UNIVERSITY OF HELSINKI
DEPARTMENT OF GENERAL LINGUISTICS

Master's Thesis

Unsupervised discovery of morphs in children's stories and their use in Self-Organizing Map -based analysis

Mikaela Klami
011929277

Supervisor: Kimmo Koskenniemi

Espoo, November 24, 2005

Acknowledgements

The work for this Master's Thesis was carried out in the Laboratory of Computer and Information Science at Helsinki University of Technology. I wish to thank my instructors, D.Sc. Krista Lagus and docent Timo Honkela, for the guidance, discussions and ideas as well as their encouragement, constructive criticism and comments that lead to the accomplishment of this Master's Thesis. I also want to thank the Laboratory of Computer and Information Science in general, for providing me with the opportunity of working here even though my allegiances have been lying at another university and department.

My gratitude also goes to the supervisor of my work, professor Kimmo Koskenniemi from the Department of General Linguistics at the University of Helsinki, especially for his willingness to put up with this extremely tight schedule. The Children are Telling group of independent researchers also has my thanks for providing us with the unique set of data used in this Thesis, and for their many questions and ideas on how to utilize it.

Further, I would like to thank my research group of Computational Cognitive Systems at the CIS laboratory, and particularly my colleague Tiina Lindh-Knuutila for proof-reading my Thesis and providing valuable comments and insights. Also, Mathias Creutz has my special thanks for his generous help with Morfessor.

Finally, my gratitude goes to my family and friends, who tolerated my endless babble about the Thesis and cheered me up when I felt out of energy. Especially, I want to thank my husband Arto, who kept me going and stood there for me all the way through this.

In Espoo, 24th November 2005

Mikaela Klami

Contents

1 Introduction 1

1.1 Problem setting . . . 1

1.2 Aim of the Thesis . . . 1

1.3 Structure of the Thesis . . . 2

2 Background 4

2.1 Linguistic concepts . . . 4

2.1.1 Linguistic structure . . . 4

2.1.2 Linguistic context . . . 6

2.2 Statistical natural language processing . . . 8

2.2.1 Supervised and unsupervised learning . . . 9

2.2.2 Emergence of linguistic structure . . . 10

2.3 Storycrafting method . . . 12

3 Methods 15

3.1 Morfessor . . . 15

3.1.1 The algorithm . . . 16

3.1.2 Other unsupervised methods for morphology induction . 20

3.2 Self-Organizing Map . . . 21

3.2.1 The algorithm . . . 22

3.2.2 Related work on SOMs in natural language processing . 25

3.2.3 Constructing word SOMs with morph features . . . 28

4 Data set 32

4.1 Description of data . . . 32

4.2 Division into categories . . . 35

4.3 Preprocessing . . . 37

4.3.1 Removing metadata . . . 37

4.3.2 Preprocessing stories . . . 38

4.4 Morphological analysis . . . 39

4.5 Suggested standardized format for stories . . . 40

5 Experiments on feature selection 43
5.1 Word SOM parameters 43
5.2 Evaluation measure 44
5.2.1 Manual baseline categorization of word forms 44
5.2.2 Evaluation algorithm 47
5.3 Feature selection experiments 48
5.3.1 Experiment results 49
5.4 Conclusions 52

6 Data analysis using the SOM 54
6.1 Analysis of children's stories data 54
6.2 Comparison: Different age categories 62
6.3 Conclusions 68

7 Discussion 70
7.1 Conclusions 70
7.2 Future work 71

List of Figures

3.1 An excerpt of the output of Morfessor Categories-ML variant on the children's stories corpus. 17

3.2 A U-matrix representation for a 12 × 9 hexagonal SOM with five clusters, and component plane representations for its features X, Y and Z. 25

3.3 An illustration of the process of calculating a feature vector when features are whole context words. 29

3.4 An illustration of the process of calculating a feature vector when features are morphs and context words have been morphologically segmented. 31

4.1 A random example of a story by an individual child. 33

4.2 A random example of a group story. 34

4.3 A suggested XML-based template for recording stories. 41

5.1 An excerpt from the list of the manually classified 200 most frequent word forms in the whole children's stories corpus. . . . 46

6.1 The U-matrix representation of the word SOM on the whole children's stories corpus. 56

6.2 Component plane images of some feature morphs active in the agent noun cluster. 57

6.3 Component plane images of some feature morphs active in the imperfect tense verb cluster. 58

6.4 Component plane images of some feature morphs active in the quantity word cluster or in the this-cluster. 60

6.5 Component plane images of some feature morphs active in the location word cluster or in the movement verb cluster. 62

6.6 The U-matrix representation of the age category word SOM on the stories of children aged from 1 to 4. 64

6.7 The U-matrix representation of the age category word SOM on the stories of children aged from 5 to 6. 65

6.8 The U-matrix representation of the age category word SOM on the stories of children aged over 6. 66

List of Tables

4.1 The number of stories and words and average story length in each of the three age categories. 36
4.2 The performance of three Morfessor variants on the children's stories corpus. 40

5.1 The evaluation results for different word SOM variants. 50

6.1 The types and sizes of the feature morph and training word sets for word SOMs of each age category. 63

Chapter 1

Introduction

In this Chapter, the framework and context of this Master's Thesis are explained briefly. The task and aims of this work are introduced, and an overview of the structure of the Thesis is presented.

1.1 Problem setting

The work and experiments performed in this Thesis fall within the broader framework of statistical natural language processing, and, more precisely, particularly in the context of the emergence of linguistic structure. The level of linguistic structure in the scope of this work is limited to studying the emergence of word-level categorizations.

Also, through the data set used in the experiments, this Thesis also relates to the study of children and the emergence of human language skills. The data set, provided by the *Children are Telling* group of independent researchers, is a collection of stories told by Finnish children aged from 1 to 14 and collected using a special method called Storycrafting, which seeks to promote equality in dialogs between children and adults.

1.2 Aim of the Thesis

In this Thesis, the children's stories text corpus is analyzed with an unsupervised learning method called the Self-Organizing Map (SOM). The aim is to have the linguistic structure that is present in the stories of young children, especially at the level of word categorizations, emerge automatically from the corpus itself.

The main innovation of this Thesis is the utilization of emergent morphology-level information as the features for constructing self-organizing maps. Compared to traditional self-organizing map -based word categorizations which use whole context words as features, the utilization of morphological information is hoped to improve the quality of the resulting word maps. In fact, Lagus et al. (2002) have successfully categorized Finnish verbs using word SOMs with morphosyntactic features, obtained with a rule-based parser for Finnish.

In this Thesis, the morphological information used in the training of the self-organizing maps is extracted automatically from the children's stories corpus itself, with a fairly recently developed unsupervised morphology induction method called Morfessor. Hence, the main goal of this Thesis is to find out whether the use of morphology-level features obtained by an unsupervised method could help in training self-organizing word maps that are of better quality than traditional whole context word -based word maps. In summary, this Thesis studies the task of categorizing Finnish words in a completely unsupervised manner.

Also, in order to find out whether the selection of different types of morphs, namely root morphs, suffixes and prefixes, for features of a self-organizing map could affect the quality of the resulting word map, experiments are performed on self-organizing word maps with different combinations of morph types as features. An evaluation method for automatically measuring the quality of word maps is developed, based on comparing part-of-speech information of word forms mapped to adjacent map nodes and calculating a kind of a density score for the word clusters on the map. Based on the experiment results, a successful combination of morphs is chosen for the features of the final self-organizing maps on the children's stories corpus. Then, the story corpus is analyzed through these self-organizing word maps, particularly from the point of view of emergent word categorizations.

1.3 Structure of the Thesis

This Thesis consists of roughly four parts. In the first one, the framework and methodology of this work are presented. In Chapter 2, both some linguistic concepts and background knowledge on statistical natural language processing and unsupervised learning methods are explained, essential for understanding the experiments performed in this Thesis. Then, in Chapter 3, the methodology used in this Thesis, namely the Morfessor morphology induction method and the self-organizing map, is described in more detail.

Chapter 4 is an introduction to the data set utilized in this work, a Finnish corpus of stories told by children aged from 1 to 14. The nature of the data and its division into subcategories is explained, and the preprocessing and morphological analysis procedures performed on the data are described. Finally, a standardized format for improving the machine-readability of the existing and future story data is presented.

In Chapter 5, the selection of morph features for a self-organizing word map is examined, and an evaluation measure is presented for enabling automatical evaluation and comparison of word maps. The evaluation results for several self-organizing word maps with both morphs and whole context words as features are studied, and observations on the performance of the different word map variants and on the optimal sets of morph features are made.

Chapter 6 contains the actual data analysis of the children's stories corpus, using self-organizing word maps with morphs as features. First, a more detailed analysis of the whole story corpus is presented. Then, self-organizing word maps are constructed on the age-based subcategories of the corpus, and comparisons between word maps on the data in the different age categories and also with the word map on the whole corpus are performed. Finally, Chapter 7 presents a summary and the conclusions on the work performed for this Thesis, and some suggestions on future work in this area are made.

Chapter 2

Background

In this Chapter, the background and framework of this Thesis are described in more detail. First, some linguistic concepts related to the area of research of the Thesis are explained. Then, the statistical natural language processing framework of the Thesis' methodology is introduced (see Chapter 3 for the actual methodology). Finally, a method called Storycrafting for collecting stories from children, used in obtaining the children's stories data set analyzed in this Thesis, is described.

