Word lookup on the basis of associations: from an idea to a roadmap

Michael ZOCK
LIMSI-CNRS
B.P. 133, 91403 Orsay,
France
zock@limsi.fr

Slaven BILAC
Tokyo Institute of Technology
Ookayama 2-12-1, Meguro 152-8552,
Japan
sbilac@cl.cs.titech.ac.jp

Abstract

Word access is an obligatory step in language production. In order to achieve his communicative goal, a speaker/writer needs not only to have something to say, he must also find the corresponding word(s). Yet, knowing a word, i.e. having it stored in a data-base or memory (human mind or electronic device) does not imply that one is able to access it in time. This is a clearly a case where computers (electronic dictionaries) can be of great help.

In this paper we present our ideas of how an enhanced electronic dictionary can help people to find the word they are looking for. The yet-to-be-built resource is based on the age-old notion of association: every idea, concept or word is connected. In other words, we assume that people have a highly connected conceptual-lexical network in their mind. Finding a word amounts thus to entering the network at any point by giving the word or concept coming to their mind (*source word*) and then following the links (associations) leading to the word they are looking for(*target word*).

Obviously, in order to allow for this kind of access, the resource has to be built accordingly. This requires at least two things: (a) indexing words by the associations they evoke, (b) identification and labeling of the most frequent/useful associations. This is precisely our goal. Actually, we propose to build an associative network by enriching an existing electronic dictionary (essentially) with (syntagmatic) associations coming from a corpus, representing the average citizen's shared, basic knowledge of the world (encyclopedia). Such an enhanced electronic database resembles in many respects our mental dictionary. Combining the power of computers and the flexibility of the human mind (omnidirectional navigation and quick jumps), it emulates to some extent the latter in its capacity to navigate quickly and efficiently in a large data base.

While the notions of *association* and *spreading activation* are fairly old, their use to support word access via computer is new. The resource still needs to be built, and this is not a trivial task. We discuss here some of the strategies and problems involved in accomplishing it with the help of people and computers (automation).

1 Introduction

We all experience now and then the problem of being unable to find the word expressing the idea we have in our mind. It we care and have time we may reach for a dictionary. Yet, this kind of resource may be of little help, if it expects from us precisely what we are looking for : a perfectly spelled word, expressing the idea we try to convey. While perfect input may be reasonable in the case of *analysis* (comprehension), it certainly is not in the case of *synthesis* (generation) where the starting point is conceptual in nature: a message, the (partial) definition of a word, a concept or a word related to the target word. The language *producer* needs a dictionary allowing for reverse access. A thesaurus does that, but only in a very limited way: the entry points are basically topical.

People use various methods to initiate search in their mind : words, concepts, partial descriptions, etc. If we want to mimic these functionalities by a computer, we must build the resource accordingly. Let us assume that the text producer is looking for a word that he cannot access. Instead he comes up with another word (or concept)[1] somehow related to the former. He may not know precisely how the two relate, but he knows that they are related. He may also know to some extent how close their relationship is, whether a given link is relevant or not, that is, whether it can lead directly (synonym,

[1] We will comment below on the difference between concepts and words.

antonym, hyperonym) or indirectly to the target word. Since the relationship between the source- and the target word is often indirect, several lookups may be necessary: each one of them having the potential to contain either the target word (direct lookup), or a word leading towards it (indirect lookup).

2 How reasonable is it to expect perfect input?

The expectation of perfect input is unrealistic even in analysis,[2] but clearly more so in generation. The user may well be unable to provide the required information: be it because he cannot access in time the word he is looking for, even though he knows it,[3] or because he does not know the word yet expressing the idea he wants to convey. This latter case typically occurs when using a foreign language or when trying to use a very technical term. Yet, not being able to find a word, does not imply that one does not know anything concerning the word. Actually, quite often the contrary is the case.

