Chapter 12

Natural language processing

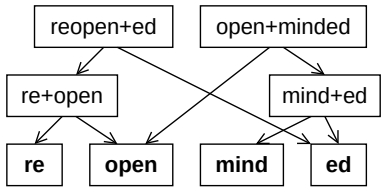
Krista Lagus, Mathias Creutz, Mikko Kurimo, Krister Lindén

12.1 Unsupervised segmentation of words into morphs

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis would be beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. Many existing applications make use of *words* as vocabulary units. However, for some languages, e.g., Finnish and Turkish, this is infeasible, as the number of possible word forms is very high. The productivity of word forming in Finnish is illustrated in Figure 12.1a, where one single word consists of six morphemes.

Morph.	kahvi	n	juo	ja	lle	kin
Transl.	coffee	of	drink	-er	for	also

(a)



(b)

Figure 12.1: (a) Morphological segmentation of the Finnish word for "also for [the] coffee drinker". (b) Hypothetical binary splitting trees for the words "reopened" and "open-minded" (segmented as "re+open+ed" and "open+mind+ed").

We have developed language-independent, data-driven methods for the unsupervised segmentation of the words in a corpus. We call the resulting segments *morphs* and we do not distinguish between categories, such as stems, suffixes, and prefixes. For us words simply consist of (possibly lengthy) sequences of morphs concatenated together. In this sense, our work differs from most previous work in the field of automated morphology learning, where more limitations are set on word structure (e.g., [1]). Instead, our work resembles algorithms for unsupervised text segmentation and word discovery (e.g., [2]).

In [3] we present a model inspired by the Minimum Description Length (MDL) principle. This means that we try to find the optimal balance between *accuracy* of representation and model *complexity*, which generally improves generalization capacity by inhibiting overlearning. In concrete terms, we construct a morph vocabulary, or a *lexicon of morphs*, so that it is possible to form any word in the corpus by the concatenation of some morphs. Each word in the corpus is rewritten as a *sequence of morph pointers*, which point to entries in the lexicon. We aim at finding the optimal lexicon and segmentation, i.e., a set of morphs that is concise, and moreover gives a concise representation for the corpus.

The optimal segmentation is obtained by splitting words recursively and trying to find common subword chunks, which are potential morphs. Figure 12.1b shows a hypothetical recursive splitting of two English words. The leaf nodes correspond to morphs discovered by the algorithm.

Results

For evaluation purposes, we have constructed a gold standard segmentation based on linguistic theory for 1.4 million Finnish and twenty thousand English word forms. When comparing our MDL-inspired method to the main other method, called Linguistica (cf. [1]), we obtain clearly better results on Finnish, and for very small data sets on English, whereas Linguistica is better on English for larger data sets. However, Linguistica utilizes some linguistic assumptions particularly suitable for Indo-European languages.

aamuyö + stä, elma + n, hyvinvointi + yhteis + kuntina, kellari + ssa + kaan, lait + tomasti + kin, miljonääri + ltä, palkka + tuloilla + an, rebrov + ia, sunnuntai + vuorossa, tuuli + potentiaali + sta, wal + ston + in, ääni + lehtenä
abandon, a + shore, book + ers, cherry, cooper + s, dinner + s, entrance, form + ing, harpsichord + ist, in + jury, learned, men + a + c + ing, n + un, pick + ers, radio + activity, sir, succeed + ing, travel + ler, war + 's

Figure 12.2: Examples of Finnish and English words segmented by our algorithm.

Some sample segmentations are shown in Figure 12.2. They include correctly segmented words, where each boundary coincides with a real morpheme boundary (e.g., "kellari+ssa+kaan", "miljonääri+ltä", "dinner+s", "form+ing", "war+'s"). In addition, there are over-segmented words, with boundaries inserted at incorrect locations (e.g., "men+a+c+ing" instead of "menac+ing"), as well as under-segmented words, where some boundary is missing (e.g., "learned" instead of "learn+ed"; "ääni+lehtenä" instead of "ääni+lehte+nä"). Sometimes many alternative segmentations seem correct: e.g., "hyvin+vointi", "hyv+in+voi+nti".

