COGSCI'02, Washington, USA, 8.-11.8.2002. p. 566-571.

Data analysis of conceptual similarities of Finnish verbs

Krista Lagus (krista.lagus@hut.fi)
Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 9800, 02015 HUT, Finland

Anu Airola (anu.airola@helsinki.fi)
Department of General Linguistics, University of Helsinki
P.O.Box 9, 00014 University of Helsinki, Finland

Mathias Creutz (mathias.creutz@hut.fi)
Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 9800, 02015 HUT, Finland

Abstract

The study of the conceptual representations that underlie the use of language is a problem motivated from both a cognitive research point of view and that of construing language models for various language processing tasks. In this work, we organized 600 Finnish verbs using the SOM algorithm. Three experiments were conducted using different features to encode the verbs: morphosyntactic properties, individual nouns, and noun categories in the context of the verb. In general, the morphosyntactic properties seem to draw attention to semantic roles, whereas nouns as features seem to highlight clusters formed on grounds of topics in the text.

Introduction

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual representations that underlie the use of language is important for applications such as speech recognition. Due to the redundancy in communication, 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. Whether this is possible, and if so, how, is an interesting and controversial question.

A central problem in learning a language or in estimating a language model[1] from data is how to generalize from particular observations to new, similar instances. Generalization requires knowledge of similarities between words, concepts and other units of language and thought, i.e., similarity representations.

The hypothesis that the semantic similarity of two words correlates strongly with the similarity of their contexts has been widely discussed in linguistics and psychology (for recent treatments, see Levin, 1993 and Miller & Charles, 1991).

It has been proposed by Gärdenfors that a central part of our conceptual reprensentations are grounded in various low-dimensional *conceptual spaces*. A conceptual space is defined as a set of quality dimensions with a geometrical structure (Gärdenfors, 2000). Examples of conceptual spaces near our perceptual apparatus are colors and the pitch of sounds. For many higher order concepts a geometric interpretation can be found, as well. For example, comparative relations such as 'longer than' can be represented as a geometric relation between two elementary length spaces. Gärdenfors proposes a subset of concepts called *natural concepts*:

> A natural concept is represented as a set of regions in a number of domains together with an assignment of salience weights to the domains and information about how the regions in different domains are correlated.

An inherent and important property of the proposed conceptual spaces is that they provide a meaning representation that is ordered and offers means for representing similarities, often in terms of some continuous-valued underlying qualities. Gärdenfors gives examples of conceptual spaces that humans are likely to have. However, an open research question remains for both brain research and the study of language use: What are the possible conceptual dimensions that humans utilize?

In this work we analyze the use of Finnish[2] verbs with the following goals in mind: (a) to uncover possible conceptual spaces, i.e., underlying, organizing semantic qualities or properties, (b) to study semantic similarities of verbs in actual language use. In particular, we examine the kinds of semantic or conceptual ordering qualities that appear to affect the distribution of features in the immediate context of a verb, in particular (1) morphosyntactic properties of nearby words, and (2) the nearby nouns and (3) unsupervised categories of nearby nouns. In effect, we rely on the redundancy in communication and assume that certain regularities observed in the distributions of verb contexts will contain significant information about the semantics of the verb as well.

[1] For an introduction to statistical language modeling see (Manning & Schütze, 1999). Their applications include speech recognition, machine translation, and dialogue agents that converse with humans in order to perform tasks such as answering questions about train schedules and booking flights.

[2] Most of the research on language is carried out using English data only, which creates a too narrow or misleading picture of the modeling apparatus underlying language learning and use.

In order to obtain a simultaneous visualization and a clustering of the verbs (and in one experiment, nouns as well), we apply the self-organizing map (SOM) (Kohonen, 1995) algorithm. The visualized ordering of the verbs is studied qualitatively to obtain an understanding of the conceptual dimensions by which the verbs are likely to be organized. Furthermore, the obtained clustering is compared to a kind of 'ground truth', namely a semantic classification of Finnish verbs suggested by Pajunen (2001).

