# Language Resources for a Network-based Dictionary

**Veit Reuer**
Institute of Cognitive Science
University of Osnabrück
Germany
vreuer@uos.de

## Abstract

In order to facilitate the use of a dictionary for language production and language learning we propose the construction of a new network-based electronic dictionary along the lines of Zock (2002). However, contrary to Zock who proposes just a paradigmatic network with information about the various ways in which words are similar we would like to present several existing language resources (LRs) which will be integrated in such a network resulting in more linguistic levels than one with paradigmatically associated words. We argue that just as the mental lexicon exhibits various, possibly interwoven layers of networks, electronic LRs containing syntagmatic, morphological and phonological information need to be integrated into an associative electronic dictionary.

## 1 Introduction

Traditional dictionaries are mainly used for language reception, even though they were also developed to be used for language production. However the form-based structure following orthographic conventions which could also be called "one-dimensional", makes it difficult to access the information by meaning. Therefore the usage of a traditional dictionary for text production is quite limited as opposed to, for example, a thesaurus. The main advantage of a thesaurus is the structuring based on the semantic relation between words in an entry. This allows for the availability of a different type of information.

Therefore our proposal is to construct an electronic dictionary which has a network-like structure and whose content is drawn from various existing lexical resources. The dictionary will represent both paradigmatic information - information about the various ways in which words are similar - as well as syntagmatic information - information about the relationships among words that appear together. Additionally information from other types of resources such as morphology and phonology will be integrated as they are also relevant in models of the mental lexicon. In these models "associations" between words are based not only on meaning but also on phonological or morphological properties of the connected words. Following Brown and McNeill (1966) and subsequent research people in the so-called "tip-of-the-tongue"-state (TOT-state) are able to clearly recall the properties of the missing word such as the number of syllables or the meaning, and can easily identify the target word when it is presented to them.

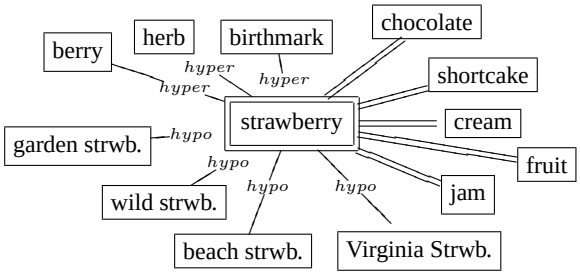


Figure 1: Example network with data from WordNet (–) and Deutscher Wortschatz (=)

Figure 1 exemplifies the visualization of a single node with related information from two LRs[1]. Here a user would be able to find the term "shortcake" even if s/he only knows only one part, namely strawberries.[2] A click on a neighbouring node should e.g. re-center the structure and hence allow the user to "explore" the network.

As mentioned above the most obvious usage seems to be in language production where information can be provided not only for words already activated in the mind of the language producer but also for alternatives, specifications or for words not directly accessible because of a TOT-state. This seems reasonable in light of the fact that speaker's passively vocabularies are known to be larger than

[1] http://www.cogsci.princeton.edu/~wn
http://www.wortschatz.uni-leipzig.de

[2] LDOCE (1987) mentions "cream" and "jam" but not "shortcake" as part of the entry for "strawberry". The entry for "shortcake" however lists specifically "strawberry shortcake".

their active vocabularies. The range of information available of course depends on the material integrated into the dictionary from the various resources which are explored more closely below.

A second area of application of such a dictionary is language learning. Apart from specifying paradigmatic information which is usually also part of the definition of a lemma, syntagmatic information representing collocations and cooccurrances is an important resource for language learners. Knowledge about collocations is a kind of linguistic knowledge which is language-specific and not systematically derivable making collocations especially difficult to learn.

Even though there are some studies that compare the results from statistically computed association measures with word association norms from psycholinguistic experiments (Landauer et al., 1998; Rapp, 2002) there has not been any research on the usage of a digital, network-based dictionary reflecting the organisation of the mental lexicon to our knowledge. Apart from studies using so called Mind Maps or Concept Maps to visualize "world knowledge"[3] (Novak, 1998) nothing is known about the psycholinguistic aspects which need to be considered for the construction of a network-based dictionary.

In the following section we will summarize the information made available by the various LRs we plan to integrate into our system. The ideas presented here were developed in preparation of a project at the University of Osnabrück.

## 2 Language Resources

Zock (2002) proposes the use of only one type of information structure in his network, namely a type of semantic information. There are, however, a number of other types of information structures that may also be relevant for a user. Psychological experiments show that almost all levels of linguistic description reveal priming effects. Strong mental between words are based not only on a semantic relationship but also on morphological an phonological relationships. These types of relationships should also be included in a network based dictionary as well.