In [4] we re-formulated the model in a probabilistic framework and studied whether prior information about morph length and frequency could be utilized to avoid over- and under-segmentation. This did not lead to considerable improvements in overall accuracy, though. So far each morph has been considered individually, irrespective of the previous and next morph. In future research, the model will be extended to learn information on morph categories and sequences of them.

Demonstration and applications

An online demonstration of the model is available at the address: `http://www.cis.hut.fi/projects/morpho/`. The demonstration allows the user to select a corpus to be used as training data and to type in words that are then segmented according to the model.

The algorithm in [3] has been applied for producing morph vocabularies for *automatic speech recognition*, both for Finnish [5] and Turkish [6] (cf. Section 13.2). Among the different vocabulary approaches tested, the ones based on morphs were most successful.

Acknowledgement

We appreciate the contribution of Sami Virpioja, who implemented the online demo.

References

[1] J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2), pages 153–198, 2001.

[2] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34, pages 71–105, 1999.

[3] M. Creutz and K. Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, 2002.

[4] M. Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, 2003.

[5] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proc. Eurospeech'03*, pages 2293–2296, Geneva, 2003.

[6] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz. On lexicon creation for Turkish LVCSR. In *Proc. Eurospeech'03*, pages 1165–1168, Geneva, 2003.

12.2 Word sense disambiguation using document maps

A single word may have several senses or meanings, for example "was *heading* south/the newspaper *heading* is", or "Church" as an institution versus "church" as a building. Word sense disambiguation automatically determines the appropriate senses of a particular word in context. It is an important and difficult problem with many practical consequences for language-technology applications in information retrieval, document classification, machine translation, spelling correction, parsing, and speech synthesis as well as speech recognition. For a textbook introduction, see [1]. In particular, Yarowsky [2] noted that words tend to keep the same sense during a discourse.

In [3] we introduce a method called THESSOM for word sense disambiguation that uses an existing topical document map, in this case a map of nearly 7 million patent abstracts, created with the WEBSOM method (see Section 8.4 or [4]). The method uses the document map as a representation of the semantic space of word contexts. The assumption is that similar meanings of a word have similar contexts, which are located in the same area on the self-organized document map. The results confirm this assumption. In this method, the existing general-purpose document map is calibrated, i.e., marked with correct senses, using a subset of data where the ambiguous words have been sense-tagged. The sense-calibrated map can then be utilized as a word sense classifier, for determining a probable correct sense for an ambiguous sample word in context. The data flow of the training and testing procedure is shown in Figure 12.3.

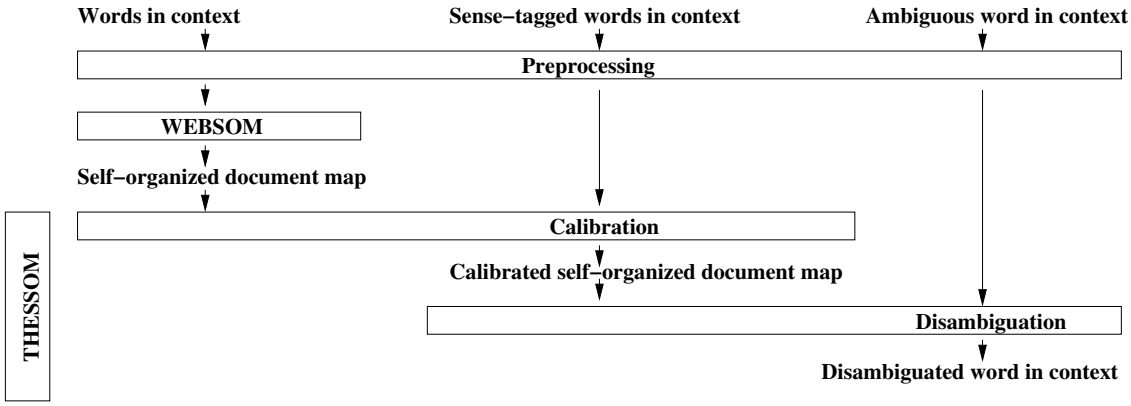


Figure 12.3: Data flow of word sense disambiguation with self-organized document maps