Suppose, you were looking for a word expressing the following ideas: *domesticated animal, producing milk suitable for making cheese.* Suppose further that you knew that the target word was neither *cow* nor *sheep*. While none of this information is sufficient to guarantee the access of the intended word *goat*, the information at hand (part of the definition) could certainly be used. For some concrete proposals going in this direction, see (Bilac et al., 2004), or the OneLook reverse dictionary.[4] Besides the definition information, people often have other kind of knowledge concerning the target word. In particular, they know how the latter relates to other words. For example, they know that *goats* and *sheep* are somehow connected, that both of them are *animals*, that *sheep* are appreciated for their wool and meet, that *sheep* tend to follow each other blindly, while *goats* manage to survive, while hardly eating anything, etc. In sum, people have in their mind lexical networks: all words, concepts or ideas they express are highly interconnected. As a result, any one of the words or concepts has the potential to evoke each other. The likelihood for this to happen depends, among other things, on such factors as *frequency* (associative strength), *saliency* and *distance* (direct vs. indirect access). As one can see, associations are a very general and powerful mechanism. No matter what we hear, read or say, any idea is likely to remind us of something else.[5] This being so, we should make use of it.[6]

3 Search based on the relations between concepts and words

If one agrees with what we have just said, one could view the *mental dictionary* as a huge semantic network composed of *nodes* (words and concepts) and *links* (associations), with either being able to activate the other.[7] Finding a

[2] Obviously, looking for "pseudonym" under the letter "S" in a dictionary won't be of great help.

[3] Temporary amnesia, known as the TOT, or tip-of-the-tongue problem (Brown and McNeill, 1996; Zock and Fournier, 2001; Zock, 2002)

[4] `http://www.onelook.com/reverse-dictionary.shtml`

[5] The idea according to which the mental dictionary (or encyclopedia) is basically an associative network, composed of nodes (words or concepts) and links (associations) is not new, neither is the idea of spreading activation. Actually the very notion of association goes back at least to Aristotle (350BC), but it is also inherent in work done by philosophers (Locke, Hume), physiologists (James & Stuart Mills), psychologists (Galton, 1880; Freud, 1901; Jung and Riklin, 1906) and psycholinguists (Deese, 1965; Jenkins, 1970; Schvaneveldt, 1989). For surveys in psycholinguistics see (Hörmann, 1972), or more recent work (Spitzer, 1999). The notion of association is also implicit in work on semantic networks (Quillian, 1968), hypertext (Bush, 1945), the web (Nelson, 1967), connectionism (Dell et al., 1999) and, of course, in WordNet (Miller et al., 1993; Fellbaum, 1998).

[6] In the preceding sections we used several times the terms *words* and *concepts* interchangeably, as if they were the same. Of course, they are very different. Yet, not knowing what a concept looks like (a single node, or every node, i.e. headword of the word's definition?), we think it is safer to assume that the user can communicate with the computer (dictionary) only via words. Hence, concepts are represented by words, yet, since the two are connected, one can be accessed via the other, which addresses the interface problem with the computer. Another point worth mentionning is the fact that associations may depend on the nature of the arguments (words vs. concepts). While in theory anything can be associated with anything (words with words, words with concepts, concepts with concepts, etc.), in practice words tend to trigger a different set of associations than concepts. Also, the connectivity between words and concepts explains to some extent the power and the flexibility of the human mind. Words are shorthand labels for concepts, and given the fact that the two are linked, one can make big leaps in no time and easily move from one plane (let's say the conceptual level) to the other (the linguistic counterpart). Words can be reached via concepts, but the latter can also serve as starting point to find a word. Compared to the links between concepts which are a superhighway, associations between words are more like countryroads.

