Chapter 1

SEMANTIC CLUSTERING OF VERBS

Analysis of Morphosyntactic Contexts Using the SOM Algorithm

Krista Lagus

Helsinki University of Technology, Neural Networks Research Centre
P.O.Box 5400, FIN-02015 HUT, Finland
Krista.Lagus@hut.fi

Anu Airola

University of Helsinki, Department of General Linguistics
P.O.Box 9, FIN-00014 University of Helsinki, Finland
Anu.Airola@helsinki.fi

Abstract Obtaining semantic or functional word categories from data in an unsupervised manner is a problem motivated both from the linguistic point of view and from that of construing language models for various language processing tasks. In this work, we use the self-organizing map algorithm to visualize and cluster common Finnish verbs based on functional and semantic information coded by case marking and function words like postpositions and adverbs. Firstly, based on a data set of over 500,000 utterances of 25 verbs, we studied (a) the base forms and (b) the most common word forms of the same verbs (4764 forms). Secondly, the first experiment was repeated on a set of 600 verbs. The results show that even the simple feature selection used in this experiment was found to be suitable for rough automatic categorization of verbs on the basis of data extracted from unrestricted texts. In particular, the results demonstrate the importance of cultural, social and emotional dimensions in lexical organization.

Keywords: distributional clustering of words, semantic representations, data exploration and visualization, self-organizing map

1. Introduction

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 a recent treatment, see Levin, 1993 and Miller and Charles, 1991). For the acquisition of information of semantic similarities of words, the crucial question then is what kind of information is relevant when a system forms contextual representations for words.

As pointed out in (Redington et al., 1998), distributional information of words is not the only source that is available to humans, and it is likely that all relevant information can not be obtained from a single source. E.g., in spoken language, other possible sources of information include prosody and semantic information from other perceptual modalities. However, since there is plenty of textual data available, and much less multimodal data of natural conversations, it remains an interesting question how much and what kind of information can be aquired from distributions of contextual features in text alone.

In actual language use there is redundancy due to the fact that several elements may contribute to communicating similar content. For example, in the sentence *I went to the store* both *went* and *to* appear to signal the presence of movement. In natural communication the redundancy is necessary for successful communication especially in environments where parts of the signal are corrupted by noise, or when it is possible that the speakers' vocabularies or otherwise their knowledge of language differs. For example, there may be subtle differences in the perceived meaning or the connotation of a word. The redundancy in expressions is useful since it provides several clues to what the speaker has in mind. Morever, in less noisy situations such redundancy allows one to learn about language itself, for example, to deduce similarities between elements of language based on distributions of contexts of their usage.

As has been shown previously, in languages with rather rigid word order, such as English, the distribution of words in the immediate context of a target word contains a considerable amount of information regarding the syntactic category of the target word, and to some degree, also semantic grouping can be observed (see Ritter and Kohonen, 1989; Finch and Chater, 1992; Zavrel and Veenstra, 1995; Redington et al., 1998; for an overview of early work, cf. Charniak, 1993).

However, for some other languages, such as Finnish, the same approach does not seem directly feasible. For example, the kind of functional information that in English is coded by independent function words such as "to" and "at" is in Finnish coded by inflectional morphology (see Section 2 for details). The result is a very large vocabulary of word forms, which causes immediate problems for reliable estimation of contextual models. Note that if words are used as features (as in all the previously cited works) the doubling of the number of words simultaneously doubles the number of features as well. Moreover, with Finnish the effect is much stronger than a doubling—a single base form of a verb or a noun may have thousands of inflected forms (Karlsson, 1983).

The efficient representation of syntactic and semantic properties of words in a language model is an important issue for building natural language applications. For example, statistical language models are models that are estimated based on large corpora and that allow one to make predictions regarding future linguistic data (e.g., future words) given some properties of previous data (e.g., previous words). Such models are applied, e.g., in speech recognition. Surprising success has been achieved with very simple models, such as N-gram models (cf. Manning and Schütze, 1999; Jurafsky and Martin, 2000) that record frequencies of all encountered combinations of n words and attempt to predict the next word based on the previous $n-1$ ones. However, these typically require vast amounts of learning data, and even then, some quite ordinary word combinations have not appeared in the data even once (e.g. Charniak, 1993).

If a word combination has not appeared in the training data, a straightforward statistical method assumes it to be extremely unlikely (generally as unlikely as any syntactically or semantically impossible combination). Syntactic and semantic categorizations may be used to ameliorate this problem. For example, let us say that we have categorized words such as *run*, *walk*, *stroll*, *dash* etc. into the same verb category. Upon observing the phrase *John runs*, we would update our belief about the likelihood of the phrase *John runs* but also of *John walks* as well as all the combinations of John and another verb in the same category as *run*. In statistical terms, we would be doing model smoothing based on semantic word categories. Model smoothing strategies help in the problem of sparsity of the data. However, as a result the likelihood estimates may as well deteriorate, and thus the task of uncovering suitable categories remains a challenge.

Examining the similarities in the use of words is even more interesting from a cognitive point of view. It seems a reasonable hypothesis that some general perceptual, cognitive, emotional, communicative, social etc. dimensions underlie our use of language, and that they must in some way be reflected in the choices of words and other features of language use. A detailed hypothesis regarding such underlying dimensions, or *conceptual spaces* has been proposed in (Gärdenfors, 2000). Even given that this hypothesis is correct, it remains to be seen (1) exactly what kind of dimensions there are, and (2) which linguistic features reflect information regarding them. We consider this an empirical question that is best answered by statistical data analysis methods. More specifically, the following question is addressed here: How much and what kind of information regarding semantic similarities of verbs can, in principle, be uncovered by assuming very little, i.e., by selecting the verbs to be studied based on frequency, and by looking only at features in their immediate contexts in a very simple and straightforward manner. In order to obtain a view that is as little affected by existing linguistic theories as possible, we try to assume very little from the linguistic analysis. It is nevertheless assumed that the system is already capable of distinguishing verbs from other syntactic word classes, and able to recognize some morphological features.[1]

We apply the self-organizing map (SOM) (Kohonen, 1982; Kohonen, 1995; Kohonen et al., 1996) algorithm for simultaneous clustering and visualization of the data.