Results on the SENSEVAL-2 corpus (from a word sense disambiguation contest) indicate that the proposed method is statistically significantly better than the baselines, and performs on an average level when compared to the total of supervised methods in the competition. The benefit of the proposed method is that a single general purpose representation of the semantic space can be used for all words and their word senses.

In [5], instead of utilizing one general-purpose document map and merely calibrating (marking) it with particular sense locations, an individual document map is created for each ambiguous word from the training material (short contexts) for that word. Moreover, advanced linguistic analysis was performed using a dependency grammar parser to produce additional features for the document vectors. The training material consisted of a total of 8611 contexts for the 73 ambiguous words, i.e., on the average 118 contexts per word. As a result, 73 maps were generated, one for each ambiguous word.

The algorithm was tested on the SENSEVAL-2 benchmark data and shown to be on a par with the top three contenders of the SENSEVAL-2 competition. It was also shown that

adding more advanced linguistic analysis to the feature extraction seems to be essential for improving the classification accuracy. We conclude that self-organized document maps have properties similar to a large-scale semantic structure that is useful for word sense disambiguation.

References

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

[2] D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, MA, 1995.

[3] K. Lindén and K. Lagus. Word Sense Disambiguation in Document Space. In *2002 IEEE Int. Conference on Systems, Man and Cybernetics*, Tunisia, October 6–9, 2002.

[4] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.

[5] K. Lindén. Word Sense Disambiguation with THESSOM. In *Workshop on Self-Organizing Maps, WSOM'03 — Intelligent Systems and Innovational Computing*, Kitakyushu, Japan, September 11–14, 2003.

12.3 Topically focusing language model

A statistical language model provides predictions for future words based on the already seen word sequence. This is important, for example, in large vocabulary continuous speech recognition (see Section 13.2) to guide the search into those phoneme sequence candidates that constitute relevant words and sentences. Especially when the vocabulary is large, say 100 000 words, the estimation of the most likely words based on the previous sequence is challenging since all possible words, let alone all word sequences, have never been seen in any data set. For example, there exist 10^{25} sequences of 5 words of a vocabulary of 100 000 words. Thus directly estimating a n:th order markov model is generally out of the question for values of n larger than 5.

In [1] we proposed a *topically focusing language model* that is built utilizing a topical clustering of texts obtained using the WEBSOM method. The long-term dependencies [2] are taken into account by focusing the predictions of the language model according to the longer-term topical and stylistic properties of the observed speech or text.

In speech recognition suitable text data or the recognizer output can be utilized to focus the model, i.e., to select the text clusters that most closely correspond to the current discourse or topic. Next, the focused model can be applied to speech recognition or to re-rank the result hypothesis obtained by a more general model.

It has been previously shown that good topically organized clustering of large text collections can be achieved efficiently using the WEBSOM method (see Section 8.4 or [3]). In this project, the clustering is utilized as a basis for constructing a focusing language model. The model is constructed as follows:

Cluster a large collection of topically coherent text passages, e.g., paragraphs or short documents using the WEBSOM method. For each cluster (e.g. for each map unit), calculate a separate, small n-gram model. During speech recognition, use transcription history and the current hypothesis to select a small number of topically 'best' clusters. Combine the language models of each cluster to obtain a focused language model. This model is thus focused on the topical and stylistic peculiarities of a history of, say, 50 words. Combine further with a general language model for smoothing. The structure of the resulting combined language model is shown in Figure 12.4.