[7] Actually, one could question the very notion of *mental dictionary* which is convenient, but misleading in as it supposes a dedicated part for this task in our brain. A

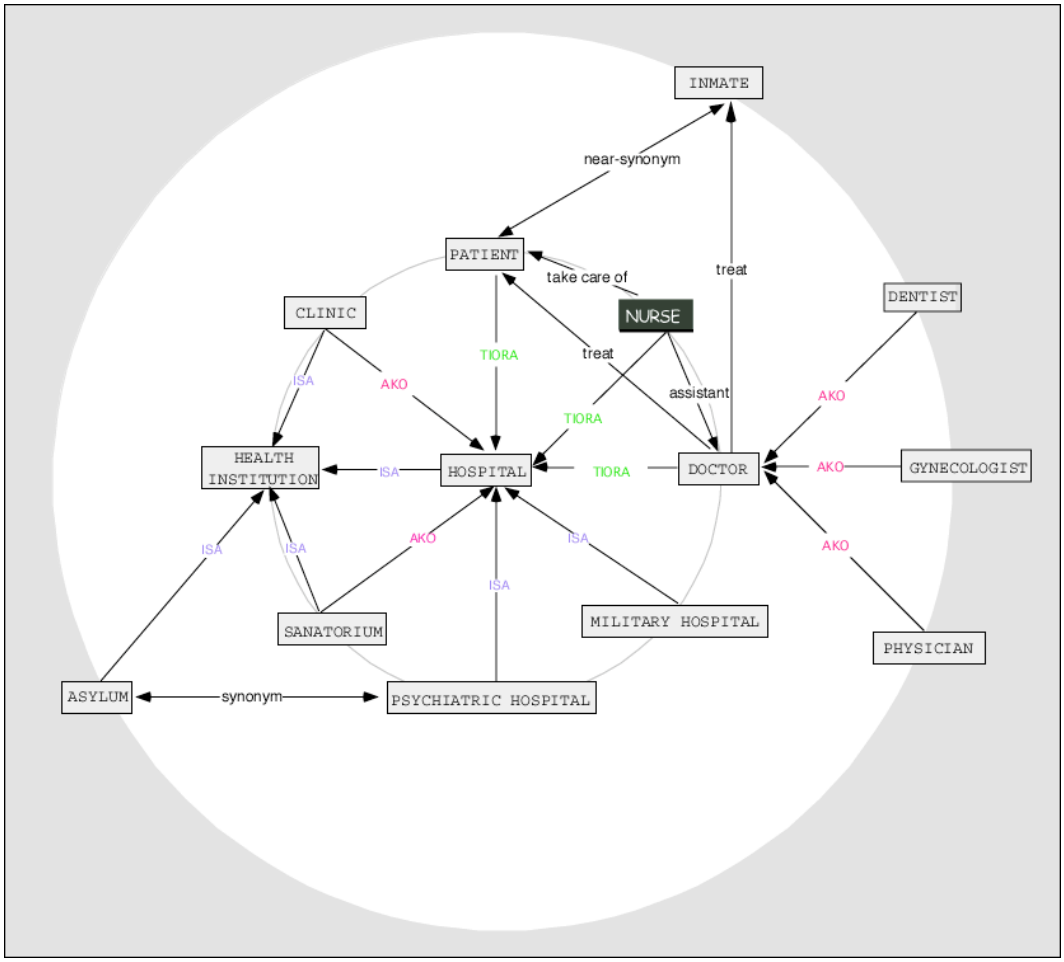

Figure 1: Search based on propagation in a network (internal representation)

word amounts thus to entering the network and following the links leading from the *source node* (the first word that comes to your mind) to the *target word* (the one you are looking for). Suppose you wanted to find the word "nurse" (*target word*), yet the only token coming to your mind were "hospital". In this case the system would generate internally a graph with the *source word* at the center and all the associated words at the periphery. Put differently, the system would build internally a semantic network with "hospital" in the center and all its associated words as satellites (figure 1).[8]