Self-Organizing Map (SOM) algorithm

The SOM (Kohonen, 1982; Kohonen, 1995; Kohonen et al., 1996) is an unsupervised neural network method that is able to arrange complex and high-dimensional data so that similar inputs are, in general, found near each other on the map. The ordered map display can then be utilized to illustrate various properties of the data set in a meaningful manner.

The algorithm automatically places a set of reference vectors—also called model vectors—into the input data space so that the data set is approximated by the model vectors. Each reference vector corresponds to a *map unit* on a two-dimensional regular grid. In effect, the grid and the vectors form a two-dimensional 'elastic net' in the high-dimensional input space: after application of the SOM algorithm, the map follows the data in a nonlinear fashion. The algorithm simultaneously obtains *a clustering* of the data onto the model vectors and a *nonlinear projection* of the input data from the high-dimensional input space onto the two-dimensional ordered map.

Prior work on unsupervised word categorization and projection

It has been shown that distributional information of word contexts can, at least for English, be used to induce syntactic categorization, and to some degree, semantic categorization as well using various methods (cf. e.g., Finch & Chater, 1992; Charniak, 1993; Honkela, 1997; Redington, Chater & Finch, 1998). The SOM has been applied to clustering English word forms based on the word forms in their immediate contexts in (Ritter & Kohonen, 1989; Honkela, 1997). Furthermore, the word categories obtained in such a manner have been used for encoding the meaning of documents e.g., in (Lagus et al., 1996).

Various alternatives to using SOM exist, including other clustering methods such as hierarchical clustering (used e.g. in Redington, M., Chater, N., & Finch, S., 1998; Pereira et al., 1993). Moreover, different metrics or clustering criteria can be applied, such as relative entropy (e.g. in Pereira et al., 1993) or minimum description length (in Li and Abe, 1998). Moreover, projection methods that do not form clusters but only project the data into a lower-dimensional space include a number of nonlinear projection methods under the name multidimensional scaling (MDS), and linear projection methods such as latent semantic analysis (LSA). Compared to these, a particular property of the SOM is that it simultaneously forms a grouping and a nonlinear projection of the data set.

In constructing the feature vectors for each word, one may look at the occurrence of individual words, or of syntactic word categories, or of morphological or derivational features (e.g. in Light, 1996). The position in which features are examined may be defined in terms of a wide window where more distant occurrences contribute less (e.g. Gallant et al. 1992; Lund and Burgess, 1996), or looking at only specific positions near the word, or according to some grammatical relationship with the word (e.g. Pereira et al., 1993; Schulte im Walde, 2000).

Experiments

We carried out three experiments on organizing and clustering verbs using the SOM algorithm. For the verb encoding, the following types of contextual features were explored: morphosyntactic properties, individual nouns and noun categories.

Corpus and data set

A corpus consisting of 13.6 million words of Finnish newspaper text [3] was used in the experiments. The examined set of verbs consisted of the 600 most frequent Finnish verbs, returned to their base forms[4]. Each context of a verb in the corpus was examined. The context was defined as the preceding and the two following words relative to the verb.

Encoding of the verbs

Morphosyntactic properties. In the first experiment, we preferred overtly marked morphosyntactic features. The selected features were 1) case endings (see example (i)), 2) endings of the nominal forms of verbs (see example (ii)), and two closed-class parts of speech, namely 3) adverbs and 4) adpositions (see example (iii)). In addition, the features NUM (for digits), and PUNCT (for punctuation marks) were included since they can be considered as visible features, too.

In Finnish, the primary means for coding various semantic-functional dependencies in a clause is the case-marking system. The case endings are added to stems, as shown in the following example:

	Lapse-t	ajoivat	kaupunki-in.
(i)	child	drive	city
	N-PL-NOM		N-SG-ILL
	'The children drove to the city.'		