2.1 Linguistic concepts

Before delving into the actual methodological framework of the Thesis, it is fit to take a look at some linguistic concepts and the nature of a natural language. First, different kinds of structure in language are viewed shortly. Then, some clarifications and explanations on the terminology that will be used later in this Thesis are presented.

2.1.1 Linguistic structure

Natural language is a system with an abundance of structure. First, the main structural distinction is the dualism between the sound and meaning of words of a natural language (Karlsson, 1998). Language is symbolic of nature, meaning that it consists of symbols (words of the language) and different combinations of these symbols. The relation between the form of the symbols (their pronunciation) and their meaning (semantics), however, is completely arbitrary. It is based only on a social convention that this or that symbol should refer to this or that referent in the world. But even if this relation between form and meaning of linguistic symbols is arbitrary, its nature of conventionality means that the relation, socially accepted, is also indispensable (Karlsson, 1998).

In addition to the minimal basic symbols of a language, more complex symbols can be constructed using the basic symbols and the special structural rules of the language (Karlsson, 1998). For example, some morphological processes, such as the one of deriving from the words “snow” and “man” a new compound word “snowman”, are very productive in natural languages. Sometimes the new compound word is simply the sum of its parts, but sometimes the new word carries a meaning that is not obvious from the original words. Derivation is a morphological process of creating new words from existing words and derivational affixes, as opposed to the process of inflection which produces inflected word forms of the same word (snow+s → “snows”).

Natural language thus has structural rules that can operate on several different levels of abstraction. The most concrete subsystem of language is *phonetics*, the study of the sound units of a language and the way they are produced and observed. All linguistic symbols consist of such sound units, called phones. The slightly more abstract study of the structure of the sound units is called *phonology*. The subsystem of the conventionalized words of a natural language is the *lexicon*, or the vocabulary of the language. The subsystem that studies the internal structure of words and their composition is called *morphology*, and *syntax* in its turn studies the combination of words into phrases and sentences. Finally, at the most abstract level, the subsystem of *semantics* involves studying the meaning of linguistic symbols.

Put together, the subsystems of phonology, lexicon, morphology and syntax are often regarded as the formal subsystems whose units have a physical phonological form (Karlsson, 1998). Their opposite is the subsystem of semantics, which is materialized through the formal subsystems, especially the lexicon. Semantics has therefore a connection to each of the other subsystems. Despite of its lack of own physical form, the semantic meaning is inseparable from the form it is realized as (Karlsson, 1998). Also, the immaterial nature of semantics doesn’t mean that it would be devoid of structure.

Each of the formal subsystems has its own units, and the categories that the units belong to (Karlsson, 1998). For example, phonetics categorizes phones, and syntax has categories for different types of phrases and sentences. However, the subsystems that are most central to the work in this Thesis are those of morphology, lexicon and semantics. Lexicon involves the categorization of words into part-of-speech classes, for example into nouns, verbs, adjectives and so on. The units of lexicon are words, or independent vocabulary items called lexemes.

Morphology, on the other hand, can have categories for example for the ending types of number, case and person of Finnish words. Other morphological categories include tense, aspect, and mode affix types of verbs, adjective comparison affixes, the many affixes of the processes of deriving new words, and so on. The units of morphology are called morphemes, and they can be divided into free or unbound morphemes and bound morphemes. Free morphemes can occur by themselves, but bound morphemes cannot as they are always attached to some other morphemes. Morphemes are regarded as the smallest linguistic units that bear a meaning (Matthews, 1991).

Before moving on, a short clarification on some morphological terms is needed. Affixes are bound morphemes that can be attached to before, after or within a root or stem. In this Thesis, the word *root* is used as referring to the portion of a word that has been stripped of *all* affixes and is not further analyzable into meaningful elements. Some word roots can appear by themselves and are thus free morphemes, but others always require affixes to be attached to them. The word *stem*, in its turn, refers to a root of a word together with some possible derivational affixes, but without inflectional affixes. Thus, the adjective "luotettava" ('reliable' or 'trustworthy' in English) is a root, but the adjective "epä+luotettava" ('unreliable' or 'untrustworthy' in English), derived from the previous, is a stem.

Finally, in this Thesis, the word *morph* is used as referring to a phonetic realization of a morpheme, as opposed to *morpheme* which means the smallest meaningful unit in a language. A morpheme, for example the Finnish suffix *-ssA* for marking the inessive case, may have more than one realizations as a morph due to allomorphy, or morphophonological variation in languages. For example, the Finnish inessive case suffix morpheme mentioned above can have two different phonetic realizations or allomorphs, namely *-ssa* ("juna+ssa" or 'in (the) train') and *-ssä* ("kynä+ssä" or 'in (the) pencil'), depending on the vowels in the root or stem it is attached to.

2.1.2 Linguistic context

One linguistic concept that will be essential in understanding the experiments described later in this Thesis is the notion of context. Basically, linguistic context refers to the language surrounding the word, phrase or whichever linguistic unit we are looking at.

The units, categories and their realizations in the different subsystems can be grouped under the term *element* (Karlsson, 1998). Elements have inherent properties, for example verbs are words that are used to express action, existence or a state of being, and noun phrases always have a noun as their head word. Elements also have a distribution, which means the linguistic environment the elements can occur in (Karlsson, 1998). The co-occurrence of an element with elements of some other type and the relations between these elements are central to the methodological framework of this Thesis.

An element is considered to be in a *syntagmatic relation* with the other elements that it can be catenated with to form a linear sequence of words (Karlsson, 1998). These sequences of elements of some level are called syntagms. The meanings of the words in the sequence are also in a syntagmatic relation with each other, for example in the sentence "The cat purred." purring is an act which is usually related to felines, and cats are animals that often express their contentedness by purring. These kinds of syntagmatic relations between words are what is used as the basis of categorizing word forms in the experiments conducted in the course of this Thesis.

Apart from syntagmatic relations, elements are also in a *paradigmatic relation* with the elements they are interchangeable with in a certain frame (Karlsson, 1998). For example, even if purring is an act usually reserved for cats, it can be used as a figure of speech to yield sentences like "The engine purred." or "The woman purred." In this frame that consists of the definite article and an inflected form of the verb "purr", the words "cat", "engine" and "woman" are in a paradigmatic relation with each other and thus form a paradigm. In the word SOMs that are trained in this Thesis, the word forms that end up in the same node on the map or very close to each other can be considered as forming a kind of such paradigm with each other (see Chapters 3 and 6 for more information on the methodology and the resulting word SOMs).

Finally, the size of the context or the frame in which the syntagmatic and paradigmatic relations of elements are studied can vary. The context may consist of only one or two elements immediately before and after the element in question, or the *context window* may extend over several words or maybe even sentences. Also, even if the context window is large, all elements that fall within its span are not necessarily taken into consideration but perhaps only a subset of them, for example every second element or only the two elements that are two steps before and after the center of the frame.

To conclude this section on the linguistic background, the aim of this Thesis is to find categorizations for words, a task which belongs traditionally to the subsystem of lexicon. Indeed, the usual way to categorize words is to use part-of-speech classes, which are the traditional categories of lexicon. The categories that emerge in the experiments of this Thesis, however, are slightly different. They have less to do with the subsystem of lexicon than has traditional word categorization, and they tend to give much more weight to the semantic similarity of words. Also, the methods which are used to construct the categories borrow information from the subsystems of morphology and, in the form of context windows, even syntax.

2.2 Statistical natural language processing

The work in this Thesis falls into the category of statistical natural language processing (statistical NLP). The word "statistical" means here simply that in this approach, NLP problems are being solved with methods that use natural language text corpora and statistical and probabilistic tools for extracting information from them. Adopting the definition of Manning and Schütze (1999), "statistical NLP comprises all quantitative approaches to automated language processing, including probabilistic modeling, information theory and linear algebra".

In short, statistical NLP usually consists of non-logical work on NLP problems. Its opposite are systems that use rules to structure linguistic expressions. Different kinds of rules on linguistic structure have a long history in linguistics and also in NLP. In the last century, however, this rule-based approach became increasingly complicated and rigorous, as detailed grammars attempting to describe what were well-formed versus ill-formed utterances of a language were constructed (Manning and Schütze, 1999).

But, as Edward Sapir (1921) already put it, "All grammars leak." It is simply not possible to provide an exact and complete characterization which would encompass all well-formed utterances of a language and which would cleanly separate them from all other sequences of words, considered ill-formed utterances (Manning and Schütze, 1999). This is due to the fact that language is not a static system but rather a tool that is constantly adapted by people to meet their current communicative goals and needs. Rigid rule systems cannot tackle such adaptiveness, and therefore a looser approach is needed.

Instead of trying to find rules to describe grammatical or ungrammatical sentences, statistical NLP aims to find the common patterns that occur in language use. The practitioners of statistical NLP are thus interested in good descriptions of the associations and preferences that occur in the totality of language use, instead of concentrating on categorical judgements about sentences that can, in reality, be very rare in actual language use (Manning and Schütze, 1999).