Obviously, the greater the number of associations, the more complex the graph. Given the diversity of situations in which a given object may occur we are likely to build many associations. In other words, lexical graphs tend to become complex, too complex to be a good representation to support navigation. Readability is hampered by at least two factors: *high connectivity* (the great number of links or associations emanating from each word), and *distribution*: conceptually related nodes, that is, nodes activated by the same kind of assocation are scattered around, that is, they do not necessarily occur next to each other, which is quite confusing for the user. In order to solve this problem we suggest to display by category (chunks) all the words linked by the same kind of association to the source word (see figure 2). Hence, rather than displaying all the connected words as a flat list, we suggest to present them in chunks to allow for categorial search. Having chosen a category, the user will be presented a list of words or categories from which he must choose. If the target word is in the category chosen by the user (suppose he looked for a hyperonyme, hence he checked the ISA-bag), search stops, otherwise it goes on. The user could choose either another category (eg. AKO or TIORA), or a word in

multiply indexed *mental encyclopedia*, composed of polymorph information (concepts, words, meta-linguistic information) seems much more plausible to us.

[8] **AKO**: a kind of; **ISA**: subtype; **TIORA**: typically involved object, relation or actor.

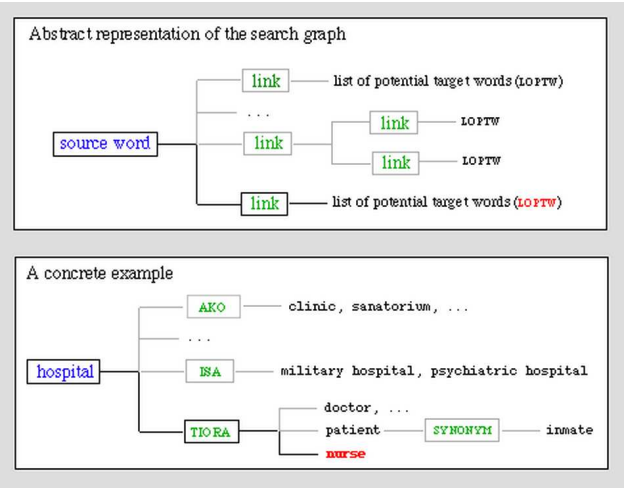

Figure 2: Proposed candidates, grouped according to the nature of the link

the current list, which would then become the new starting point.

4 A resource still to be built

The fact that the links are labeled has some very important consequences. (a) While maintaining the power of a highly connected graph (possible cyclic navigation), it has at the interface level the simplicity of a tree: each node points only to data of the same type, i.e. same kind of association. (b) Words being presented in clusters, navigation can be accomplished by clicking on the appropriate category. The assumption being that the user generally knows to which category the target word belongs (or at least, he can recognize within which of the listed categories it falls), and that categorical search is in principle faster than search in a huge list of unordered (or, alphabetically ordered) words.

Word access, as described here, amounts to navigating in a huge associative network. Of course, such a network has to be built. The question is how. Our proposal is to build it automatically by parsing an existing corpus containing sufficient amount of information on world knowledge (for example, an encyclopedia). This would yield a set of associations (see below),[9] which still need to be labeled. A rich ontology should be helpful in determining the adequate label for many, if not most of the links. Unlike private information,[10] which by definition cannot and should not be put into a public dictionary,[11] encyclopedic knowledge can be added in terms of associations, as this information expresses commonly shared knowledge, that is, the kind of associations most people have when encountering a given word. Take for example the word *elephant*. An electronic dictionary like Word Net associates the following gloss with the headword: *large, gray, four-legged mammal*, while Webster gives the following information:

> A mammal of the order Proboscidia, of which two living species, Elephas Indicus and E. Africanus, and several fossil species, are known. They have a proboscis or trunk, and two large ivory tusks proceeding from the extremity of the upper jaw, and curving upwards. The molar teeth are large and have transverse folds. Elephants are the largest land animals now existing.

While this latter entry is already quite rich (trunk, ivory tusk, size), an encyclopedia contains even more information.[12] If all this information were added to an electronic resource, it would enable us to access the same word (e.g. *elephant*) via many more associations than ever before. By looking at the definition here above, one will notice that many associations are quite straightforward (color, size, origin, etc.), and since most of them appear frequently in a pattern-like manner it should be possible to extract them automatically (see footnote 18 below). If one agrees with these views, the remaining question is how to extract this encyclopedic information and to add it to an existing electronic resource. Below we will outline some methods for extracting associated words and discuss the feasibility of using current methodology to achieve this goal.