1.1 Structure of the article

Section 3 gives a brief introduction to the Self-Organizing Map algorithm.

In section 2 we describe some properties of Finnish morphosyntax that are relevant to the experiments.

Next, three different experiments are discussed. In the first experiment in section 4, each of the 25 verbs is examined as a distinct equivalence class. By equivalence class we mean that each different form of the same verb is treated alike. In particular, we look at the immediate right context of the verb and average the contextual information over all instances of that verb in a corpus, resulting in a single feature vector. Then we utilize the SOM for clustering and visualizing these vectors.

The second experiment described in section 5 complements the first by focusing on the *variation in behavior* between the different forms of each verb. According to our hypothesis different forms of the same verb have somewhat different behaviour profiles. The purpose of the second experiment was thus to examine these differences.

The third experiment in section 6 is basically a repetition of the first one, but on a larger set, namely 600 verbs. Moreover, in the third experiment different context window widths were tried.

Finally, conclusions are presented in section 7.

2. On Finnish morphosyntax

The meaning-bearing elements in language expressions are often described as constructions consisting of the following elements: the words themselves (representing lexical concepts), morphological endings and markers, the order of words, and the intonation (see e.g. Goldberg, 1995; Tomasello, 1998; Croft, 1999).

In Finnish, the primary means for coding various grammatical relations between the verb and the nominal constituents in a clause is the case-marking system (see Table 1.1). The case endings are added to stems, as shown in (1), where the subject is marked by the nominative, and the end point of the movement indicated by the verb *ajaa* 'to drive' is marked by the illative.

(1)	Lapse-t	ajoivat	kaupunki-in.
	child	drive	city
	N-PL-NOM		N-SG-ILL

'The children drove to the city.'

In addition to the so-called local cases, location or movement can be coded by using an adposition (2), too. Adpositions include postpositions (PSP) and prepositions (PRE). They take a complement either in the genitive or in the partitive.

(2)	Lapse-t	ajoivat	kaupunki-a	kohti.
	child	drive	city	toward
			N-SG-PTV	PSP

'The children were driving towards the city.'

Adpositions, adverbs, and local cases are grammatical categories that are at least partly functionally overlapping. However, as examples (1) and (2) show, they construe the situation differently. The two clauses show an aspectual opposition; the illative in (1) indicates that the action is completed, while the postposition *kohti* 'towards' in (2) expresses duration without specifying completion.

Finnish has an extensive system of nominal (non-finite) verb forms, including four infinitives. The infinitives function in a sentence as nouns (3).

(3)	Halua-n	lähte-ä	syö-mä-än.
	want	go	eat
	V-PRES-SG1	V-INF1	V-INF3-ILL

'I want to go to eat.'

In Finnish, the phrase-level constituents, e.g., NPs (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 (4). Due to this redundancy, the function of a dependent NP can be inferred before hearing or seeing the head of the phrase.

Table 1.1. The Finnish case system.

Case		**Endings**	**Meaning**	**Example**
		GRAMMATICAL CASES		
nominative	NOM	- (pl. -t)	basic form	auto 'car'
partitive	PTV	-a/-ä; -ta/-tä;	indefinite	maito+a '(some) milk'
		-tta/-ttä	quantity	vet+tä '(some) water'
genitive	GEN	-n; -den, -tten	possession	auto+n 'of the car'
		LOCAL CASES		
inessive	INE	-ssa/-ssä	inside	auto+ssa 'in the car'
elative	ELA	-sta/-stä	out of	auto+sta 'out of the car'
illative	ILL	-Vn,	into	auto+on 'into the car'
		-hVn,		maa+han 'into the country'
		-seen, -siin		Lontoo+seen 'to London'
adessive	ADE	-lla/llä	on; instrument	pöydä+llä 'on the table'
ablative	ABL	-lta/ltä	off	pöydä+ltä 'off the table'
allative	ALL	-lle	onto	pöydä+lle 'onto the table'
		OTHER CASES		
essive	ESS	-na/-nä	state	opettaja+na 'as a teacher'
translative	TRA	-ksi	change of state	opettaja+ksi '(become) a teacher'
comitative	COM	-ine-	accompanying	vaimo+ine+ni 'with my wife'
instructive	INS	-n	(idiomatic)	jala/n 'on foot'

Cf. Karlsson, 1987: 22–23.

(4)	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.'			

In the clause-level, the basic, or default, word order is SVO. According to (Hakulinen et al., 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.

3. 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.

3.1 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 (Finch and Chater, 1992; Charniak, 1993; Honkela, 1997; Redington et al., 1998). The SOM has been applied to clustering English words based on the words in their immediate contexts in (Ritter and Kohonen, 1989; Honkela, 1997). The word categories obtained in such a manner have been used for encoding the meaning of documents e.g., in (Kaski et al., 1998).

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, 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 multi-dimensional 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 visualization of the data set.

4. Experiment 1: A map of 25 Finnish verbs

For this study, we selected 25 frequent Finnish verbs, shown in Table 1.2. In particular, we were interested in whether the contextual window of only one word and the set of morphological features would contain sufficient information for obtaining automatically an interesting sematic ordering or clustering for the verb base forms. As a basis for comparison, we used Pajunen's verb classification (see Table 1.2).

4.1 Selected verbs and the corpora

All the occurrences of the verbs, a total of about 500,000 samples, were extracted from two sub-corpora of written standard Finnish (CSC, 2001), namely 'Newspapers' (13.6 million words) and 'Books' (4.1 million words). The morphological analysis of the data was conducted by the Conexor FDG, a functional dependency parser for Finnish (Tapanainen and Järvinen, 1997)[2]. No effort was made to detect possible parsing errors. Furthermore, the verbs were not disambiguated semantically.

4.2 Contextual morphological features

To describe a verb we collected features from a window of a single word to the right of the target word, regardless of the syntactic function of the contextual word. The following morphosyntactic properties were included: 1) a set of case endings (Table 1.1), 2) adpositions, i.e., postpositions and prepositions (see example 2), 3) adverbs, and 4) two nominal forms of verbs (i.e., the 1st and the 3rd infinitive, see example 3). Thus, the only information the system 'knows' is whether the next word immediately to the right of the target word bears some of the selected features.