Statistical NLP has always had quite an applied character to it. This is quite natural, given the fact that it usually tries to find solutions to real NLP problems, some of which may have eluded solution for a long time when using traditional methods. Much of the skepticism and criticism towards probabilistic models for language stem from the fact that the well-known early probabilistic models in the 1940s and the 1950s were extremely simplistic of nature (Manning and Schütze, 1999). But as Manning and Schütze (1999) argue, complex probabilistic models can be just as explanatory as complex non-probabilistic models – but with the added advantage that they can also explain phenomena that involve uncertainty and incompleteness of information, which occur so frequently in human cognition and particularly in language.

2.2.1 Supervised and unsupervised learning

Statistical natural language processing usually involves some kind of machine learning. Machine learning means positing some general form of model and then using training patterns to learn or estimate the unknown parameters of the model. Learning, in turn, refers to some form of algorithm for reducing the error on a set of training data (Duda et al., 2001).

Machine learning algorithms can be roughly classified into supervised and unsupervised algorithms, depending on the task and the nature of the data which is used to train them. The distinction is that with supervised learning, we know the actual status for each piece of data on which we train; a category label for each pattern in a training set is provided in advance. With unsupervised learning, however, we do not know the classification of the data in the training sample beforehand. There is no explicit "teacher", and the algorithm forms its own clusters or "natural groupings" of the input patterns (Duda et al., 2001). Unsupervised learning can thus often be viewed as a clustering task, while supervised learning can be seen as a classification task (Duda et al., 2001). In supervised learning, we typically have a manually annotated text corpus or some other pieces of information that have usually involved human effort, and the aim is to have the algorithm learn to repeat the annotation. An unsupervised learning algorithm, in turn, attempts to learn to extract information automatically from an unannotated text corpus. The methodology used in the experiments of this Thesis belongs to the latter category of unsupervised learning.

Using unsupervised learning algorithms in statistical NLP can thus help save human effort in solving an NLP task. Of course, there are some already annotated text corpora[1] distributed freely for statistical NLP research purposes, but sometimes the existing annotated corpora simply cannot satisfy the need at hand. This is the case with for example the children's stories corpus used in this Thesis. Being an instance of the actual use of language of small children, with its slangy and particular expressions and words, no existing annotated corpus[2] would be of much help in training an algorithm for the task of categorizing the

[1] See for example the Brown Corpus (Francis and Kucera, 1964), the British National Corpus (Burnard, 1995) or the Penn Treebank (Marcus et al., 1993).

[2] With the possible exception of the CHILDES database (MacWhinney and Snow, 1985), which contains transcripts of conversations with young children. This corpus, however, has the fault of neither being really textual data but a collection of audio recordings with tran-

words of this data. Also, making an own annotated corpus for this particular task would be a tedious and extensively time-consuming job.

Further, with the era of the Internet with its vast, constantly expanding amounts of text data, it would be wasteful not to be able to utilize such huge, free text resources in NLP tasks. Algorithms that can learn on unannotated text data are thus a great asset which enable the harnessing of the potential of unprecedentedly large text collections.

Finally, sometimes a categorization made by an unsupervised learning algorithm is exactly what we hope to achieve. The unsupervised method may find in the data some patterns that would have been missed using the pre-determined classes of a supervised learning algorithm. An unsupervised learning algorithm may succeed in extracting from the corpus information of a completely different type or on a completely different basis than what its creators did or did not originally have in mind. This could help give totally new viewpoints into the data, and reveal some facts about it that would perhaps have otherwise been missed. This, together with the fact that material for supervised categorization of the words in the children's stories corpus was not even available, were the main motivations for turning to unsupervised methods rather than supervised in this Thesis.

2.2.2 Emergence of linguistic structure

Applying unsupervised learning methods to natural language processing tasks in the purpose of finding implicit patterns in the data can also be viewed as *emergence of linguistic structure*. When an unsupervised learning algorithm extracts its own categorizations from the data, these categories are considered to be emergent, something that emerged from the data itself, as opposed to the predefined classes of a supervised learning algorithm. The emergent structure can either be in correspondence with some existing linguistic theory (for example a theory on word categorization), or it can also represent a categorization of a completely new type, based on phenomena which may have been previously ignored or which may have passed unrecognized until now.

Linguistic structure can emerge from data on several different levels, corresponding to the subsystems of language described earlier (see Section 2.1.1). For example, on the level of morphology, there have been several efforts to extract morphological information automatically from text corpora. One such method, namely the *Morfessor* family of algorithms (Creutz and Lagus, 2005a), was also used for providing the morphological information utilized in the ex-

scripts, and nor having been collected from Finnish-speaking children like the data set used in this Thesis.

periments and data analyses performed in this Thesis (see Chapter 3 for more information on Morfessor and other morphology extraction algorithms).

The research on the emergence of word categorizations and semantics is even more abundant. One method which is claimed to find semantic emergent representations is the *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990), which is a statistical technique for extracting and representing the similarity of meaning of words and passages by analysis of large bodies of text. The idea is to use singular value decomposition to reduce a very large matrix of word-by-context data into a considerably smaller and more compact representation. This resulting representation has been shown to mimic closely the way humans judge meaning similarity (Landauer and Dumais, 1997). As Landauer et al. (1998) point out, the similarity estimates derived by LSA are not based on simple frequencies or co-occurrences but they depend on a deeper statistical analysis – on the "Latent Semantics", an instance of the emergence of linguistic structure.

More recently, an algorithm called *Independent Component Analysis* (ICA) (Hyvärinen et al., 2001) has been used for a similar task. ICA is a statistical and computational technique for revealing hidden factors that underlie in multivariate data. The variables in the data are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The aim of ICA is to to find these latent variables, called the sources or the independent components of the observed data. In a more linguistic context, ICA has been applied by Honkela et al. (2005) on word context data to extract distinct features or categories that reflect syntactic and semantic categories of words.

In this Thesis, however, yet another method was adopted for the task of word category emergence. Like LSA and ICA, also *Self-Organizing Maps* (SOMs) (Kohonen, 2001) can be used to construct a representation of the input text data based on word contexts (Ritter and Kohonen, 1989). Apart from generating representations that are conceptually intuitive, SOMs also have the additional advantage of building an efficient visualization of the emergent contextual relations of words. More information on SOMs can be found in Chapter 3 of this Thesis.

Finally, emergence of structure has also been researched at the level of syntax, and, overlapping with the field of research of language evolution, even from the point of view of the emergence of an entire language. This kind of research typically involves simulations with populations of individual learners, often called agents. For example, in the computational model of Kirby (2000), syntactic rules are shown to emerge from unstructured data in a population of learners through observational learning, without natural selection of learners.

The success in this kind of experiments with social and cultural approaches to language evolution has been taken as an argument against the theories that the human language abilities would be genetically encoded and that language would have emerged just as a response to the pressures of natural selection. Rather, as Smith et al. (2003a) argue, language should be seen as a system which arises from the interaction of the three complex adaptive systems of biological evolution, learning and culture. The later model by Kirby (2001), called the *Iterated Learning Model* (ILM), has been proposed by Smith et al. (2003b) as a framework for new research on the cultural evolution of language. The simulation of language learning in agent populations has also been studied by for example Honkela and Winter (2003) and later by Lindh-Knuutila (2005), who use self-organizing maps to represent an agent's semantic memory or conceptual map.

2.3 Storycrafting method

The children's stories corpus used and analyzed in this Thesis was provided by a group of independent researchers called the *Children are Telling* group. The stories in the data set were collected between 1994 and 2001 using a method called Storycrafting (in Finnish, 'sadutus'). It is therefore fit to devote a section to describing in more detail the background and aims of this method.

The Storycrafting method is a Finnish invention that promotes equal possibilities for the participants in a dialog (Riihelä, 1991). It was developed especially for helping to transform the status of children in the society; to listen to what the children have to say. The Storycrafting method turns the focus to the person who tells the story – the child. Using the Storycrafting method, children can be heard the way children want to be heard: the children can choose the words, drawings and acts they want to use to express themselves (Riihelä, 2001). Also, the children may freely choose the subject or topic of their stories; adults are just to listen what they have to say, on whichever topic they choose.

The idea of the Storycrafting method is simple. The adult, or the storycrafter, asks the child to tell a story, and says that he or she will write it down exactly as the child will tell it. When the story is finished, the storycrafter will read it aloud to the child, who can then make any corrections or changes to the story if he or she wants to.

It is important to write the story down exactly as the child tells it, resisting the urge to correct any mistakes or slangy use of language by the child. The purpose is to make it clear to the child that the adult is specifically interested in the child's own story; the aim is to inspire the child to tell about his or her own world and thoughts (Riihelä, 2001). Contrary to the traditional relations between a caregiver and a child or an educator and a child, in the Storycrafting method it is the child who takes the lead and the adult should just follow behind and document the process. In accepting to write the story down exactly as he or she hears it, the storycrafter also accepts to respect the way of self-expression the child chooses to use, and not to change or add anything to it in the process (Karlsson, 2000).

Also, it is important to convey the feeling that the children have a copyright to their own work and that it is not just being used for the purposes of the adult (Riihelä, 2001). The adult can of course ask the child to give him or her a copy of the story, but above all, the story should belong to the child, to be his or hers to do as he or she pleases.