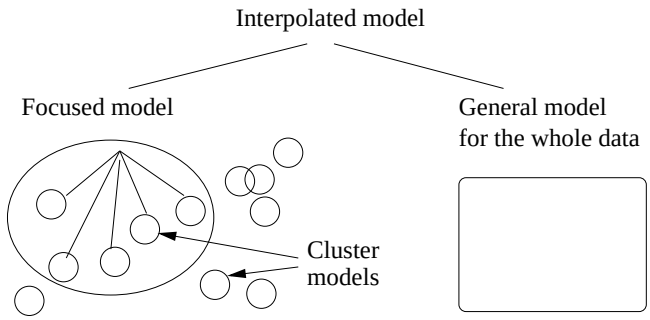


Figure 12.4: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

As the cluster-specific models and the general model we have used n-gram models of various orders. However, other types of models descriBing the short-term relationships between words could, in principle, be used as well. The combining operation amounts to a linear interpolation of the predicted word probabilities.

The models were evaluated using perplexity[1] on independent test data averaged over

[1]Perplexity is the inverse predictive probability for all the words in the test document.

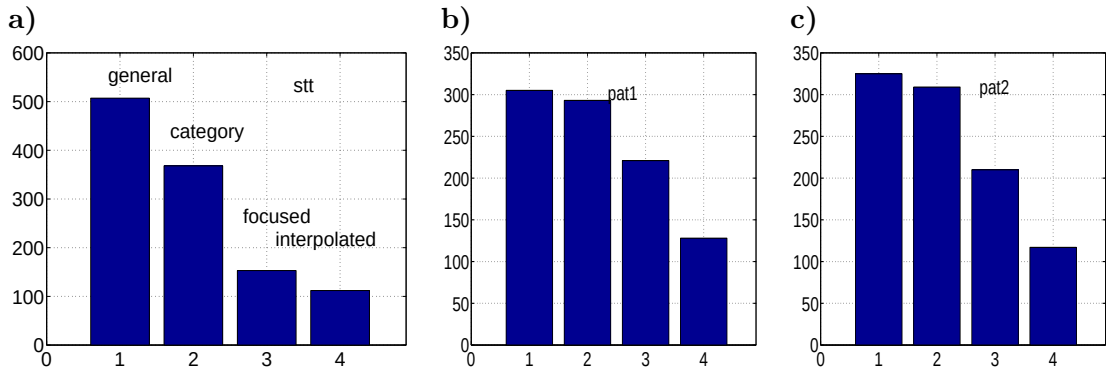


Figure 12.5: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model using prior text categories, 3. focusing model using unsupervised text clustering, and 4. the focusing model interpolated with the general model.

documents. The results for the Finnish and English text corpora in Figure 12.5 indicate that the focusing model is superior in terms of perplexity when compared to a general "monolithic" trigram model of the whole data set [4]. The focusing model is, as well, significantly better than the topic category specific models where the correct topic model was chosen based on manual class label on the data. One advantage of unsupervised topic modeling over a topic model based on fixed categories is that the unsupervised model can achieve an arbitrary granularity and a combination of several sub-topics. Finally, the lowest perplexity was obtained by a linear interpolation of word probabilities between the focusing model and the general model.

The first experiments to apply the focusing language models in Finnish large-vocabulary continuous speech recognition are reported in [5]. The results did not show clear improvements over the baseline, but by using a local LM of small but relevant text material, we see, however, that lattice rescoring can decrease the error rate. The preliminary English speech recognition tests indicate as well, that an interpolated model between a huge general LM and a small local LM performs better than the general LM alone. While there are clearly improvements to be made in language modeling, for example, to collect larger amounts of relevant text training data, maybe the most important result of the Finnish speech recognition tests is that the topical focusing works and does not slow down the whole recognition process.

References

[1] V. Siivola, M. Kurimo, and K. Lagus. Large vocabulary statistical language modeling for continuous speech recognition in Finnish. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 737–730, 2001.

[2] R.M. Iyer and M. Ostendorf, "Modeling long distance dependencies in language: Topic mixtures versus dynamic cache model," *IEEE Trans. Speech and Audio Processing*, 7, 1999.