A number of LRs that are suitable in this scenario already provide some sort of network-like structure possibly closely related to networks meaningful to a human user. All areas are large research fields of their own and we will therefore only touch upon a few aspects.

[3]The maps constitute a representation of the world rather than reflecting the mental lexicon.

### 2.1 Manually Constructed Networks

Manually constructed networks usually consist of paradigmatic information since words of the same part-of-speech are related to each other. In ontologies usually only nouns are considered and are integrated into these in order to structure the knowledge to be covered.

The main advantage of such networks, since they are hand-built, is the presumable correctness (if not completeness) of the content. Additionally, these semantic nets usually include typed relations between nodes, such as e.g. "hyperonymy" and "is_a" and therefore provides additional information for a user. It is safe to rely on the structure of a network coded by humans to a certain extend even if it has certain disadvantages, too. For example networks tend to be selective on the amount of data included, i.e. sometimes only one restricted area of knowledge is covered. Furthermore they include basically only paradigmatic information with some exceptions. This however is only part of the greater structure of lexical networks.

The most famous example is WordNet (Fellbaum, 1998) for English – which has been visualized already at http://www.visualthesaurus.com – and its various sisters for other languages. It reflects a certain cognitive claim and was designed to be used in computational tasks such as word sense disambiguation. Furthermore ontologies may be used as a resource, because in ontologies usually single words or NPs are used to label the nodes in the network. An example is the "Universal Decimal Classification"[4] which was originally designed to classify all printed and electronic publications in libraries with the help of some 60,000 classes. However one can also think of it as a knowledge representation system as the information is coded in order to reflect the knowledge about the topics covered.

### 2.2 Automatically Generated Paradigmatic Networks

A common approach to the automatic generation of semantic networks is to use some form of the so called vector-space-model in order to map semantically similar words closely together in vector space if they occur in similar contexts in a corpus (Manning and Schütze, 1999). One example, Latent Semantic Analysis (Landauer et al., 1998, LSA) has been accepted as a model of the mental lexicon and is even used by psycholinguists as a basis for the categorization and evaluation of test-items. The results from this line of research seem to describe not only relations between words but seem to provide

[4]http://www.udcc.org

the basis for a network which could be integrated into a network-based dictionary. A disadvantage of LSA is the positioning of polysemous words at a position between the two extremes, i.e. between the two senses which makes the approach worthless for polysemous words in the data.

There are several other approaches such as Ji and Ploux (2003) and the already mentioned Rapp (2002). Ji and Ploux also develop a statistics-based method in order to determine so called "contexonyms". This method allows one to determine different senses of a word as it connects to different clusters for the various senses, which can be seen as automatically derived SynSets as they are known from WordNet. Furthermore her group developed a visualization tool, that presents the results in a way unseen before. Even though they claim to have developed an "organisation model" of the mental lexicon only the restricted class of paradigmatic relations shows up in their calculations.

Common to almost all the automatically derived semantic networks is the problem of the unknown relation between items as opposed to manually constructed networks. On the one hand a typed relation provides additional information for a user about two connected nodes but on the other hand it seems questionable if a known relation would really help to actually infer the meaning of a connected node (contrary to Zock (2002)).

### 2.3 Automatically Generated Syntagmatic Networks

Substantial parts of the mental lexicon probably also consist of syntagmatic relations between words which are even more important for the interpretation of collocations.[5] The automatic extraction of collocations, i.e. syntagmatic relations between words, from large corpora has been an area of interest in recent years as it provides a basis for the automatic enrichment of electronic lexicons and also dictionaries. Usually attempts have been made at extracting verb-noun-, verb-PP- or adjective-noun-combinations. Noteworthy are the works by Krenn and Evert (2001) who have tried to compare the different lexical association measures used for the extraction of collocations. Even though most approaches are purely statistics-based and use little linguistic information, there are a few cases where a parser was applied in order to enhance the recognition of collocations with the relevant words not being next to each other (Seretan et al., 2003).

The data available from the collocation extraction research of course cannot be put together to give a complete and comprehensive network. However certain examples such as the German project "Deutscher Wortschatz"[6] and the visualization technique used there suggest a network like structure also in this area useful for example in the language learning scenario as mentioned above.