5 Automatic extraction of word associations

Above we outlined the need for obtaining associations between words and using them to improve dictionary accessibility. While the associations can be obtained through association experiments with human subjects, this strategy is

[9] The assumption being that every word co-occurring with another word in the same sentence is a candidate of an association. The more frequently two words co-occur in a given corpus, the greater their associative strength.

[10] For example, the word *elephant* may remind you of a specific animal, trip or location (zoo, country in Africa).

[11] This does not (and should not) preclude the possibility to add it to one's personal dictionary.

[12] You may consider taking a look at Wikipedia (`http://en.wikipedia.org/wiki/`) which is free.

not very satisfying due to the high cost of running the experiments (time and money), and due to its static nature. Indeed, given the costs, it is impossible to repeat these experiments to take into account the evolution of a society. Hence, the goal is to automatically extract associations from large corpora. This problem was addressed by a large number of researchers, but in most cases it was reduced to extraction of collocations which are a proper subset of the set of associated words. While hard to define, collocations appear often enough in corpora to be extractable by statistical and information-theory based methods.

There are several basic methods for evaluating associations between words: based on *frequency counts* (Choueka, 1988; Wettler and Rapp, 1993), *information theoretic* (Church and Hanks, 1990) and *statistical significance* (Smadja, 1993). The statistical significance often evaluate whether two words are independant using hypothesis tests such as t-score (Church et al., 1991), the X^2, the log-likelihood (Dunning, 1993) and Fisher's exact test (Pedersen, 1996). Extracted sets for associated words are further pruned using numerical methods, or linguistic knowledge to obtain a subset of collocations.

The various extraction measures have been discussed in great detail in the literature (Manning and Schütze, 1999; McKeown and Radev, 2000), their performance has been compared (Dunning, 1993; Pedersen, 1996; Evert and Krenn, 2001), and the methods have been combined to improve overall performance (Inkpen and Hirst, 2002). Most of these methods were originally applied in large text corpora, but more recently the web has been used as a corpus (Pearce, 2001; Inkpen and Hirst, 2002). Collocation extraction methods have been used not only for English, but for many other languages: French (Ferret, 2002), German (Evert and Krenn, 2001) and Japanese (Nagao and Mori, 1994), to cite but those.

The most obvious question in this context is to clarify to what extent available collocation extraction techniques fulfill our needs of extracting and labeling word associations. Since collocations are a subset of association, it is possible to apply collocation extraction techniques to obtain related words, ordered in terms of the relative strength of association.

The result of this kind of numerical extraction would be a large set of numerically weighted word pairs. The problem with this approach is that the links are only labeled in terms of their relative associative strength, but not categorically, which makes it impossible to group and present them in a meaningful way for the dictionary user. Clusters based only on the notion of association strength are inadequate for the kind of navigation described here above. Hence another step is necessary: qualification of the links according to their types. Only once this is done, a human being could use it to navigate through a large conceptual-lexical network (the dictionary) as described above. Unfortunately, research on automatic link identification has been rather sparse. Most attempts have been devoted to the extraction of certain types of links (usually syntactic type (Lin, 1998) or on extensions of WordNet with topical information contained in a thesaurus (Stevenson, 2002) or on the WWW (Agirre et al., 2000). Additional methods need to be considered in order to reveal (automatically) the kind of associations holding between words and/or concepts. Earlier in this paper we have suggested the use of an encyclopedia as a source of general world knowledge. It should be noted, though, that there are important differences between large corpora and encyclopedias. Large corpora usually contain a lot of repetitive texts on a limited number of topics (e.g. newspaper articles) which makes them very suitable for statistical methods. On the other hand, while being maximally informative and comprehensive, encyclopedias are written in a highly controlled language, and their content is continually updated and re-edited, with the goal to avoid unnecessary repetition. While most of the information contained in an entry is important, there is a lack of redundancy. Hence, measures capable of handling word pairs with low appearance counts (e.g. log-likelihood or Fisher's exact test) should be favored. Also, rather than looking at individual words, one might want to look at word patterns instead.