The feature set used also includes two nominal (non-finite) verb forms, namely the 1st and the 3rd infinitive. The infinitives function in a sentence as nouns:

[3] Corpus by CSC, http://www.csc.fi/kielipankki/.

[4] The morphosyntactic analysis of word forms was performed using the Conexor FDG parser for Finnish, ©Conexor Oy and Anu Airola.

(ii)	Halua-n	lähte-ä	syömä-än.
	want	go	eat
	V-PRES-SG1	V-INF1	V-INF3-ILL
	'I want to go to eat.'		

Location or movement can be coded by using an adposition (iii) instead of case endings. Adpositions include postpositions (PSP) and prepositions (PRE):

(iii)	Lapse-t	ajoivat	kaupunki-a	kohti.
	child	drive	city	toward
			N-SG-PTV	PSP
	'The children were driving towards the city.'			

The lower-level constituents, e.g., noun phrases, are predominantly head-final, the order being modifier-before-head. Adjectives agree in number and case with the head-noun when they occur as attributes (iv). Due to this redundancy, the function of a dependent noun phrase can be inferred before hearing or seeing the head of the phrase.

(iv)	Emmi	ajaa	punaise-lla	auto-lla.
	Emmi	drive	red	car
	N-PROP	V-SG3	A-SG-ADE	N-SG-ADE
	'Emmi drives in a red car.'			

At the clause level, the basic, or default, word order is subject-verb-object. According to (Hakulinen, Karlsson & Vilkuna, 1980), subjects precede finite verbs in 61% of all sentences in standard written prose. However, as Vilkuna (1989) points out, clause-level word order in Finnish shows great freedom; for example, in a simple sentence consisting of a subject, an object, a verb and one or two adverbials, all permutations are at least grammatically possible.

To sum up, we used a set of 21 mostly non-overlapping morphosyntactic properties, collected from three different textual positions relative to the verb. Each verb was thus encoded using a 63-dimensional feature vector. Averaged over a large number of samples, the value of a dimension in the vector is the conditional probability of a particular morphosyntactic feature in a particular contextual position given that verb.

Individual nouns as features. In the second experiment, instead of morphosyntactic properties individual nouns were used as features. The feature set consisted of the 10,000 most frequent nouns, returned to their base forms.

In order to keep the size of the feature vector reasonable, random projection was applied: Each of the 10,000 nouns was represented as a 500-dimensional vector, where 5 randomly selected positions were set to 1 and the values of the other dimensions were 0. The correlations thus introduced between words are in general negligible: random projection has been shown to roughly preserve distances between vectors, if the dimension of the projected vectors is sufficiently large. A theoretical treatment is presented in (Kaski, 1998); for empirical results on the use of random projection in the representation of documents, see (Kohonen et al, 2000).

As in the previous experiment, each contextual position was encoded as a separate part of the feature vector. The resulting dimensionality of the vector was 1,500.

Table 1: Sample noun categories.

Finnish nouns	Translations	Characterization
Matti, Jukka, Riitta, ...		first names of persons
maanantai, tiistai, ...	Monday, Tuesday	week days
kirja, levy, kokoelma, näytelmä, ...	book, record, collection, play	products of art
syy, pakko, tarkoitus, taipumus, ...	reason, obligation, intention, inclination	modalities

Noun categories as features. In the third experiment, noun categories were used as features. The categories were obtained by using the SOM algorithm to cluster the set of the 10,000 nouns from the previous experiment. The nouns were clustered based on verbs appearing in the same sentence at a maximum distance of five words. The position of verbs within the window was not taken into account. Verbs occurring at least 20 times in the corpus were considered as features, which yielded a total number of 3,089 verbs. Again, random projection was applied to reduce the dimensionality of the vectors.

A noun map consisting of 160 units was constructed. Each map unit was regarded as a noun category. Some examples of the resulting noun categories are shown in Table 1.