### 2.4 Phonological/Morphological Networks

Electronic lexica and rule systems for the phonological representation of words can be used for spell-checking as has been done e.g. in the Soundex approach (Mitton, 1996). In this approach a word not contained in the lexicon is mapped onto a simplified and reduced phonological representation and compared with the representations of words in the lexicon. The correct words coming close to the misspelled word on the basis of the comparison are then chosen as possible correction candidates. However this approach makes some drastic assumptions about the phonology of a language in order to keep the system simple. With a more elaborate set of rules describing the phonology of a language a more complex analysis is possible which even allows the determination of words that rhyme.[7] Setting a suitable threshold for some measure of similarity a network should evolve with phonologically similar words being connected with each other. A related approach to spelling correction is the use of so called "tries" for the efficient storage of lexical data (Oflazer, 1996). The calculation of a minimal editing distance between an unknown word and a word in a trie determines a possible correct candidate.

Contrary to Zock (2002) who suggests this as an analysis step on its own we think that the phonological and morphological similarity can be exploited to form yet another layer in a network-based dictionary. Zock's example of the looked-for "relegate" may than be connected to "renegade" and "delegate" via a single link and thus found easily. Here again, probably only partial nets are created but they may nevertheless help a user looking for a certain word whose spelling s/he is not sure of.

Finally there are even more types of LRs containing network-like structures which may contribute to a network-based dictionary. One example to be mentioned here is the content of machine-readable

[5] We define collocations as a syntactically more or less fixed combination of words where the meaning of one word is usually altered so that a compositional construction of the meaning is prevented.

[6] Note however that the use of the term "Kollokation" in this project is strictly based on statistics and has no relation to a collocation in a linguistic sense (see figure 1).

[7] Dissertation project of Tobias Thelen: personal communication.

dictionaries. The words in definitions contained in the dictionary entries – especially for nouns – are usually on the one hand semantically connected to the lemma and on the other hand are mostly entries themselves which again may provide data for a network. In research in computational linguistics the relation between the lemma and the definition has been utilized especially for word sense disambiguation tasks and for the automatic enrichment of language processing systems (Ide and Véronis, 1994).

## 3 Open Issues and Conclusion

So far we have said nothing about two further important parts of such a dictionary: the representation and the visualization of the data. There are a number of questions which still need to be answered in order to build a comprehensive dictionary suitable for an evaluation. With respect to the representation two major questions seem to be the following.

- As statistical methods for the analysis of corpora and for the extraction of frequent cooccurrance phenomena tend to use non-lemmatized data, the question is whether it makes sense to provide the user with the more specific data based on inflected material.
- Secondly the question arises how to integrate different senses of a word into the representation, if the data provides for this information (as WordNet does).

With regard to visualization especially the dynamic aspects of the presentation need to be considered. There are various techniques that can be used to focus on parts of the network and suppress others in order to make the network-based dictionary manageable for a user which need to be evaluated in usability studies. Among these are hyperbolic views and so-called cone trees

As we have shown a number of LRs, especially those including syntagmatic, morphological and phonological information, provide suitable data to be included into a network-based dictionary. The data in these LRs either correspond to the presumed content of the mental lexicon or seem especially suited for the intended usage. One major property of the new type of dictionary proposed here is the disintegration of the macro- and the micro-structure of a traditional dictionary because parts of the micro-structure (the definition of the entries) become part of the macro-structure (primary links to related nodes) of the new dictionary. Reflecting the structure of the mental lexicon this dictionary should allow new ways to access the lexical data and support language production and language learning.

## References


R. Brown and D. McNeill. 1966. The 'tip-of-the-tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5:325–337.

C. Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

N. Ide and J. Véronis. 1994. Machine readable dictionaries: What have we learned, where do we go. In *Proceedings of the post-COLING94 international workshop on directions of lexical research*, Beijing.

H. Ji and S. Ploux. 2003. A mental lexicon organization model. In *Proceedings of the Joint International Conference on Cognitive Science*, pages 240–245, Sydney.

B. Krenn and S. Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL-Workshop on Collocations*, pages 39–46, Toulouse.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

LDOCE. 1987. *Longman Dictionary of Contemporary English*. Langenscheidt, Berlin, 2. edition.

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

R. Mitton. 1996. *English Spelling and the Computer*. Longman, London.

J. D. Novak. 1998. *Learning, Creating and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Erlbaum, Mahwah, NJ.

K. Oflazer. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89.

R. Rapp. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proc. 19th Int. Conference on Computational Linguistics (COLING)*, Taipeh.

V. Seretan, L. Nerima, and E. Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the International Conference on Recent Advances in NLP (RANLP-2003)*, Borovets, Bulgaria.

M. Zock. 2002. Sorry, but what was your name again, or, how to overcome the tip-of-the tongue problem with the help of a computer? In *Proceedings of the COLING-Workshop on Building and Using Semantic Networks*, Taipeh.