6 Discussion and Conclusion

We have raised and partially answered the question of how a dictionary should be indexed in order to support word access. We were particularly concerned with the language producer, as his needs (and knowledge at the onset) are quite different from the ones of the language receiver (listener/reader). It seems that, in order to achieve our goal, we need to do two things:

add to an existing electronic dictionary information that people tend to associate with a word, that is, build and enrich a semantic network, and provide a tool to navigate in it. To this end we have suggested to label the links, as this would reduce the graph complexity and allow for type-based navigation. Actually our basic proposal is to extend a resource like WordNet by adding certain links, in particular on the horizontal axis (syntagmatic relations). These links are associations, and their role consists in helping the encoder to find ideas (concepts/words) related to a given stimulus (brainstorming), or to find the word he is thinking of (word access).

One problem that we are confronted with is to identify possible associations. Ideally we would need a complete list, but unfortunately, this does not exist. Yet, there is a lot of highly relevant information out there. For example, Mel'cuk's lexical functions (Mel'cuk, 1992), Fillmore's FRAMENET[13], work on ontologies (CYC), thesaurus (Roget), WordNets (the original version from Princeton, divers EuroWordNets, BalkaNet), HowNet[14], the work done by MICRA, the FACTOTUM project[15] or the Wordsmyth dictionary/thesaurus combination[16]. Of course, one would need to make choices here and probably add links. Another problem is to identify *useful* associations. Not every possible association is necessarily plausible. Hence, the idea to take as corpus something that expresses shared knowledge, for example, an encyclopedia. The associations it contains can be considered as being plausible. We could also collect data by watching people using a dictionary and identify search patterns.[17] Next, we could run psycholinguistic experiments.[18] While the typical paradigm has been to ask people to produce a response (red) to some stimulus (rose), we could ask them to identify or label the links between words (e.g. *apple-fruit*, *lemon-yellow*, etc.). The ease of labeling will probably depend upon the origin of the words (the person asked to label the link or somebody else).

Another approach would be to extract collocations from a corpus and label them automatically. There are tools for extracting cooccurrences (see section 5.5), and ontologies could be used to qualify some of the links between collocational elements. While this approach might work fine for couples like *coffee-strong*, or *wine-red* (since an ontology would reveal that *red* is a kind of *color*, which is precisely the link type: i.e. association), one may doubt that it could reveal the nature of the link between *smoke* and *fire*. Yet, most humans would immediately recognize this as a *causal* link. As one can see, there are still quite a few serious problems to be solved. Nevertheless, we do believe that these obstacles can be removed, and that the approach presented here has the potential to improve word access, making the whole process more powerful, natural and intuitive, hence efficient.

[13] `http://www.icsi.berkeley.edu/~framenet/`

[14] `http://www.keenage.com/html/e_index.html`

[15] `http://humanities.uchicago.edu/homes/MICRA/`

[16] `http://www.wordsmyth.com/`

[17] One such pattern could be: give me the word for a bird with yellow feet and a long beak, that can swim. Actually, word access problems frequently come under the form of questions like: What is the word for X that Y?, where X is usually a hypernym and Y a stereotypical, possibly partial functional/relational/case description of the target word.

[18] Actually, this has been done for decades, but with a different goal in mind (Nelson, 1967), `http://cyber.acomp.usf.edu/FreeAssociation/`.

References

E. Agirre, E. Hovy O. Ansa, and D. Martinez. 2000. Enriching very large ontologies using the WWW. In *Proc. of ECAI Ontology Learning Workshop*.

S. Bilac, W. Watanabe, T. Hashimoto, T. Tokunaga, and H. Tanaka. 2004. Dictionary search based on the target word description. In *Proc. of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004)*, pages 556–559.