In selection of the feature set we preferred overtly marked features that can be identified by a system able to carry out some segmentation. These include case endings and the endings of the nominal forms of verbs. Moreover, by including 'adverb' and 'adposition' as features we assumed that the system is already capable of distinguishing parts of speech.

4.3 Formation of the feature vectors

Let V be the set of 25 verbs, $v_i \in V$, $i = 1 \ldots 25$; and M the set of 18 morphological tags, $m_j \in M$, $j = 1..18$. In the data vector $\mathbf{x}_i$ for verb v_i a separate input dimension is reserved for representing each morphological tag. The value of $x_{i,j}$ is the frequency of context $v_i m_j$ in the corpus normalized by the frequency of verb v_i. In effect, $\mathbf{x}_i$ contains estimates of the conditional probabilities $P(m_j(t+1)|v_i(t))$ for each j (t is an index over the sequence of words). In other words, the vector $\mathbf{x}_i$ describes how typical each of the morphological tags is in the word following the verb v_i.

4.4 The organization and clustering of the data

After forming the data vectors, various data analysis methods can be applied. We organized the verb representations automatically with the self-organizing map (SOM) (Kohonen, 1995) algorithm. The SOM is

Table 1.2. Baseline categorization.

Verb	**freq.**	**Translation**
1. PSYCHOLOGICAL STATES AND PROCESSES, MODAL VERBS		
alkaa	17,987	begin, start, commence
haluta	14,637	want; wish; like (how do you like your tea); care for (would you care for a cup of tea?)
nähdä	14,746	see, spot
pitää	38,299	I) hold, keep, retain II) must, have to, be compelled to, be obliged to, shall; be supposed to
saada	57,815	get; receive; may, can, be allowed to, be permitted to; have to
tietää	12,678	know, be aware of, be conscious of
voida	46,474	can, be able, be capable, may
2. MOTION VERBS		
ajaa	6,372	drive; ride; run
käydä	14,447	go; walk
laskea 1	5,150	I) tr. lower, let down, haul down II) itr 1. fall, decline, go down, drop, decrease, diminish
laskea 2		count; calculate
lähteä	12,921	go, leave
mennä	18,486	go
nousta	10,889	rise; go up; ascend; climb; arise
päästä	14,809	a) get (into/out of); arrive, reach b) be let (into a room); be allowed (to enter); be admitted
tulla	43,972	come; arrive; become, get, turn, grow into
tuoda	6,840	bring
3. SPEECH ACT VERBS		
kertoa	21,474	tell; narrate, relate, recount; inform
puhua	10,389	speak, talk; tell; say; discuss
sanoa	43,015	say
vastata	7,554	answer, respond, reply, give an answer, make a reply, counter
4. ACTION VERBS		
käyttää	12,035	use (a knife)
tehdä	27,195	do; perform; commit; make
5. VERBS OF POSSESSION		
antaa	14,269	give; let; allow
kuulua	12,425	belong; be heard, be audible
ottaa	18,501	take

The verbs are divided into five categories according to Pajunen's ontological-semantic verb classification (Pajunen, 1999; Pajunen, 2001). The numbers after the verbs show the frequency of occurrences of the verb in the studied corpora. Translations are from (Hurme et al., 1984). Note that many of the verbs are polysemous and different senses should, in fact, be placed into different categories.

an unsupervised neural network algorithm (Haykin, 1999) that is able to arrange complex and high-dimensional data so that similar inputs are, in general, found near each other on the map. Furthermore, the organized map display can 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. The model vectors are constrained to a (usually two-dimensional) regular grid that in effect forms an 'elastic network': after application of the SOM algorithm, the map follows the data in a nonlinear fashion. The algorithm obtains simultaneously *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 lattice formed by the model vectors.

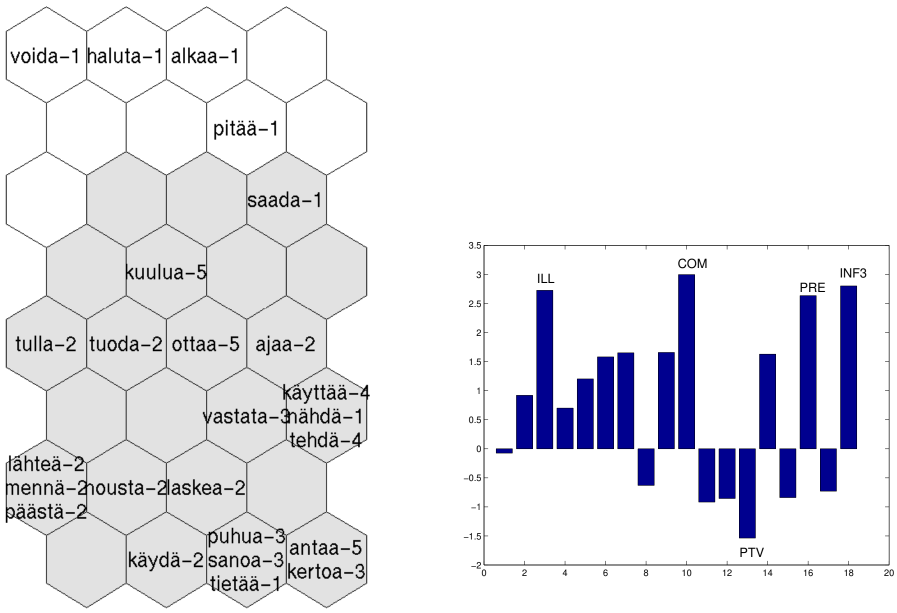
The SOM ToolBox (Vesanto et al., 2000) program package was utilized to obtain the clustering and the visualizations. Default values were used for SOM learning parameters. K-means[3] was utilized for clustering the map vectors after creation of the map. The K-means was applied in order to use the SOM also for displaying some of the more coarse-grained cluster structure that lies on top of the more fine-grained distribution of the data into individual map units.