[3] M. Kurimo and K. Lagus. An efficiently focusing large vocabulary language model. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'02)*, pages 1068–1073, Madrid, Spain, 2002.

[4] K. Lagus and M. Kurimo. Language model adaptation in speech recognition using document maps. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pages 627–636, Martigny, Switzerland, 2002.

12.4 Semantic analysis of Finnish words and sentences

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual, cognitive representations that underlie the use of language is important for applications such as speech recognition. By studying large amounts of data it may be possible to induce the conceptual, system-internal representations which provide a grounding for meanings of words.

As shown in [1,2], the self-organizing map (Chapter 8) can be applied for clustering English word forms based on the words that have appeared in their immediate contexts. In Finnish, however, the rich inflectional morphology poses a challenge as the vocabularies built on inflected word forms are typically very large. Moreover, also the inflections, some of which correspond to prepositions and function words in English, carry relevant semantic information [3]. Furthermore, the much less restricted word order compared to English is likely to cause more variation in the immediately nearby words.

In this project, we have applied the SOM algorithm to visualize and cluster common Finnish verbs based on averaged contextual features. The verb category was selected for study for two reasons: there exists a semantic reference classification of Finnish verbs [3] for comparing the results, and the semantic representation of verbs is considered an interesting problem in linguistics [3,4] because of the richness and variability of information that is connected to different verbs. In collecting information about the verbs, both morphosyntactic properties (inflections) [5] and noun base forms [6] were examined, and the resulting categorizations were compared.

An example of a map where 600 Finnish verbs were organized based on their contextual morphological features is shown in Figure 12.6. In contrast, the use of nouns as features produced for example the kinds of verb categories shown in Table 12.1.

Table 12.1: Sample verb categories based on noun categories.

Finnish verbs	Translations	Topic
myydä, ostaa, tuottaa, palkata, työllistää, kattaa, vuokrata	sell, buy, produce, hire, employ, cover, rent	business
nousta, laskea, kasvaa, pudota, vähentyä, kohota, pienentyä, supistua, noutaa, kallistua	rise, decrease, grow, fall, diminish, rise, get smaller, contract, fetch, go up in price	stock rates
kuolla, hukkua, ampua, surmata, hyökätä, loukkaantua, menehtyä	die, drown, shoot, kill, attack, get hurt, pass away	dying

The results in all the experiments show that even the simple contextual features used, collected over a large number of instances, can be suitable for obtaining automatically a semantic clustering and organization of verbs. In general, morphosyntactic properties seem to push the categorization towards the direction of linguistic semantics, while categorization based on nouns or noun categories is more a reflection of the topics of their