Next, the feature vectors for the 600 verbs were created. The encoding was identical to that of the first experiment except that instead of morphosyntactic properties, noun categories were used as features. The resulting feature vectors were 480-dimensional.

Creation of verb maps

In each of the three experiments, the feature vectors representing our selection of 600 frequent verbs were organized on a map of 140 units using the SOM ToolBox (Vesanto et al., 2000). As a consequence, verbs having similar feature vectors, and hopefully similar semantic representations, can be found close to each other on the map.

Results

The verb maps generated using different features were evaluated in two ways: by comparing to an

Table 2: Quantitative comparison with Pajunen's classification.

Exp. no.	Type of features	Precision
1	Morphosyntactic	35.7%
2	Individual nouns	23.6%
3	Noun categories	27.5%

existing classification, and by exploring the ordering of the verbs on the visualized maps.

Comparison to an existing verb classification

We compared the obtained clustering to Pajunen (2001), which is the most comprehensive semantic classification of Finnish verbs available. The semantic classification Pajunen presents is based both on conceptual classes, that is abstract schemas of states of affairs, and the theory of semantic fields (about field-theory, see e.g. Lyons, 1977). If compared e.g. to Levin's (1993) large-scale classification of English verbs, it can be seen that both Pajunen and Levin rely on the notion of semantic determination, i.e., the assumption that semantics determines syntax. However, while Pajunen explains form in terms of meaning, Levin (1993: 5) assumes that 'verbs that fall into classes according to shared syntactic behavior would be expected to show shared meaning components.'

Only 200 of our 600 verbs are mentioned in Pajunen's classification. These are divided into 54 classes, with 1-13 verbs per class. The comparison was carried out as follows: The set of verbs that formed a class by Pajunen were considered as correct hits for each other. For each of the 200 verbs, the map unit of that verb was examined, and the *precision* was the number of hits divided by the total number of verbs in that map unit[5].

For each experiment, the precisions, averaged over the set of 200 verbs are reported in Table 2. For the experiments 2 and 3, in which random projection was used in feature encoding, the results are averages of five runs with different random seeds. A paired t-test showed significant differences for each pair of experiments (p=0.9997 between 1 and 2, p=0.9706 between 2 and 3, and p=0.9853 between 1 and 3). The experiments, ordered by similarity to Pajunen's classification, were: morphosyntactic properties, noun categories, and individual nouns.

Visual inspection of the maps

The map obtained in the first experiment is shown in Figure 1. The organization of the map seems to highlight the importance of cultural, social and emotional aspects in lexical organization. Dimensions of social interaction, wielding of power, the will of an individual person, and manipulative behavior between people all occupy rather strong regions on the map.

[5]The verb itself was excluded, as well as all the verbs that were not mentioned by Pajunen.

Table 3: Sample verb categories based on noun categories (Experiment 3).

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, ammua, hyökätä, loukkaantua, menehtyä	die, drown, shoot, kill, moo, attack, get hurt, pass away	dying

The maps based on the distribution of individual nouns and noun categories (not shown) seem to be organized more according to a continuum from verbs describing subjective cognitive events, e.g., the states of mind of an individual, to the verbs expressing mostly social actions. There were clear categories that appeared to be clustered together on grounds of a common topic in the text. Some examples are shown in Table 3.

Note that the verb 'to moo' has apparently fallen into the 'dying' category due to an incorrect morphological analysis made by the FDG parser. The verbs *ampua*, 'to shoot', and *ammua*, 'to moo', have many common forms that are homonyms, e.g. *ammun*, 'I shoot' or 'I moo'. In fact, the FDG is reported to make 2–7 % of disambiguation errors for words in general.

Discussion

Based on visually examining the map and the clusters and a quantitative comparison to Pajunen's categories, the results are promising. It is nevertheless possible that some improvement may be achieved by (1) correcting the errors in preprocessing, (2) trying yet different types of features, feature windows, or feature encodings. Moreover, when the purpose is only to obtain individual clusters and not a visualization or projection (as done by the SOM), different clustering methods chould be examined as well.