4.5 Results

For each verb, the best-matching map unit (bmu) was found by locating the model vector closest to the verb vector. The verb was then written onto the bmu on the map lattice in Figure 1.1. The data vectors of verbs related to two selected map units are visualized in Figure 1.2. The values of individual features have been plotted on the map in Figure 1.3.

When comparing the obtained organization to Pajunen's ontological-semantic verb classification, there seems to be a reasonable amount of agreement: verbs with the same category number are rather often found near each other on the map display.

There are also some interesting differences between Pajunen's classification and the emergent ordering. For example, in the lowest part of the map, there is a map unit of three verbs, namely *puhua* 'to speak', *sanoa* 'to say', and *tietää* 'to know'. However, in the baseline categorization, *puhua*, *sanoa*, and *tietää* are placed in two different categories, namely a) speech act verbs, and b) verbs indicating psychological states and processes[4].



(a) (b)

Figure 1.1. Analysis of the 25 verbs. (a) The SOM lattice where each hexagon marks a map unit. The map units belonging to cluster 2 are shaded (bottom of the map). The best-matching unit, bmu, has been automatically searched for each of the 25 verb vectors, and each verb has been written on its bmu. For example, the bmu for the verbs *lähteä, mennä,* and *päästä* was the unit near the lower left corner of the map. The number after the verb refers to the verb class of Table 1.2. (b) The difference between the map unit of verbs *lähteä, mennä, päästä* and the rest of the map for each input feature. The values are normalized with standard deviation of the feature. If the value for a feature (say COM) is above zero, that feature is more frequent in this map unit than in general, and below zero means that it is here less frequent than in general. The features that most distinguish this unit from other units are ILL, COM, PRE, INF3, and the lack of PTV.

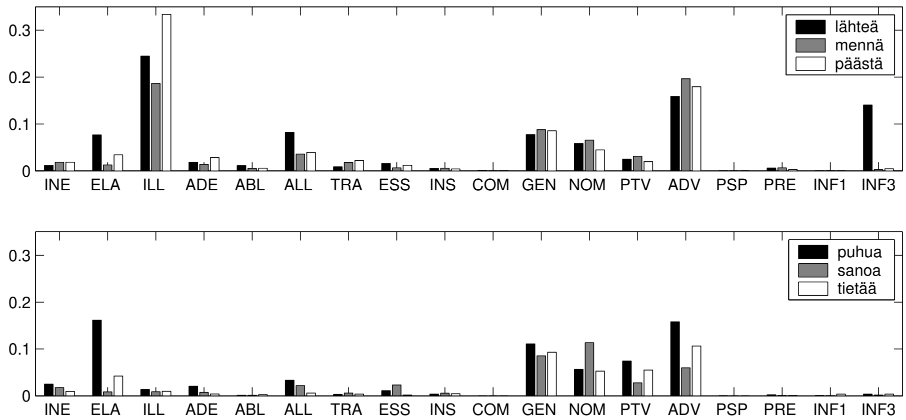


Figure 1.2. Data vectors for verbs in two different map units: the value for each morphological form is visualized as the height of the bar.

While Pajunen starts by defining ontological-semantic classes and continues by placing verbs into them, our method operates on low-level morpho-syntactic features and as a result, implicit categories emerge. Concerning the cluster with *puhua-sanoa-tietää*, the common dimension seems to be related to possessing information. On the sentence level, this means that the verbs *puhua*, *sanoa*, and *tietää* prefer next to their right phrases that indicate either the CONTENT (see a below) or the SOURCE (see b below) of the message.

(a) Hän oli aivan oikeassa ja *tiesi* **mitä puhui**. (CSC, 2001:Demari 1995.)
"She was absolutely right and *knew* **what she talked about**."

(b) "Täällä voin antaa tyttärelleni hyvää ruokaa", *sanoo* **Sergei Agalakov**. (CSC, 2001:Keskisuomalainen 1995.)
"Here I can give my daughter good food", *says* **Sergei Agalakov**.

The neighboring unit on the lower right corner of the map is occupied by two verbs, i.e. *antaa* 'to give', and *kertoa* 'to tell'. Pajunen places *antaa* as a 'verb of giving', while *kertoa* in 'verbs of saying'. However, these verbs share a more abstract cognitive property, namely *change of possession*. In the case of *antaa* 'to give', the object of possession may be either concrete or abstract, e.g. information like for the verb *kertoa*, too (see c and d below).

(c) Junan kuljettaja oli havainnut radalla olevan miehen ja *antanut* **äänimerkin**. (CSC, 2001:Iltalehti)

"The train driver had noticed the man on the track and *given* **a signal**."

(d) Pienet vauvat eivät tietenkään voi *kertoa* **lihaskivuistaan tai muista oireistaan**. (CSC, 2001:Iltalehti)
"Of course, small babies are not able to *tell* **about their muscular pains or other symptoms**."

Furthermore, it is interesting to observe that the above discussed two map units are next to each other on the map, i.e., there is a gradual change from possessing information to the change of possession.

Pajunen's classification exhibits a certain ontological view of the world. As the comparison with our experiment suggests, when one concentrates on what is typical on actual language use, an alternative perspective emerges. Moreover, our results seem to be in accord with one of the iconic principles that is formulated e.g. by Givón (2001: 35) as follows: "information chunks that belong together conceptually are kept in close spatio-temporal proximity."

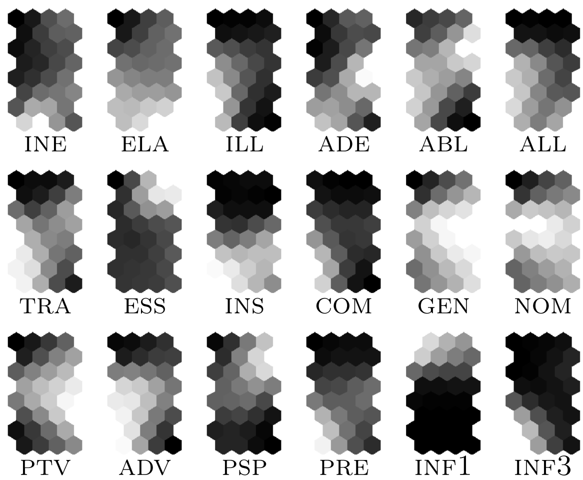


Figure 1.3. The values of individual features have been visualized on the map—bright colors signal large values. For example, the most prominent region of the feature INF3 is the lower left corner of the map. By comparing the visualizations of two features, dependencies between them can be identified. For example, features COM, PRE, and INF3 seem strongly correlated.