The Storycrafting method has been used most extensively in the Storyride network project co-ordinated by Finnish National Research and Development Centre for Welfare and Health (Stakes). The project started in 1995 on collaboration with 23 Finnish municipalities and professionals in social and health care, parishes, individual daycare centres and family daycare units and other institutions (Riihelä, 2001). In this network, the method for Storycrafting was further refined, and the construction of a children's own network of stories was begun. The project continues even today in collaboration with universities, colleges, daycare centers and cultural organizations in the Nordic countries, and it has received support from the Nordic Council of Ministers. Further information and a quantitative and qualitative evaluation of the Storyride project can be found in for example Karlsson (1999) and Karlsson (2000).

One purpose of the Storyride project has also been to create an own network of direct contacts between children. In the project, the stories told by a child or a group of children will be sent to another group of children in a different daycare center, school, orphanage etc., either in their own country or abroad. There, the story or stories will be read to a new audience, and in response, the listeners will tell their own stories based on their reactions to the received story. These new stories are then sent back to the original group, forming a kind of a story circle between the groups. Like this, the children are given the opportunity to produce their own culture, which is documented and published along the way (Riihelä, 2001). The children also get an opportunity to hear about other children's thoughts from different parts of their own country and from abroad.

The applications of the Storycrafting method are many. It can be used with one person or with groups, at home or at school, daycare center or some other institution, in parental advice, in special education, in social work or even in adult education. It can be used as an interview method, or as a therapeutic method e.g. for children who have experienced some kind of a traumatic event. It can be used to handle many problems, like speech disabilities, insulting treatment, physical and psychic illness, or simply to improve co-operation between adults and children or to change working practices towards some more client-centered habits. (Riihelä, 2001)

Further, the stories collected by using the Storycrafting method could also be of great help for research concerning the language of children, forming a valuable text corpus of the actual use of language of younger and older children. Particularly, stories from children of many different ages could help to understand how the human language skills develop throughout the childhood. In this Thesis, it is indeed from these points of view of (socio)linguistics and language development that the children's stories corpus provided by the Children are Telling group will be analyzed.

Chapter 3

Methods

In this Chapter, the methodology used in this Thesis is examined in more detail. First, the Morfessor family of algorithms for unsupervised extraction of morphological information from text corpora is introduced, and some other work on unsupervised induction of the morphology of a language is also briefly described. Then, the principles of the main method used in this Thesis, namely the Self-Organizing Map (SOM), are presented. Also, some applications of SOMs in natural language processing are viewed. Finally, the procedures for constructing a word SOM are explained, for both the traditional word SOMs with whole context words as features as well as for the morph-featured word SOMs which are the main innovation of this work.

3.1 Morfessor

Morfessor[1] (Creutz and Lagus, 2005a) is an unsupervised data-driven algorithm for inducing the morphology of a language. Inducing refers here to the emergence of morphological information from the text data itself, and by unsupervised it is meant that the algorithm is provided with no or very little morphological or other linguistic knowledge related to the task.

The aim of Morfessor is to segment words of an unlabeled text corpus into morphemes or morpheme-like units, and also to be applicable especially to highly inflecting, morphologically rich languages like Finnish. Also, the Morfessor morphology extraction method not only seeks to find the most accurate segmentation possible, but it also learns a representation of the language from the data it was applied to, namely an inventory of the morphs of the language.

The output of Morfessor is a lexicon of the words from the corpus, segmented at proposed morpheme boundaries into morpheme-like units called morphs.

[1] The Morfessor family of algorithms was first named Morfessor in (Creutz and Lagus, 2005b). The software is available at `http://www.cis.hut.fi/projects/morpho/`.

Since Morfessor does not, at least for the time being, recognize allomorphic variation, the units produced by its segmentation cannot really be called morphemes, but rather they should be regarded as something closer to morphs[2]. However, the morph lists produced by Morfessor are not necessarily even meant to be linguistically correct. When utilized the way described in this Thesis, for example – as input to another unsupervised learning algorithm, namely the self-organizing map – the question whether the morphs extracted by Morfessor actually strictly correspond to linguistically accepted Finnish morphs seems less important.

Morfessor has been tested on Finnish and English text corpora, with good results (Creutz and Lagus, 2004). Compared with other unsupervised morphology extraction tools, Morfessor seems to have a good performance on corpora both in the morphologically rich Finnish and in the less inflecting English language. The morphological analysis produced by Morfessor has been applied to speech recognition (Siivola et al., 2003; Hacioglu et al., 2003) and to improving language models (Virpioja, 2005). In the future, the tasks of for example machine translation and information retrieval could conceivably benefit from using automatically extracted morphological information. In fact, the recent Master's Thesis by Ville Turunen (2005) studies the use of Morfessor-extracted morphs in spoken document retrieval.

In this Thesis, the Morfessor algorithm is applied to the children's stories corpus in order to produce a morphological segmentation of the words in the data, which is then used in calculating the feature vectors for the morph-featured self-organizing maps presented in this work. An emergentist approach to acquiring a morphological analysis of the data was adopted because the colloquial, non-orthographical nature of the language in the children's stories corpus would have seriously challenged the capabilities of any non-statistical morphological analyzer for Finnish. An example excerpt from a Morfessor output morph lexicon, obtained by morphologically analyzing the children's stories corpus, can be found in figure 3.1. In the example, the numbers on the left refer to the frequency of the word form in the corpus, and each morph has been labeled as being either a root (STM), a prefix (PRE) or a suffix (SUF).

3.1.1 The algorithm

Morfessor is actually more like a family of algorithms than one specific method. The three current variants of the Morfessor approach to morphology induction are called, retroactively, Baseline, Categories-ML and Categories-MAP. The

[2] See Section 2.1.1 for a further terminological clarification concerning morphs and morphemes.

```
3 äiti/PRE + pupu/STM
1 äiti/PRE + roisto/STM
1 äiti/PRE + roisto/STM + a/SUF
1 äiti/STM + s/SUF
2 äiti/STM + si/SUF
3 äiti/STM + stä/SUF
1 äiti/PRE + sud/STM + e/SUF + lle/SUF
3 äiti/PRE + susi/STM
1 äiti/STM + t/SUF
3 äiti/PRE + tonttu/STM
1 äiti/PRE + vala/STM + s/SUF
51 aivan/STM
9 aivast/STM + i/SUF
2 aivast/STM + i/SUF + vat/SUF
1 aivast/STM + uksen/SUF
2 aivo/STM + kääpiö/STM
5 aivo/STM + t/SUF
1 aja/STM
36 aja/STM + a/SUF
4 aja/STM + an/SUF
1 aja/STM + i/SUF
2 aja/STM + ja/SUF
1 aja/STM + ja/SUF + lle/SUF
1 aja/STM + ja/SUF + n/SUF
1 aja/STM + ja/SUF + t/SUF
2 aja/STM + ksi/SUF
1 ajamaa/STM
40 aja/STM + maan/SUF
1 ajamisen/STM
34 aja/STM + n/SUF
```

Figure 3.1: An excerpt from a list of morphologically segmented word forms, extracted from the children's stories corpus. The numbers on the left denote the frequency of the word form in the corpus, and each morph has been labeled as either a root (STM), a prefix (PRE) or a suffix (SUF). This segmentation was obtained by using the Categories-ML variant of Morfessor.

Baseline method (Creutz and Lagus, 2002) utilizes the minimum description length (MDL) principle, i.e. it is based on minimizing the sum of the length of the model and the length of the data as measured using the model. The Baseline algorithm uses an incremental online learning approach to learning a morph lexicon of the data, analyzing each example word according to the model that had been built up so far.

The Categories-ML variant of Morfessor (Creutz and Lagus, 2004) uses simply a maximum likelihood (ML) estimation of the data, instead of measuring the minimum description length of the model. It also uses batch learning, a type of learning where, alternatingly, all the words in the data are first split according to a fixed model, and then the model is updated. The Categories-MAP (Creutz and Lagus, 2005a) model is similar to the Categories-ML model, but it uses maximum a posteriori (MAP) estimates of the parameters instead of ML estimates, and it is computationally slower. The Categories-MAP model also has a hierarchical morph lexicon, and it utilizes corpus frequency to decide when not to split a segment. In this Thesis, the method for providing the morphological analysis required for the experiments was chosen to be the Categories-ML model, since this Morfessor variant had the best performance of the three on the children's stories corpus (see Section 4.4). Consequently, only the Categories-ML model will be described in more detail here. The presentation follows the article of Creutz and Lagus (2004).

Unlike the Baseline method, the Categories-ML model also labels the morphs it segments, assigning them to the morph category of either roots[3] (STM), prefixes (PRE) or suffixes (SUF). The Categories-ML variant uses a Hidden Markov Model (HMM) to model morph sequences. These morph sequences are allowed to be quite long, making the algorithm especially applicable to highly inflecting languages like Finnish. In this task of learning the morphology from text data, neither the segments (morphs), nor their labels (morph categories) are known in advance.

In order to facilitate the task, some linguistic assumptions are made. First, as explained in the previous paragraph, morphs are assumed to fall into the two main categories of roots and affixes as far as sequential behaviour is concerned. However, roots and affixes are not allowed to be combined into just any sequence of morphs, but there should be some restrictions on the form of a legal morph sequence in order to prevent some sequences, like words starting with a suffix, from emerging. These restrictions, called morphotactic rules, can

[3] It should be noted that in the original Morfessor papers, the word *stem* is used instead of *root* when referring to the portion of a word that has been stripped of all affixes (see Section 2.1.1 for a further terminological clarification). For this reason, the label attributed by Morfessor to root morphs is called "STM".

be summarized as the regular expression:

$$\texttt{word = (prefix* root suffix*)+} \tag{3.1}$$

Finally, each category of morphs is assumed to be associated with some set of likely properties. For example, affixes are likely to occur together with many different morphs and more commonly than roots, and roots are probably morphs that are not very short.