Morphosyntactic properties appear to correspond most closely to the grounds by which Pajunen's classification is formed: both categorizations emphasize

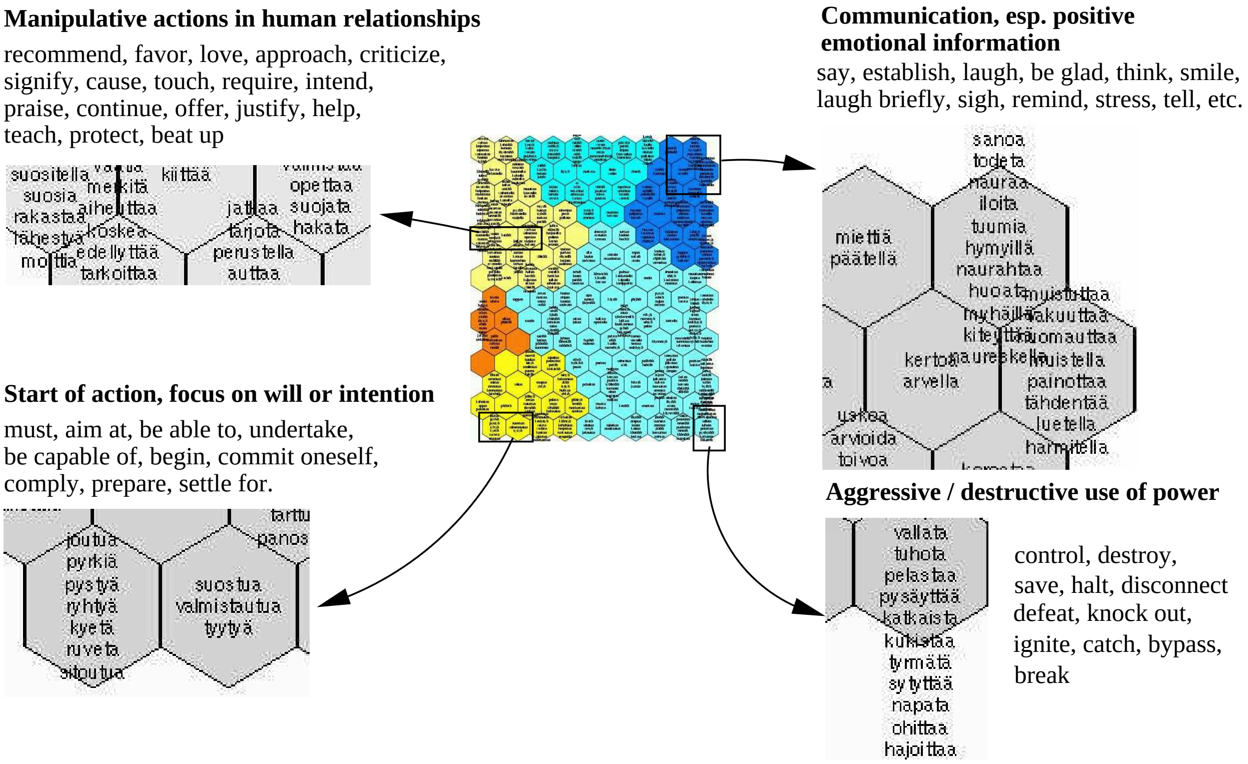

Figure 1: A map of the 600 verbs organized based on the distribution of morphosyntactic properties (non-finite forms of the verb were excluded). Properties of the preceding word and the two following words were considered. The contents of four sample map regions are shown in the insets. Many of the obtained categories correspond to categories defined by Pajunen. However, in Pajunen's classification the verbs 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 (Levin, 1993) for English verbs.

the semantic roles a noun phrase may bear in its clause. This seems reasonable, since in Finnish, the primary means to express semantic roles (e.g. AGENT and PATIENT) is the case system. Even though semantic roles cannot be simply derived from morphosyntactic cases, a strong correlation can still be assumed.