5. Experiment 2: A map of 4764 verb forms

The purpose of the second experiment was to study more closely the 21 verbs in the second cluster (the lower part of the map obtained in the first experiment, see Figure 1.1). In particular, of interest were the various inflected verb forms, since it is to be expected that different forms of the same verb may have quite different behavior profiles.

5.1 Formation of the feature vectors and the clustering

The data vectors were formed similarly as in experiment 1, the only difference being that a vector was now collected per each verb form. To avoid unreliable estimates, verb forms with fewer than 20 occurrences were discarded. As a result, 4764 data vectors were obtained. Next, a map of 11×14 units was automatically created with the SOM ToolBox. Best-matching units were then located for each data vector.

5.2 Results

To obtain an overall view of the map organization of the verb form map, we plotted all the occurrences of verb forms of a single base form onto a single map display. Several examples are depicted in Figure 1.4. Figure 1.5 shows more closely the behavior of different verb forms, such as the contrast between finite and certain non-finite forms of verbs.

The spread of a verb shows how coherently the forms of that verb behave in terms of the applied feature selection. By comparing spreads of two different verbs, say *mennä* and *lähteä*, one quickly obtains some overview of the differences and similarities between the verbs—in this case, that the verbs are rather similar. However, a difference is implied since the verbs fall on adjoining regions instead of overlapping.

By looking at the values of the morphological features on the map (not shown here) it can be seen that the feature INF3 (3rd infinitive) is especially strong on the region overlapping with the verb *lähteä* but not on the adjoining region of *mennä*. The illative form of the 3rd infinitive can be used after all motion verbs, indicating an action which is about to begin, as in (5):

	Lähden / Menen	nukku-ma-an.
(5)		V-INF3-ILL
	'I'll go to sleep.'	

However, as the experiment shows, the 3rd infinitive illative is particularly typical of the verb *lähteä*. In contrast, the verbs *mennä* and *päästä* are used more often to indicate a 'pure' change of location. The difference in use is depicted in Table 1.3.

Intuitively, many Finnish speakers would probably consider the verbs *lähteä* and *mennä* as synonyms at least in some textual contexts. However, as our data reveals, there nevertheless appears to be differences in

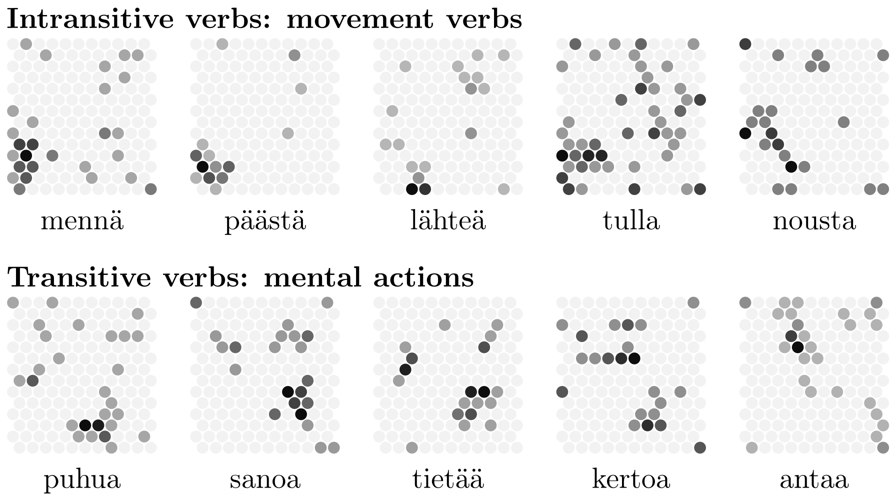


Figure 1.4. Verb distributions on the verb form map. The first plot depicts the occurrences of verb forms with the base form *mennä*—dark colour corresponds to a large number of occurrences. Each plot can be viewed as a "higher-order fingerprint" of the verb—similar fingerprints correspond to similar distributions of word forms in terms of their morphological contexts. The plots on the first row depict verbs that were found on the lower left side of the map in the previous experiment (see Figure 1.1), whereas the second row consists of the verbs in the lower right corner of the same map.

Table 1.3. A particular finding of Experiment 2.

Going verb	***lähteä***	***mennä***
Possible complements	3rd-INF-ILL NP-ILL	3rd-INF-ILL NP-ILL
Findings	**3rd-INF-ILL**	**NP-ILL**
Example	Lähden mene-**mään.** 'I'll start to go.'	Menen koti-**in.** 'I'll go home.'
Semantic property of the verb-complement construction	Beginning of an action or a process	Change of location