For the sequences of morph categories occurring in a word, a first-order Markov chain (a bigram model) is assumed. For each category, there is a separate probability distribution over the set of possible morphs. Thus, the probability of a particular segmentation of the word w into the morph sequence $\mu_1, \mu_2 ... \mu_k$ is

$$p(\mu_1, \mu_2 ... \mu_k | w) = \left[\prod_{i=1}^{k} p(C_i | C_{i-1}) \cdot p(\mu_i | C_i) \right] \cdot p(C_{k+1} | C_k) \ . \tag{3.2}$$

In the equation, $p(C_i|C_{i-1})$ denotes a bigram model on categories, determining for example how likely it is that a prefix should follow another prefix. $p(\mu_i|C_i)$ is the probability that the category C_i should generate the morph μ_i, and $p(C_{k+1}|C_k)$ is the probability that a word ends with a morph of category C_k.

The actual Categories-ML algorithm proceeds as follows:

1. **Produce a baseline segmentation.** The Baseline variant of Morfessor is used to obtain a good initial morph segmentation of the data.

2. **Initialize $p(\mu_i|C_i)$ and $p(C_i|C_{i-1})$, and do EM.** The probability $p(\mu_i|C_i)$ for each given morph to be in a particular category is calculated using the left/right perplexity of the morph for affixes and the length of the root for roots. Right (or left) perplexity of a morph refers to a measure of the difficulty of predicting the morph that follows (or precedes) this particular morph. In order to help the optimization of the three probabilities for root/suffix/prefix-likeness, a fourth category of noise morphs is introduced.

3. **Remove redundant morphs, and do EM.** If there are morphs which can be split into submorphs that already exist, then they should be split. If there are multiple choices, the most likely one is chosen.

4. **Remove noise morphs, and do EM.** Noise morphs are usually short, and a result of over-segmentation. They are removed by merging them with adjacent morphs, according to some joining preference heuristics.

At the end of each step from 2 to 4, the probabilities of the model are re-estimated by using Expectation Maximization (EM). That is, the categories of all morphs are re-tagged using the Viterbi algorithm by maximizing the equation 3.2. The probabilities $p(\mu_i|C_i)$ and $p(C_i|C_{i-1})$ are then re-estimated from the tagged data, and this process is repeated until the probabilities converge. Basically, the EM makes things that have been observed frequently more likely, and things that have been observed infrequently less likely. After the final re-estimation of morph categories in step 4, all the words in the data are finally re-segmented using the newest model probabilities.

3.1.2 Other unsupervised methods for morphology induction

The method adopted in this Thesis for automatically extracting morphological information from the children's stories corpus belongs to the Morfessor family (the Categories-ML variant of Morfessor). However, there are also some other unsupervised methods which could have been applied to a similar task.

The work of Harris (1955) may be regarded as a basic approach to unsupervised induction of morphology. He proposes the use of so-called successor frequencies, stored in a trie structure, to find word and morpheme boundaries in phoneme utterances. The idea is that a word or a morpheme boundary is suggested at locations where the predictability of the next letter in a letter sequence is low – that is, where there is a peak in the successor count. Inside a word unit, the choices of successors are more limited, but at the boundaries of two units, the choice is typically much less restricted.

The approach is quite simplistic and obviously has its limitations, but some of them have been solved by Harris himself or by for example Hafer and Weiss (1974). They extend the work of Harris by proposing four different basic strategies for segmentation: segmenting according to a cutoff value for successor count, according to the peak and plateau strategy of the original work of Harris, according to a strategy favoring matches to complete corpus words, or according to a cutoff value of the successor entropy. They apply the output of their system to an information retrieval task.

Harris' system has even inspired some more recent methods for morphology extraction. Déjean (1998) presents a method where segmentation occurs when the successor count is greater than a threshold, set to be half of the number of letters in the alphabet of the language. Also, Goldsmith (2001) uses the successor and predecessor counts presented in Hafer and Weiss (1974) in his system, called Linguistica. He assumes that roots form groups that he calls signatures, and that each signature shares a set of possible affixes.

Schone and Jurafsky (2000) adopt a different kind of an approach to the problem of unsupervised morphology induction. They consider also the semantic content of words (in the form of word contexts) in determining the "morphologically relatedness" of word pairs sharing a set of hypothesized candidate affixes which may be morphological variants. Also Schone and Jurafsky use trie structures in identifying their candidate affixes. In their system, the semantic representations of terms, needed for comparing the semantical similarity of their contexts, are obtained using singular value decomposition, a matrix factorization method used in Latent Semantic Analysis. Schone and Jurafsky (2001) extend the previous work by for example adding support for circumfixation and for frequency similarity features, and by using transitivity to help find morphological variants otherwise unrecognized.

Yarowsky and Wicentowski (2000) and Wicentowski (2002) also use context similarity in determining morphological variants. Their system, which they call "minimally supervised", combines a few different unsupervised models to predict inflection–root alignments from an unlabeled corpus. The alignments are used to train a probabilistic string transduction model, whose output, in turn, is used to further refine the parameters of the unsupervised models. This process is iterated until the output converges. The unsupervised alignment models are based on for example the similarity between inflected forms and their citation form, the context similarity of morphological variants, or the distributional similarity exhibited by morphological variants. When using only an unannotated text corpus, the algorithm is unsupervised, but to improve its performance, Yarowsky and Wicentowski present ways of providing the algorithm with optional resources, thus increasing its level of supervision.

3.2 Self-Organizing Map

The Self-Organizing Map (SOM) is an artificial neural network algorithm developed by Teuvo Kohonen (1982). One of the main assets of the model is its ability to efficiently visualize data sets as two-dimensional, usually hexagonal map grids onto which the input data samples are projected. The samples, as well as the units on the map grid, are represented as *feature vectors* which consist of values for the features chosen to represent the data set. The relative distances of the samples on the resulting map reflect their similarity according to the chosen feature set, so that samples that have very similar values for the features will end up close to each other on the grid.

Being an unsupervised learning algorithm, SOM requires no teacher to define the correct output for a given input. This naturally makes it highly applicable to any set of data that hasn't been examined and classified beforehand, for example unannotated text data. Also, since the categorization of the input samples emerges from the data set itself, SOM can also be used in the purpose of finding categorizations typical for a particular set of data.

3.2.1 The algorithm

The map grid of a SOM consists of cells or nodes, each of which corresponds to a prototype vector having the same dimensions as the input sample vectors. Prototype vectors are denoted here by m_i, where i corresponds to the index of the prototype. Initially, these prototype vectors will have been initialized according to some method, usually random or linear initialization. During the training process of the SOM, sample vectors (denoted by x_j) are compared to the prototype vectors, and the Best Matching Unit (BMU) on the map grid is chosen for the sample vector. More precisely, the winning prototype vector, denoted by index c, is determined by the formula

$$c(x_j) = \arg\min_i d(x_j, m_i) \ , \tag{3.3}$$

where $d(x_j, m_i)$ denotes the distance between the sample vector x_j and the map unit prototype vector m_i. The distance between the vectors is calculated using some distance metric, typically the Euclidean distance, and the prototype vector which has the smallest distance to the sample vector at hand will be chosen as its BMU. This kind of learning process is called *competitive learning*, as the prototype vectors compete against each other over the sample.

Having found the BMU for the sample, the algorithm will adapt the BMU's vector and also the vectors of its neighboring map nodes so that they will become slightly more like the sample under consideration. The amount of adaptation of the neighboring node prototype vectors depends on their distance from the BMU; the closest neighbors are adapted more than those further away on the map. The idea is that in the early phases of the training process, the amount of adaptation will be larger, and it will affect a larger number of neighboring map nodes. This will serve to perform a rough, global initial ordering of the map. But as the learning process continues, the amount of the adaptation and the size of the neighborhood affected will decrease, subjecting the map to finer, more local ordering.

Stated in a more explicit manner, the prototype vectors are adapted according to the function

$$m_i(t+1) = m_i(t) + h_{c(x),i}(t)\,(x(t) - m_i(t)) \ , \tag{3.4}$$

where m_i denotes the ith map unit, $x(t)$ the input sample vector and t the discrete time coordinate, and $h_{c(x),i}(t)$ is the neighborhood function which determines the size of the neighborhood. The typical neighborhood function used is the Gaussian function

$$h_{c(x),i} = \alpha(t) \exp\left(-\frac{\|r_i - r_c\|^2}{2\sigma^2(t)}\right) , \tag{3.5}$$

where $0 < \alpha(t) < 1$ is the learning-rate factor and $\sigma^2(t)$ the radius of the neighborhood affected. Both the learning-rate factor and the neighborhood radius continue to decrease during the learning process. The variables r_c and r_i correspond to the locations of the prototype vectors on the grid.

Finally, a SOM can be trained with two different types of training algorithms. The training procedures described above follow the usual *sequential training* algorithm, in which sample vectors are fed to the algorithm one by one and the prototype vectors of the map are adapted after each input. This sequential training process is typically iterated thousands or tens of thousands of times, and each sample of the data set may be utilized hundreds of times during the process.

