

[Brown et al. 1992] Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C. and Mercer, R. L.: "Class-based n-gram models of natural language", *Computational Linguistics*, Vol. 18, No. 4, pp. 467-479, 1992.

[Brown et al. 1993] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. and Mercer, R. L.: "The mathematics of statistical machine translation: Parameter estimation", *Computational Linguistics*, Vol. 19, No. 2, pp. 263-311, 1993.

[Church and Hanks 1990] Church, K. W. and Hanks, P.: "Word association norms, mutual information, and lexicography", *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29, 1990.

[Church et al. 1991] Church, K. W., Gale, W., Hanks, P. and Hindle, D.: "Using statistics in lexical analysis", *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Zernik U (ed.), Lawrence Erlbaum Associates, pp. 115-164, 1991.

[Ehara et al. 1990] Ehara, T., Ogura, K. and Morimoto, T.: "ATR dialogue database", *Proc. of the 1990 International Conference on Spoken Language Processing*, pp. 1093-1096, 1990.

[Hindle and Rooth 1993] Hindle, D. and Rooth, M.: "Structural ambiguity and lexical relations", *Computational Linguistics*, Vol. 19, No. 1, pp. 103-120, 1993.

[Jelinek 1990] Jelinek, F.: "Self-organized language modeling for speech recognition", *Readings in Speech Recognition*, Waibel, A. and Lee, K. F. (eds.), Morgan Kaufmann Publishers, pp. 450-506, 1990.

[Kita 1992] Kita, K.: *A Study on Language Modeling for Speech Recognition*, Ph.D Thesis, Waseda University, 1992.

[Kita et al. 1993] Kita, K., Ogura, K., Morimoto, T. and Yano, Y.: "Automatically extracting frozen patterns from corpora using cost criteria", *Transactions of Information Processing Society of Japan*, Vol. 34, No. 9, pp. 1937-1943, 1993. (in Japanese)

[Kita et al. 1993b] Kita, K., Hayashi, T. and Yano, Y.: "Corpus-based language learning: Towards practical language learning systems", *Proc. of the 1993 International Conference on Computers in Education*, pp. 355-357, 1993.

[Kupiec 1992] Kupiec, J.: "Robust part-of-speech tagging using a hidden Markov model", *Computer Speech and Language*, No. 6, pp. 225-242, 1992.

[Magerman and Marcus 1990] Magerman, D. M. and Marcus, M. P.: "Parsing a natural language using mutual information statistics", *Proc. of the Eight National Conference on Artificial Intelligence*. pp. 984-989, 1990.

[Nagao 1984] Nagao, M.: "A framework of a mechanical translation between Japanese and English by analogy principle", *Artificial and Human Intelligence*, Elithorn, A. and Banerji, R. (eds.), Elsevier Science Publishers, pp. 173-180, 1984.

[Smadja and McKeown 1990] Smadja, F. A. and McKeown, K. R.: "Automatically extracting and representing collocations for language generation", *Proc. of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 252-259, 1990.

[Smadja 1991] Smadja, F. A.: "Macrocoding the lexicon with co-occurrence knowledge", *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Zernik, U. (ed.), Lawrence Erlbaum Associates, pp. 165-189, 1991.

[Smadja 1993] Smadja, F.: "Retrieving collocations from text: Xtract", *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177, 1993.

[Sumita and Iida 1992] Sumita, E. and Iida, H.: "Example-based NLP techniques: A case study of machine translation", *Proc. of the AAAI Workshop on Statistically-Based NLP Techniques*, pp. 90-97, 1992.

[Wolff 1991] Wolff, J. G.: *Towards a Theory of Cognition and Computing.* Ellis Horwood, 1991.