The other feature types, i.e. individual nouns and noun categories, exhibit fairly similar information regarding the 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 subject matters communicated through texts.

In some cases, the morphosyntactic map distinguishes verbs based on the *kind* of the patient (upper left corner with human→{human} relationships vs. lower right corner with human→{nation, abstraction} relationships). This result confirmed our expectations, and the understanding that the type of consequence of the action for the patient is not reflected in the morphological features. On the other hand, the maps based on noun features seem to make distinctions based on both the topic, the consequences for the patient {dying, creation, change of possession, change of state}, and the kind of patient {human, food, artefact}. In this way, the different types of features highlight different relevant aspects of categorizing verbs.

Conclusions

Different feature selections correspond to different assessments of what is important in the categorization of verbs. The categorization most similar to Pajunen's was obtained with morphosyntactic features. In contrast, it appeared that the noun features bring out the similarities between verbs in a richer and more useful manner.

It is interesting to consider whether the ordering qualities observed on the map could count as quality dimensions of the conceptual spaces suggested by

Gärdenfors. The observed ordering qualities seem to reflect various higher-level cognitive, emotional, or social dimensions. These could be emergent properties of representations on some lower, more basic level. In fact, the emergent properties of one representation level or process are likely to be used as input features of another level or process.

By looking at the maps it seems clear that there are many relevant aspects for categorizing verbs, and any single categorization or ordering is of necessity reduced into considering only some of these. However, to obtain a more accurate representation, instead of a single categorization or projection, one should create several simultaneous categorizations (representations), induced using different kinds of features.

References

Charniak, E. (1993). *Statistical Language Learning.* MIT Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., and Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Finch, S., & Chater, N. (1992). Unsupervised methods for finding linguistic categories. In Aleksander, I., & Taylor, J. (Eds.), *Artificial Neural Networks, 2*, pp. II–1365–1368. North-Holland.

Gallant, S. I., Caid, W. R., Carleton, J., Hecht-Nielsen, R., Pu Qing, K., and Sudbeck, D. (1992). HNC's MatchPlus system. *ACM SIGIR Forum*, 26(2):34–38.

Gärdenfors, P. (2000). *Conceptual Spaces.* MIT Press.

Hakulinen, A., Karlsson, F., & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus.* Publications of the Department of General Linguistics, Yliopistopaino, Number 6, Helsinki.

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

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proc. COLING-00*, pp. 747–753.

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of IJCNN'98, Intl Joint Conference on Neural Networks*, vol. 1, pp. 413–418. IEEE Service Center, Piscataway, NJ.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biol. Cybern.*, 43(1):59–69.

Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin, 3rd edition 2001.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1996). SOM_PAK: The Self-Organizing Map program package. TR A31, Helsinki University of Technology, Laboratory of Computer and Information Science.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, V., & Saarela, A. (2000). 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.

Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In Simoudis, E., Han, J., & Fayyad, U., (Eds.), *Proc. KDD-96*, pp. 238-243. AAAI Press, Menlo Park, CA.

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

Li, H. and Abe, N. (1998). Word clustering and disambiguation based on co-occurrence data. In *36th Annual Meeting of the ACL, COLING-98*, pp. 749–755.

Light, M. (1996). Morphological cues for lexical semantics. In *ACL 34*, pp. 25–31.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

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

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Pajunen, A. (2001). *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* SKS, Helsinki.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of English words. In *30th Annual Meeting of the ACL*, pp. 183–190.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biol. Cybern.*, 61:241–254.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). Som toolbox for matlab 5. TR A57, Helsinki Univ. of Technology, Neural Networks Research Centre, Espoo, Finland.

Vilkuna, M. (1989). *Free Word Order in Finnish. Its Syntax and discourse functions.* SKS, Helsinki.