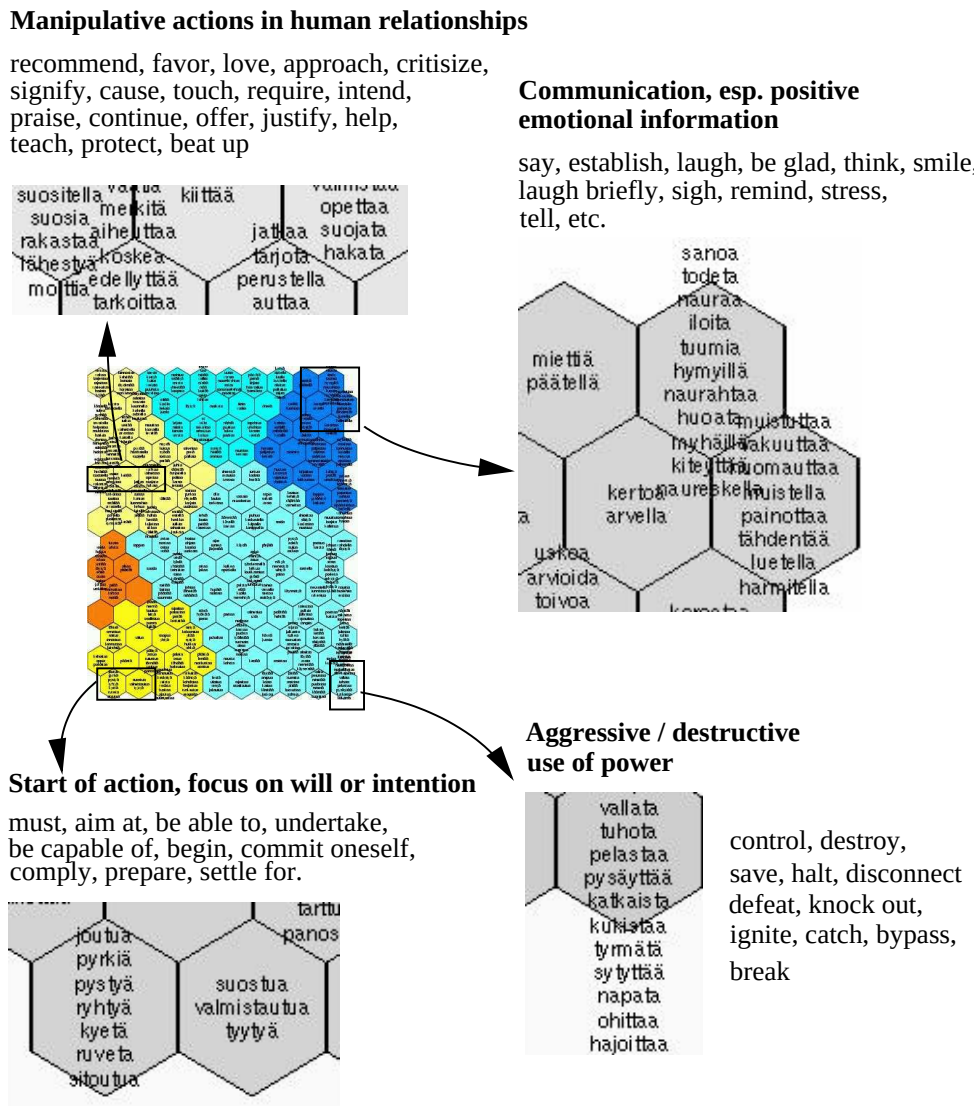


Figure 12.6: A map of the 600 most frequent verbs (base forms) in the Newspaper corpus. The verbs were organized on the basis of the distribution of morphological features in one preceding and two succeeding words, collected over all instances of the verb in any inflected form. The contents of four sample map regions are shown in the insets. In the reference classification (pp. 157–165 in [3], many of the verbs e.g. in the lower right corner indicating 'destructive use of power' are further divided into two specific categories, namely (1) break verbs (*tuhota* 'destroy', *katkaista* 'break', *hajoittaa* 'break down') and (2) fight verbs (*pysäyttää* 'stop', *kukistaa* 'defeat', *tyrmätä* 'knock out'). Similar categories can be found in [4] for English verbs.

corresponding texts. When compared to a reference classification of Finnish verbs [3], clustering shows a somewhat different perspective or world view than Pajunen's. In particular, the organization of verbs on the map reflects the importance of cultural, social, and emotional dimensions in lexical organization.

References

[1] T. Honkela. *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki University of Technology, 1997, Espoo, Finland.

[2] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 1989; 61:241-254

[3] A. Pajunen. *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* Suomalaisen Kirjallisuuden Seura, 2001.

[4] B. Levin. *English Verb Classes and Alternations: a Preliminary Investigation.* The University of Chicago Press, Chicago and London, 1993.

[5] K. Lagus and A. Airola. Analysis of functional similarities of Finnish verbs using the self-organizing map. In *ESSLLI'01 Workshop on The Acquisition and Representation of Word Meaning,* August 2001.

[6] K. Lagus, A. Airola, and M. Creutz. Data analysis of conceptual similarities of Finnish verbs. In *Proceedings of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society*, pages 566–571. Fairfax, Virginia, August 7–10, 2002.