In this Thesis, however, another approach to training a SOM was adopted, namely the *batch training* algorithm. In batch training, the whole data set is presented to the map before any adaptation of prototype vectors. Each training step consists of calculating the BMUs for every sample in the entire data set and adapting the prototype vectors of the map according to the samples. In batch training, the vector adaptation is determined by the formula

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{c(x),i}(t) x_j}{\sum_{j=1}^{n} h_{c(x),i}(t)} . \tag{3.6}$$

This training step is iterated until convergence or for a sufficiently long time, each time finding the BMUs for all the training samples in the data set and adapting their vectors and neighboring vectors. Batch training has the advantage of being significantly faster than the sequential training algorithm, especially when using MATLAB functions.

Visualization and analysis of a SOM

The efficiency of SOMs in visualizing data sets owes much to the many visualization methods developed for them. These visualization methods are usually based on drawing an image of the map grid of the SOM and then presenting some type of information in the nodes of the grid. For example, if the amount of features chosen to represent the data set is small (the feature and prototype vectors are short), one might want to see a map grid where each node contains the prototype vector associated with it.

However, when the set of data gets larger, the feature set for representing the data samples also becomes larger. Typically, the feature vector for a sample may contain values for hundreds of different features, in which case visualizing the resulting SOM by presenting the prototype vectors in each node isn't really much use for anything.

One of the most commonly used visualization methods for a SOM is the U-matrix (Ultsch, 1993), which detects topological relations among nodes and infers about the structure of the input data. The U-matrix algorithm generates a matrix in which each value is a kind of a distance node, a distance measure between two adjacent map nodes. For each map node, a distance value consisting of the average on the distances to all its neighboring nodes is calculated. These values are used to draw a display in which map nodes and distance nodes alternate, and each node is coloured according to its value (see figure 3.2).

Different colour scales can be used to colour the nodes, but they all have in common the purpose of distinguishing nodes with high values from nodes with low values. The chosen colour scale is usually provided on the side of the U-matrix. In the example U-matrix in figure 3.2, the colour scale passes from blue to red, with blue marking nodes with low values and red representing nodes with high values. The regions of low-valued nodes on the U-matrix can be considered clusters, groupings of similar nodes. On the other hand, the regions of high-valued nodes, usually emerging in between the clusters, are regarded as frontiers which separate the clusters from each other. Thus, the U-matrix display shows low values inside a cluster, and high values in the areas between the clusters.

Another useful tool for visualizing the data in a SOM are the component plane representations (Kohonen, 2001) of its features. Each component plane shows the values of a particular feature throughout the map grid. The component plane images are especially useful in examining the behaviour of the data in correspondence to an individual feature from the feature set, and they may also be used as a tool for evaluating the efficiency and the contribution to the SOM of each one of the chosen features. Component planes also help detect emerging patterns of data distribution on the SOM grid (Kohonen, 2001) and correlations between the features.

A component plane representation for a feature is obtained by colouring each map node according to the value of the feature in that node. As with U-matrices, the colour scale may vary, but the adopted colour scale is usually provided together with the component plane image. Figure 3.2 shows the component plane images for the three features, named X, Y and Z, that were used to train the example SOM, also displayed as a U-matrix in the figure. As can be seen from the figure, the data samples that had for example high values in feature X were mapped to the upper half of the SOM and especially to the nodes in its left upper corner, whereas samples mapped to the lower half of the

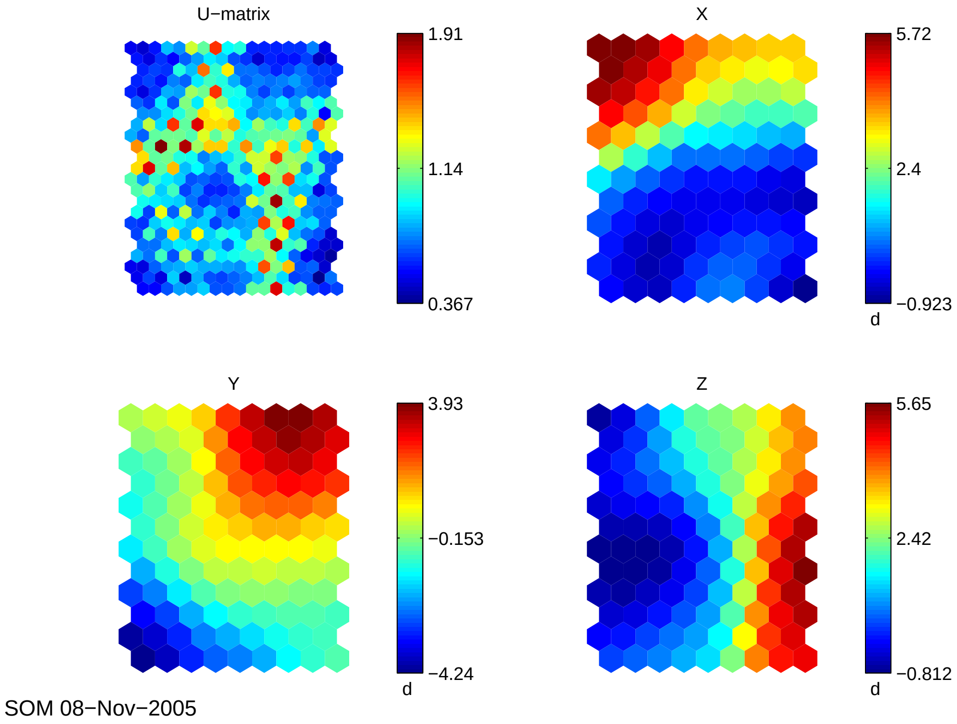

Figure 3.2: A U-matrix representation for a 12 × 9 hexagonal SOM with five clusters, and component plane representations for the features of the SOM, called X, Y and Z.

SOM seem to have in common a low value for feature X. The samples mapped to the left upper corner also seem to have relatively high values for feature Y, but low values for feature Z. This kind of analysis will help to determine what type of data was mapped to each region on the map, and what was the contribution of each individual feature in training the SOM.

3.2.2 Related work on SOMs in natural language processing

The first application area of self-organizing maps in natural language processing was speech recognition, or, more accurately, speech-to-text transformation (Kohonen et al., 1984; Kohonen, 1988). However, most of the SOM work relevant to natural language processing has been in the area of word category maps and in performing automatic statistical lexical analysis based on the SOM. The basic method for training word maps was described by Ritter and Kohonen (1989). In a word SOM, the word contexts have been reduced to a two-dimensional grid representation, in which the relative distances of data words on the map reflect the actual semantic relationships of the words in the input text (Ritter and

Kohonen, 1989). Words that are semantically or conceptually similar (words that have similar contexts in the data) will appear close to each other on the resulting word SOM, forming clusters of words. These areas or regions on a word SOM can be considered implicit categories that have emerged during the learning process (Honkela et al., 1995).

Honkela et al. (1995) have applied the SOM to analyzing contextual relations of words in Grimm tales. Miikkulainen (1990, 1993, 1997) has extensively researched the use of SOM in creating a model of story comprehension and in performing conceptual analysis of words. Miikkulainen (1997) has also presented a model of aphasia (the loss of ability to use and understand language due to brain injury or disease) based on the SOM. One advantage of the word categorizations emerging from a SOM is that they can be considered “soft”; words on a SOM are not categorized strictly as being just something or the other, but rather words are viewed as resembling each other to a certain degree, which can be either more or less.

Word category SOMs have also been applied to the Finnish language by Lagus et al. (2002). They organize the 600 most frequent Finnish verbs in a newspaper text corpus of 13.6 million word forms using their contexts in the text. In their experiments, verb categorizations by word SOMs with different kinds of features are compared to an existing semantic classification of Finnish verbs. The fact that makes this work particularly important from the point of view of this Thesis is that in one of their experiments, they use morphosyntactic features, obtained by a supervised parser for Finnish, as the features of a word SOM. In this Thesis a similar experiment is conducted, but this time with Morfessor-extracted unsupervised morphological information as features. Another difference is that here the word forms being categorized are not limited to just verbs, and the data set is also of a different type.

SOMs have also been applied the problem of word sense disambiguation (WSD). For example, Pulkki (1995) has presented a method for modeling ambiguity with SOMs. Scholtes (1992) as well as Gallant (1991) have also used neural network -based approaches to resolving ambiguity, and Mayberry and Miikkulainen (1994) present a model for lexical disambiguation in which a recurrent network parser combines one word at a time the frequency calculations of the contexts of ambiguous words, producing as output the most likely interpretation of the current sentence.

Another application area of SOMs in natural language processing has been the exploration and data mining of text documents. In the WEBSOM method (Honkela et al., 1997; Kohonen et al., 2000; Lagus et al., 2004), documents are arranged onto a two-dimensional grid based on their levels of similarities. WEBSOM can thus organize miscellaneous text documents into meaningful collections of text for exploration and search. Once one interesting document is found, other related documents, mapped close to the first one on the document map, are found as well. The WEBSOM document exploration tool has been applied to for example organizing a massive document collection of 7 million patent abstracts in Finnish (Kohonen et al., 2000).