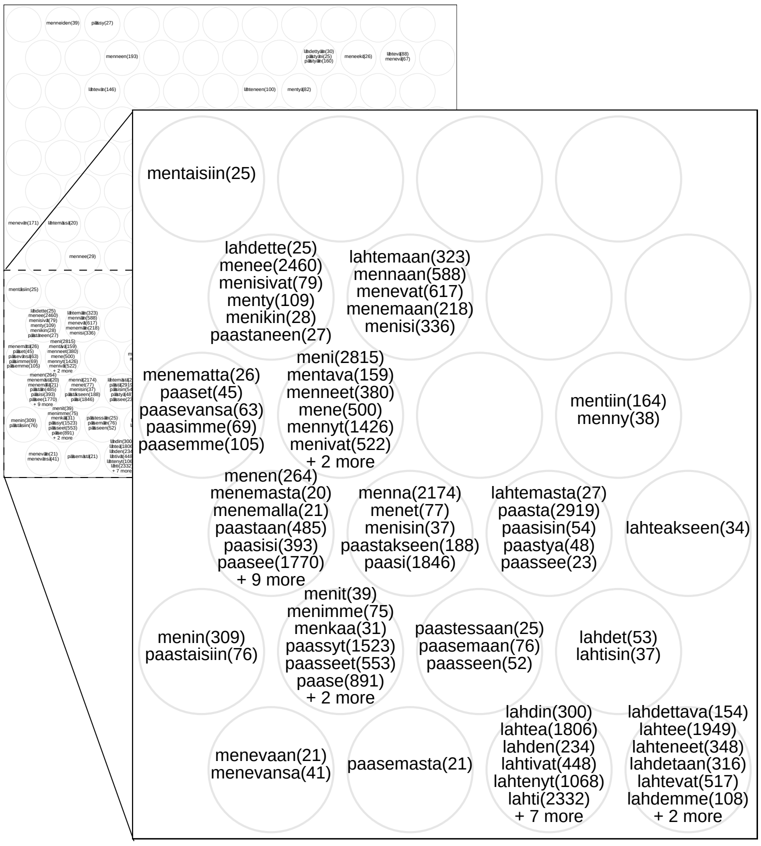


Figure 1.5. A close-up on the lower left corner of the map. The individual verb forms whose base form was *lähteä* 'to leave,' *mennä* 'to go,' or *päästä* 'to get into/out of' have been plotted on the map to their best-matching units (the scandinavian characters *ä* and *ö* are shown here as *a* and *o*). The numbers of occurrences of each form are shown in parenthesis. The verb forms in this region are mostly finite forms of the verbs. They seem to be in contrast to the non-finite forms, which appear as another, weaker cluster in the upper right corner of the map (not visible).

their contexts of actual usage. E.g. the verb *lähteä* is used more often to signal a following action or a process (indicated formally by the following word being in the 3rd infinitive in the illative) whereas *mennä* is used more for a change of location (indicated by an NP in the illative), and less for an action or a process. This is an interesting finding because e.g. according to a former study by (Ikola et al., 1989), in standard written Finnish, the two verbs seem to have had quite similar behavioral profiles at least in relation to the following 3rd infinitive in the illative. Moreover, the intuitions of different language speakers seem to differ regarding this finding. In particular, some older Finnish speakers find the so called inchoative[5] use of the verb "lähteä" quite strange in expressions like e.g. "lähteä menemään" 'to start to go.'

6. Experiment 3: A map of 600 Finnish verbs

6.1 Data and features

The experiment on verb base forms in Section 4 was repeated on a larger set of verbs, that is, the 600 most frequent Finnish verbs in the 'Newspaper' sub-corpus (13.6 million words). The data were extracted from the same corpus.

The original set of features was complemented with three additional features that can be considered as visible features, namely NUM (for digits), PUNCT (for punctuation marks), and ABE (for the abessive). The abessive is a rather rare case meaning 'without' and marked with '-tta/'-ttä'. Usually it is replased by the preposition *ilman*, 'without' (*Hän lähti raha-tta/ilman rahaa* 'He left without money').

In this experiment, different context window widths were tried, up to three words on both sides of the verb. In collection of the contextual information, sentence borders or punctuation marks were not crossed. The feature vectors were formed as before, the value for each dimension being the conditional probability of a feature in a specific position with regard to the verb.

6.2 Creation of the map

A number of SOMs were created where the context window and the SOM learning parameters were varied.

The results were evaluated by comparing the baseline classification to the obtained clustering. Each map unit was taken as a cluster, and the maps were compared to baseline classification using measures adopted from the information retrieval community, namely *precision* (or accuracy) and *recall* (or completeness)[6]). The best map, shown in Figure 1.6, was then selected for closer, qualitative examination.

On the selected map the context width was one previous word and two succeeding words. Each context position corresponded to an individual segment of 21 features in the data vector, resulting in a vector length of 63. The size of the map was 14×10 units. Also 3-dimensional SOM lattices were experimented with, but no clear improvement was observed in this case. SOM Toolbox was again utilized for construction of the maps, with the following learning parameters. Gaussian neighborhood function was utilized to achieve good global organization of the map. The length of neighborhood radius was reduced from 6 to 1 and then from 1 to 0.9 with corresponding training lengths of 10 and 40 epochs. The rather slow training schedule was applied to obtain maps that are near convergence. Apparent convergence was confirmed by observing the quantization error on the training data.

6.3 Results

The obtained map seems to depict a kind of macrolevel perspective on the organization of the Finnish verbal lexicon. Especially, the dimensions of social interaction, the wielding of power, the will of an individual person, and the manipulative behavior between people all occupy rather strong regions on the map (see 1.6). Moreover, the ordering seems to reflect the emotional load the lexical items carry, as exemplified by the regions found on two corners of the map. On the upper right corner there is a region occupied by verbs that express the communication of positive attitudes and information between people, whereas on the lower right corner there are verbs that are used to describe mostly very destructive use of power. In newspaper texts, the verbs in the upper right corner are often used to describe how the interviewee reports on an event. By the selection of the reporting verb, the writer can comment on the attitude or mood of the quoted person. This is exemplified in a) and b) below. In a), by using the verb *naurahtaa* 'to laugh briefly', and in b), by using the verb *huoata* 'to sigh', the writer shows his/her subjective interpretation of the attitude the speaker has towards his/her own message.

(a) "Ja nämä venetsialaiset," **hän** *naurahtaa* **.** "He ovat tosi maalaisia, maalaisempia kuin ihmiset Georgiassa, josta minä tulen, ha-haa." (CSC, 2001:Helsingin Sanomat 1990.)
"And these Veniciens," **he** *laughes briefly*. "They are real peasants, more peasants than people in Georgia, from where I come, ha-haa."

(b) "Voi näitä 1970-luvun lauluja. Ne ovat ihania**,**" *huokaa* **taiteilija** itsekin. (CSC, 2001:Aamulehti 1995.)