R. Brown and D. McNeill. 1996. The tip of the tonuge phenomenon. *Journal of Verbal Learning and Verbal Behaviour*, 5:325–337.

V. Bush. 1945. As we may think. *The Atlantic Monthly*, 176:101–108.

Y. Choueka. 1988. Looking for needles in a haystack. In *Proc. of the RIAO Conference on User-Oriented Context Based Text and Image Handling*, pages 609–623.

K. Church and P. Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16:22–29.

K. Church, W. Gale, P. Hanks, and D. Hindle. 1991. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrance Erlbaum Associates.

J. Deese. 1965. *The structure of associations in language and thought*. Johns Hopkins Press.

G. S. Dell, F. Chang, and Z. M. Griffin. 1999. Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, 23:517–542.

T. Dunning. 1993. Accurate methods for statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

S. Evert and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proc. of the 39th Annual meeting of Association of Computational Linguistics (ACL 2001)*, pages 188–195.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database and some of its Applications*. MIT Press.

O. Ferret. 2002. Using collocations for topic segmentation and link detection. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 261–266.

S. Freud. 1901. *Psychopathology of everyday life*. Payot, 1997 edition.

F. Galton. 1880. Psychometric experiments. *Brain*, 2:149–162.

H. Hörmann. 1972. *Introduction à la psycholinquistique*. Larousse.

D. Z. Inkpen and G. Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *Proc. of Unsupervised Lexical Acquisition Workshop of the ACL SIGLEX*, pages 67–76.

J. J. Jenkins. 1970. The 1952 minnesota word association norms. In L. Postman and G. Kepper, editors, *Norms of Word Association*, pages 1–38. Academic Press.

C. G. Jung and F. Riklin. 1906. Experimentelle unter-suchungen über assoziationene gesunder. In C. G. Jung, editor, *Diagnostische Assoziationsstudien*, pages 7–145. Barth.

D. Lin. 1998. Extracting collocations from text corpor. In *First Workshop on Computational Terminology*.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

K. R. McKeown and Dragomir R. Radev. 2000. Collocations. In H. Moisl R. Dale and H. Somers, editors, *Handbook of Natural Language Processing*, pages 507–523. Marcel Dekker.

I. Mel'cuk. 1992. *Dictionnaire Explicatif et Combinatoire du français contemporain: recherche lexicosémantique III*. Les presses de l'université de Montréal.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and Katherine Miller, editors. 1993. *Introduction to WordNet: An On-line Lexical Database*. Cognitive Science Laboratory, Princeton University.

M. Nagao and S. Mori. 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In *Proc. of the 15th International Conference on Computational Linguistics (COLING 1994)*, pages 611–615.

T. Nelson. 1967. Xanadu projet hypertextuel.

D. Pearce. 2001. Synonymy in collocation extraction. In *Proc. of NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Ted Pedersen. 1996. Fishing for exactness. In *Proc. of the South-Central SAS Users Group Conference*, pages 188–195.

R. Quillian. 1968. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 216–270.The MIT Press. Cambridge, MA.

R. Schvaneveldt, editor. 1989. *Pathfinder Associative Networks: studies in knowledge organization*. Norwood.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

M. Spitzer. 1999. *The mind within the net: models of learning, thinking and acting*. MIT Press.

M. Stevenson. 2002. Augmenting noun taxonomies by combining lexical similarity metrics. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 953–959.

M. Wettler and R. Rapp. 1993. Computation of word associations based on the co-occurrences of words in large corpora. In *Proc. of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives*.

M. Zock and J.-P. Fournier. 2001. How can computers help the writer/speaker experiencing the tip-of-the-tongue problem ? In *Proc. of RANLP*, pages 300–302.

M. Zock. 2002. Sorry, what was your name again, or how to overcome the tip-of-the tongue problem with the help of a computer? In *Proc. of the SemaNet workshop COLING2002*.