Word SOMs

In this Thesis, the focus is on word category SOMs. Consequently, a subsection is devoted to introducing them in a more detailed manner. The general idea of word SOMs is to have implicit word categorizations emerge from the input data. The word SOMs are trained on sample word forms from the text corpus, and these same word forms are usually projected on the resulting word SOMs.

The choice of the set of training words is usually based on word form frequencies; for example, the 200 or 400 most frequent word forms in the data set might be chosen for training the SOM. This is due to the fact that with more infrequent word forms, there might not be enough occurrences of the word to yield reliable calculations based on its different contexts in the corpus.

When the set of training words has been chosen, the next step is to decide on the means of representing the words. In the approach adopted in this Thesis, the representation of a word is based on its contextual information in the text. More precisely, the representation or the feature vector of the word form consists of information on the occurrences of so-called feature elements in its context. In traditional word SOMs, the feature elements used have been whole context words. These feature words, too, are usually chosen according to the list of the most frequent word forms in the data. Infrequent word forms would probably make bad features due to their small number of occurrences in the corpus; the values for such features would be unreliable.

Next, the length of the word context, called context window, should be settled. The length of the context window determines the number of context words which will be taken into consideration when counting the occurrences of feature words in the context, and calculating the feature vectors based on these occurrences. Typically, the context window consists of 1–3 words before and after the word form under examination, but it can also be much larger, encompassing even hundreds of words.

With the sets of training words and representative features ready and the context window decided, the feature vectors for each training word may be calculated. An illustration of the process of calculating a feature vector can be found in figure 3.3. As can be seen from the figure, for each occurrence of a particular training word (in this case, the word form “lapset”), the word form and its left and right contexts are extracted from the data, and the context is searched for feature words. The binary representation of this word occurrence will contain a "1" for each feature word that was found its context, and a "0" for all the other feature words. When all the occurrences of the training word in the corpus have been processed in this manner, a feature vector for the word form under examination is calculated as the average over the binary representations of all its individual occurrences. The calculation of a feature vector x_j for the jth training word can be summarized as the formula

$$x_j = \frac{\sum_i^{N_j} x_j^i}{N_j} \,, \tag{3.7}$$

where x_j^i is the binary representation of the ith individual word occurrence and N_j denotes the number of times the word occurred in the data. The resulting feature vector is the final representation of the training word.

The length of the feature vector representing each training word depends on the length of the context window. For example, with a context window of a length of one word into both directions, and with a feature set of 100 word forms, the length of the feature vector for a training word would be 200 (each of the 100 feature words both in the left and the right context position of the word). Simplified examples of binary representations and a feature vector for the word form "lapset" can be found in figure 3.3. Since the context length in this example is 1 and there are only 6 feature words, the binary representations of individual occurrences of "lapset" as well as the final feature vector for the word form have a length of 12 components.

When the feature vectors for each training word have been calculated, the training of the SOM may begin. The training proceeds as described in Section 3.2.1, and when the word map is ready, it can be visualized with a U-matrix and component plane images. In order to be able to analyze the distribution of the training words on the map, the training words are usually also projected on the U-matrix representation to the nodes they were mapped to during the training. This is called labeling the map with words.

Traditional word category SOMs with whole context words as features, as described above, are used as a basis for comparison in this Thesis. However, the main innovation of this work are word SOMs which utilize as features morphological information obtained by an unsupervised morphology induction method. The construction of such morph-featured word SOMs will be described in more detail in the following Section.

3.2.3 Constructing word SOMs with morph features

In the case of word SOMs with morphs as features instead of whole context words, the main task still remains the same. The aim still is to produce word

			Left context						Right context					
			silloin	kohta	mutta	lähtivät	lähtivätkin	menivät	silloin	kohta	mutta	lähtivät	lähtivätkin	menivät
sillon	lapset	ottivat	0	0	0	0	0	0	0	0	0	0	0	0
kohta	lapset	näkivät	0	1	0	0	0	0	0	0	0	0	0	0
sitte	lapset	meni	0	0	0	0	0	0	0	0	0	0	0	0
pian	lapset	lähtivät	0	0	0	0	0	0	0	0	0	1	0	0
mutta	lapset	lähtivätkin	0	0	1	0	0	0	0	0	0	0	1	0
	lapset		0	0.2	0.2	0	0	0	0	0	0	0.2	0.2	0

Figure 3.3: An illustration of the process of calculating a feature vector when features are whole context words. The training word whose occurrences are under examination is "lapset" ('children' in English). Notice that the context word "sillon" and the feature word "silloin" do not match, producing a 0 in the binary representation of this occurrence of the word "lapset". This is also the case for the context word "meni" and the feature word "menivät". These context words are an instance of a somewhat slangy use of language, which is quite abundant in the children's stories corpus. The vector at the bottom of the figure represents the final feature vector for the word form "lapset", calculated as an average over the binary representations of the individual occurrences of the word form in this tiny exemplary data set.

SOMs, i.e. SOMs that organize word forms based on their contextual information in a text corpus. The set of training words, i.e. the words that are to be organized and the contexts of which will be analyzed, is chosen based on the word form frequency list just like in the previous Section. In the word SOM experiments performed in this Thesis, the set of training words usually consists of the 200 most frequent word forms in the children's stories corpus.

The set of features for representing the training words is what makes the morph-featured word SOMs different from traditional ones. In the word SOMs described in the previous Section, the words occurring in the context of a training word were matched against feature words as such, but now the context words have been morphologically analyzed by the Morfessor tool and segmented into morphs, labeled as roots, suffixes or prefixes. The set of features also consists of these morphs, again typically chosen from the top of a morph frequency list, calculated from the morphologically segmented text data. The subproblem of selecting optimal combinations of different types of morphs (roots, suffixes and prefixes) into the feature set will be addressed further in Chapter 5 of this Thesis.

With morphologically segmented context words and a set of morphs as features, the feature morphs are now matched against the segmented parts of words appearing in the context of an occurrence of the training word. As before, the length of the context window may vary. An illustration of the process of calculating a feature vector for the word form “lapset” in the case of morph features can be found in figure 3.4[4]. Where feature morphs match morphs found in the context words, the component of the binary representation of the word occurrence is marked with a “1”, and where feature morphs cannot be found in context words, it is marked with a “0”. Again, the final feature vector for the training word is calculated as an average over these binary representations of the individual occurrences of the word in the corpus.

As can be seen from the example in figure 3.4, using morph features seems to have many advantages over traditional word SOMs with whole words as features. With context words segmented into morphs, the feature vectors seem to be much less sensitive to the variation in word forms due to for example word inflection or slangy use of language. In the use of language of young Finnish-speaking children, for example, plural subjects are often followed by a verb in a singular form, even though the number of the verb should of course agree with that of its subject, at least according to the established Finnish grammar.

Also, being a highly inflectional and affixing language, Finnish word forms often include plenty of inflectional or derivational suffixes. If words like this are used as features for a word SOM as such, many word forms that are very close to the ones included in the feature set, but still different on some parts, will pass unnoticed as being actually just slightly differently inflected forms of a feature word. Using morphs as features, however, seems to be something of a cure for these kinds of ailments of the contextual feature -based representation of training words.

After calculating the feature vectors for each training word, the training of the word SOM proceeds just as described earlier. Again, the resulting SOM may be visualized with a U-matrix labelled with the training words and some component plane images showing the contribution of individual morph features to the SOM.

[4] The morphological segmentation of the context and feature words in this example was taken from a real morphological analysis of the children’s stories corpus, performed with a Morfessor variant called Categories-ML.

			Left context					Right context				
			sillo/STM	lähti/STM	meni/STM	vat/SUF	vät/SUF	sillo/STM	lähti/STM	meni/STM	vat/SUF	vät/SUF
sillo/STM + n/SUF	lapset	otti/STM + vat/SUF	1	0	0	0	0	0	0	0	1	0
kohta/STM	lapset	näki/STM + vät/SUF	0	0	0	0	0	0	0	0	0	1
sitte/STM	lapset	meni/STM	0	0	0	0	0	0	0	1	0	0
pian/STM	lapset	lähti/STM + vät/SUF	0	0	0	0	0	0	1	0	0	1
mut/STM + ta/SUF	lapset	lähti/STM + vät/SUF + kin/SUF	0	0	0	0	0	0	1	0	0	1
	lapset		0.2	0	0	0	0	0	0.4	0.2	0.2	0.6

Figure 3.4: An illustration of the process of calculating a feature vector when features are morphs and context words have been morphologically segmented. The training word whose occurrences are under examination is again “lapset”. Notice that, for example, the root of the morphologically segmented context word “sillo/STM + n/SUF” and the feature morph “sillo/STM” now match. Also, the morphs in the context word “lähti/STM + vät/SUF” now match both the feature morph “lähti/STM” and the feature morph “vät/SUF”. The vector at the bottom of the figure represents the final feature vector for the word form “lapset”, calculated as an average over the binary representations of the individual occurrences of the word form in this tiny exemplary data set.

Chapter 4

Data set

In this Chapter, the children's stories data set is described in more detail. Also, the division of the data into age categories is explained, and the preprocessing and morphological analysis procedures necessary for utilizing the corpus are introduced. Finally, a simple XML-like format for writing down new stories is suggested, in order to make easier the automatic processing of future story data.