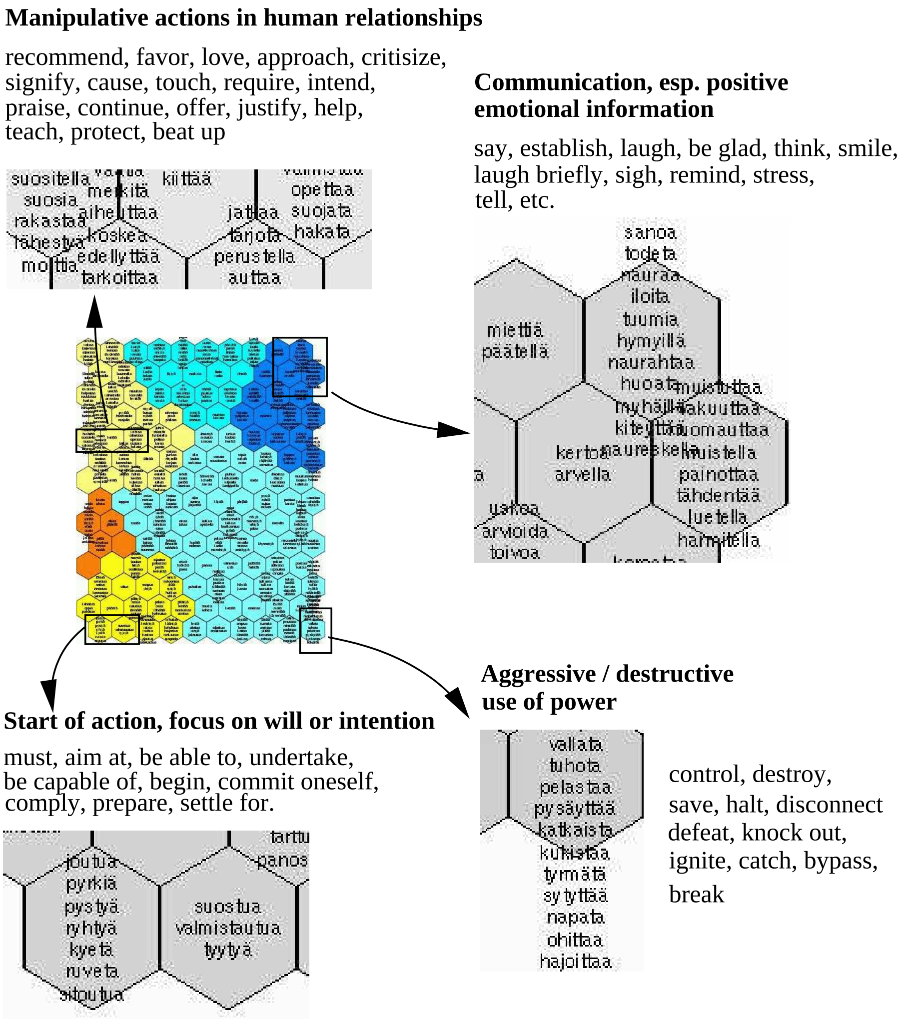


Figure 1.6. A map of the 600 most frequent verbs (base forms) in the Newspaper corpus. The verbs were organized on the basis of the distribution of morphological features in one preceding and two succeeding words, collected over all instances of the verb in any inflected form. The contents of four sample map regions are shown in the insets. In the baseline classification (pp. 157–165 in Pajunen, 2001), 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 (Levin, 1993) for English verbs.

"Oh, these songs from the 1970s. They are fantastic," *sighs* **the artist** himself, too.

The examples reveal also some typical contexts the verbs in the upper right corner of the map share, i.e. the subject in the nominative (bolded NPs), and a punctuation mark. In a), the verb *naurahtaa* is preceded by the nominative subject *hän*, i.e. the third person singular pronoun 'he/she', and followed by a full stop. In b), the reverse order of the contextual features is shown, i.e. the nominative subject *taiteilija* 'artist' follows the verb *huoata*, while there is a comma immediately to the left of the target verb. To obtain a more thorough picture of the features linkin up the verbs in a category, or to compare two categories in different map units, one can examine the data vectors calculated for each map unit. E.g. in Figure 1.2, data vectors for two categories are shown, i.e. for the category *lähteä-mennä-päästä*, and *puhua-sanoa-tietää*.

It is noteworthy that in choosing the input features the only cognitively relevant principle was that of *closeness*: i.e. the iconic principle mentioned already in Section 4.5 according to which close spatio-temporal proximity correlates with conceptual proximity. As a result we were able to uncover emotional and cognitive dimensions behind the words.

We consider it likely that such dimensions do emerge if one examines different text types, too, or even spoken input. Among different text types, different features may be prominent for a particular category, but they are nevertheless likely to be sufficiently consistent within the particular text type to enable discovery by automatic data analysis. In particular this is the case if one is of the opinion that cognitive representations form the groundwork of our language ability, and that linguistic structures have evolved merely as a means for solving the problem of how to put our possibly non-sequential thoughts into some sequential communicative expression in an orderly fashion.

The two analyses, the baseline categorization and the clustering method described here, can be seen as two complementary perspectives on lexical organization, focusing on different research questions as well. Pajunen starts from analysing verbs and grouping them into ontological-semantic classes, and then draws category-specific syntactic generalizations regarding the argument structure of verbs. As Pajunen (2001:33) herself puts it, the semantic classification she 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). In contrast, the method described here represents a bottom-up approach to lexical acquisition.

A further interesting question is what kind of lexical organization one would end up with if the method described here were applied to new languages with different typological properties such as function words instead of the case marking system.

7. Conclusions and Discussion

On the outset it may have seemed surprising that in Experiment 1 we decided to use only the word immediately following the verb for collecting information. After all, (1) many verbs take several arguments, and (2) since the word order in Finnish is free—or better, discourse-conditioned (Vilkuna, 1989)—the arguments may in some cases appear quite far apart from the verb, either before or after the verb. Nevertheless, comparison between the map ordering and the baseline verb categorization shows that the selected morphosyntactic features of the next word indeed are statistically quite indicative of the functional-semantic properties of the verb, if the information is accumulated over a large number of individual samples. In such an accumulation, the typical cases will dominate and the less typical cases, many of which may be random noise, are less frequent and become averaged out over time. Nevertheless, as indicated by Experiment 3, a wider context seems to add useful information.

From a methodological point of view, the visualizations obtained with the SOM can be extremely valuable, e.g., for a linguist interested in actual language use. For example, in Experiment 2 the obtained overview of the behavior of the verb forms draws attention to some interesting differences in word usage that are otherwise hard to notice. As pointed out in (Kaski, 1997), using traditional statistical methods it is relatively easy to answer well-defined specific questions such as 'How frequent is the use of the first and the second person pronouns in spoken language compared to their frequency in newspaper texts?'. However, when the domain is not known very well, or the questions are not initially very specific, or when one wishes to find novel, unexpected details, the traditional methods are insufficient. In such cases data exploration methods, such as the SOM, can be particularly useful in helping to understand the complex relationships in the data.

The Experiment 3 highlights the importance of cultural, social and emotional aspects in lexical organization. As is well known, cultural variation of lexical organization has been the focus of a number of linguistic studies, e.g., studies concentrating on the speaking of emotions (see e.g. Athanasiadou and Tabakowska, 1998, Siiroinen, 2001), spatial language, kinship terms, the systems of color terminology, etc. (see e.g. Foley, 1997 for further references about the topics mentioned).

The massive computer readable corpora now available have the potential to present a challenge to the way we think about the linguistic phenomena, too. For example, in theoretical corpus linguistics, one sense of the term 'data-driven' is to find new ways of deriving the theory from the data and thus bring new insights into linguistic theory itself (cf. Sinclair, 1991; Stubbs, 1996). This question is all the more relevant given the recent advances in the development of data analysis methods that are capable of visualizing the complex relationships in the data.

Notes

1. Morphological features are often regarded as "surface features" since they are to some degree apparent in the actual words.
2. ©Conexor oy and Anu Airola, <http://www.conexor.fi>
3. K-means algorithm is obtained as a special case of the SOM by disregarding the neighborhood in the organization.
4. In fact, since Pajunen's classification system is highly hierarchical, in it the verbs *puhua* and *sanoa* are further classified into two minor subclasses of speech act verbs, namely the 'verbs of speaking', and the 'verbs of saying'.
5. Inchoative aspect: expresses the very beginning of a state or activity. (See e.g. Trask, 1993.)
6. For a similar comparison between an existing syntactic categorization and a clustering, see (Redington et al., 1998.

References

Athanasiadou, A. and Tabakowska, E., editors (1998). *Speaking of Emotions. Conceptualisation and Expression*. Mouton de Gruyter, Berlin.

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

Croft, W. (1999). Typology and cognitive linguistics. In Janssen, T. and Redeker, G., editors, *Cognitive Linguistics: Foundations, Scope, and Methodology*, pages 61–94. Mouton de Gruyter, Berlin.

CSC (2001). CSC—tieteellinen laskenta oy. Finnish Language Text Bank: Corpora 'Books' and 'Newspapers'. Assemblers: Department of General Linguistics, University of Helsinki, Research Institute for the Languages of Finland, and CSC. http://www.csc.fi/kielipankki/.

Helsingin sanomat (1990) A collection of 3066995 words of Finnish electronic documents.

http://www.csc.fi/kielipankki/.

AL:Corpora The Finnish Parole-corpus. 20 million words of Finnish electronic documents, collected in LE-PAROLE -project in years 1996-1998. Assemblers: Department of General Linguistics of the University of Helsinki, Research Institute for the Languages of Finland. http://www.csc.fi/kielipankki/.

20 million words ofjoonan sanan laajuinen kokoelma suomenkielisid sdhkvisid dokumentteja kerdtty LE-PAROLE -hankkeessa vuosina 1996-1998. Koostajat: Helsingin yliopiston yleisen kielitieteen laitos ja Kotimaisten kielten tutkimuskeskus. Saanti: http://www.csc.fi/kielipankki/

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

Foley, W. A. (1997). *Anthropological Linguistics. An Introduction.* Blackwell Publishers, Oxford, UK.

Givón, Talmy (2001). *Syntax. An Introduction.* Volume I. John Benjamins Publishing Company. Amsterdam/Philadelphia.

Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture. The University of Chicago Press, Chicago.

Gärdenfors, P. (2000). Conceptual spaces. MIT Press.

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

Haykin, S. (1999). *Neural Networks—A Comprehensive Foundation*. Prentice Hall, 2nd edition.

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

Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulié, F. and Gallinari, P., editors, *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, volume II, pages 3–7, Paris. EC2 et Cie.

Hurme, R., Malin, R.-L., and Syväoja, O. (1984). *Uusi suomi-englanti suursanakirja. Finnish-English General Dictionary*. WSOY, Porvoo.

Ikola, O., Palomäki, U., and Koitto, A. (1989). *Suomen murteiden lauseoppia ja tekstikielioppia*. Suomalaisen Kirjallisuuden Seura, Mänttä.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall.

Karlsson, F. (1983). *Suomen kielen äänne- ja muotorakenne*. WSOY, Porvoo. See also http://www.ling.helsinki.fi/˜fkarlsso/bookpaprep.html.

Karlsson, F. (1987). *Finnish Grammar*. WSOY, Juva, the second edition edition.

Kaski, S. (1997). Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*. DTech Thesis, Helsinki University of Technology, Finland.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21:101–117.

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

Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin.

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

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

Lyons, J. (1977). *Semantics.*, volume 1. Cambridge University Press, Cambridge.

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

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

Pajunen, A. (1999). *Suomen verbirektiosta. Verbin argumenttirakenteen jäsenten valinta.* Yleisen kielitieteen julkaisuja 1. Åbo Akademis tryckeri, Åbo.

Pajunen, A. (2001). *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* Suomalaisen kirjallisuuden seura.

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

Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254.

Siiroinen, M. (2001). *Kuka pelkää, ketä pelottaa.* SKS, Helsinki.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford University Press, Oxford.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture.* Blackwell, Oxford.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C. Association for Computational Linguistics.

Tomasello, M., editor (1998). *The New Psychology of Language.* Lawrence Erlbaum Associates, Mahwah, NJ.

Trask, R. L. (1993). *A Dictionary of grammatical Terms in Linguistics.* Routledge, London.

Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (2000). SOM ToolBox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

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

Zavrel, J. and Veenstra, J. (1995). The language environment and syntactic word-class aquisition. In *Proceedings of the Groningen Assembly on Language Aquisition (GALA95)*, pages 365–374.