4.1 Description of data

The children's stories corpus somewhat resembles another corpus with data from children, the CHILDES database (MacWhinney and Snow, 1985) of audio data and transcripts of conversations with young children. The corpus used in this Thesis, however, is in Finnish, and it consists of only textual data (although its texts were transcribed from stories that were originally told orally by children). Also, the children's stories corpus was collected using a special method called Storycrafting (see Chapter 2 for a description of the method), a technique of Finnish invention which seeks to promote equality in dialogs between children and adults.

The corpus consists of 2842 stories told by children aged from 1 to 14. There are stories from both boys and girls, and both from individual children and groups of children. The group stories may have been collected from small groups of for example two or three children, or they may also have been group stories by a daycare center group or even by an entire school class. Further, the groups may have been composed of only boys or girls, or they may have been mixed groups with both boys and girls.

A couple of example stories (in Finnish) can be found in figures 4.1 and 4.2, the first one a story by an individual child and the other one a group story. As even these randomly chosen examples would suggest, the group stories are

```
Fi_fi_y_19970120_1_285_2_av_11_10_s_0411_NAME

NAME 4v 11kk
[ei pvämäärää] klo 13.05
10. satu, satukirje nro 11
Tila: ruokahuone
Kirjannut: NAME, pk Kanerva, Kotka
Mukana piirustus


Tyttö meni metsään

Olipa kerran tyttö ja hän meni metsään. Ja sitten hän näki ketun. Ja sitten
tyttö sanoi: "Mikä sinun nimesi on?" Ja kettu sanoi: "Kettu." Ja sitten
tyttö meni keräämään sieniä koriin ketun kanssa. Sitten tyttö meni kotiin
ketun kanssa. Sitten tyttö muisti, että äiti on allerginen ketuille. Sitten
tyttö meni keräämään metsästä kukkia. Sitten hän teki majan metsään ja haki
kotoonta eväät. Sitten he sytyttivät nuotion. Loppu.
```

Figure 4.1: A random example of a story told by an individual child (girl, age 4 years 11 months). All names have been removed from the original file and substituted with the text “NAME”.

often much longer than the stories told by just one child. These two examples also show a glimpse of the variety of metadata that the story files may contain. The problem of metadata will be addressed further in Section 4.3.

A majority of 93%[1] of all the stories in the data set are in Finnish, but there are also some 7% in Swedish. For the experiments in this Thesis, only the 2642 stories that were in Finnish were chosen, the number of stories in Swedish being so small that reliable experiment results could not be guaranteed for them. Of all the stories, 51% were from individual girls, 38% from individual boys, 5% from mixed groups and 3% from girl and 3% from boy groups. The stories were collected between years 1994 and 2001, and 89% of the stories were told by only one child. 9% were told by small groups of 2-5 children, and 2% were from large groups of 6-20 children.

In total, the chosen 2642 Finnish stories form a text corpus of 198 036 word forms. This word count was obtained after the preprocessing of the data set, explained in Section 4.3. Due to the problem that many of the story files contain unstructured metadata, it is impossible to get a reliable word count of the data before preprocessing it.

[1] All percentages in this paragraph were taken from a handout on statistics on the children’s stories corpus, handed out by Monika Riihelä in a Children are Telling group meeting in 2004.

```
fi_fi_xy_19960912_4_257_1_HP_15_0_0605_0604_0605_0601_NAME_NAME_NAME_NAME

9609xxr.hp
12.9.1996
NAME 6 v 5 kk, 3.satu
NAME 6 v 4 kk, 1 satu
NAME 6 v 5 kk, 7. satu
NAME 6 v, 1. satu
Ryhmässä
Köpaksen päiväkoti/Masala
Kirjasi: NAME
Kuva: Kyllä
Satuketju nro 15

Lintu vauvansa kanssa

Olipa kerran lintu, jolla oli vauva. Sitte se meni etsimän ruokaa.
Sitten tuli haukka, joka yritti napata vauvaa. Ja sitten haukka pysähtyi
ja laskeutui puuhun. Ja sitten se alkoi syöksyä nopeeta, sitten se kun se
syöksyi nopeeta vauhtia se laskeutui emon päälle. Ja sitten tuli toinen
haukka, joka olikin vähän rohkeampi. Ja sitte se rohkea haukka pystyi
ottamaan emon ja toinen otti vauvan. Ja sitten se vauva rääkyi. Ja ne oli
vierekkäin ne haukat, se yksi oli aika viisas. Ne yritti panna siivet
yhteen. Se rohkeampi haukka otti emolinnun niskasta ja sitten se aikoi
laskeutua omalle maalle ja sitten se meni pesälle ja sitten se aikoi taas
mennä saalistamaan ja sitten se näki kuolleen jäniksen ja otti sen. Sitten
se kun se alko syödä sitä jänistä molemmat haukat kuoli. Ja sitten kaikki
mitkä siellä oli ne toiset linnut pääsi vapaaksi ja vauva osasi jo lentää.
Ja sitten emo kuljettti sen pesälle saakka ja emo meni hakemaan lisää ruokaa.
Ja sitten se oli niin viisas, että se saalisti matoja ja eläimiä. Ja sitten
se aikoi vielä neljä matoa saada. Se sai yhden, sitten se sai vielä toisen
ja toisen ja vielä yksi eläin ja sitten mato ja kaikki muut madot jäi
loukkuun. Sen pituinen se.
```

Figure 4.2: A random example of a story told by a group of children (boy, age 6 years 5 months; boy, age 5 years 4 months; girl, age 6 years 5 months and boy, age 6 years 1 month). All names have been removed from the original file and substituted with the text "NAME".

Finally, it should be noted concerning the nature of the corpus that it is a product of young children's use of language, and as one would expect, it contains many instances of spoken language or word forms that could be considered non-standard or slangy. This challenging, non-orthographical nature of the children's stories data set is one of the reasons why statistical methods, like the ones used in this Thesis, should prove especially useful in analyzing it. Indeed, using traditional rule-based methods in analyzing this corpus would probably have lead to many problems.

4.2 Division into categories

The data set originally came on a CD with 2842 uncategorized story files (2642 in Finnish, 198 in Swedish, 1 in English, 1 in Russian). The question soon arose about whether the story files could be divided into categories according to some criterion, so that word SOMs on different subsets of the data set could be compared to each other. Such criteria for dividing this story data set could be, for example, the age of the children who told the stories, the gender of the children, or the fact whether the story was told by an individual child or by a group of children.

Interesting as it would be to compare the resulting word SOMs of stories told by boys and girls or stories by individual children and larger groups of children, it was decided at this point to divide the stories using the age criterion. The data set was thus divided into three categories: stories by children aged from 1 to 4 years, stories by children aged from 5 to 6 years and stories by children that were older than 6 years. The age categories were determined partly based on the author's concept of what would be good and natural points for dividing children into age groups, and partly according to the fact that this particular division seemed to yield the best balance between the sizes of the different subsets.

A script was created for automatically dividing the 2642 Finnish story files into the three age categories. In the resulting division, the category of children aged from 1 to 4 years has 760 stories, the category of children aged from 5 to 6 years has 1434 stories and the last category of children aged over 6 years consists of 443 story files (see table 4.1). There was a number of files that could not be classified into any category. This is due to the fact that some group story files lacked information on the age of the children (probably because there were too many of them, for example an entire school class), and some story files just failed to conform in any way to the predefined encoding format. In total, there were 69 of such unclassified story files.

Age category	Stories	Word count	Average story length
1 to 4 year-olds	770	42 449	55.1
5 to 6 year-olds	1433	119 289	83.2
Over 6 year-olds	442	46 453	105.1

Table 4.1: The number of stories and words and the average story length in each of the three age categories. The length of the stories seems to correlate with the age of the children, which compensates for the small number of stories in the category of the oldest children.

The division of group stories into age categories was especially problematic, since a group may consist of children of different ages. For example, should a group story told by four children, aged 6, 5, 3 and 4 years respectively, be classified into the category of 1 to 4 year-olds or rather to the category of 5 to 6 year-olds? In this work, a choice was made to classify group stories into all of the categories they matched, so that the previous hypothetical example story would end up both in the category of 1 to 4 year-olds and in the category of 5 to 6 year-olds. This decision was based on the fact that at this point, it is no longer possible to distinguish the contributions of each individual child to the story, but that these kinds of group stories are rather something that emerged from the dialog and collaboration of the entire group. Other strategies for categorizing the group stories would have included for example classifying according to the age of the oldest child, or according to the age category that a majority of the children in the group fell into, or perhaps ignoring completely those group stories that were told by children from different age categories.

As a consequence of this approach adopted to categorizing group stories by children from different age categories, a total of 64 group stories were classified into two distinct age categories, and 4 stories even ended up in each of the three categories. Notice that because of this, and because of the 69 story files that could not be categorized at all, the total category word count summed over the three age categories (208 191 words) differs from the number of word forms in the total corpus of 2642 stories in Finnish, which is 198 036.

As can be seen from these numbers, the amount of stories in each subset is still not very balanced. This is due to the fact that more than half of all the stories in the data set were collected from children that were 5 or 6 years old. The explanation for this bias is that these ages are typical for children in Finnish daycare centers, which is indeed where most of the stories were collected.

The number of stories from the youngest and the oldest children being considerably smaller, it is of course justified to ask whether these categories really contain enough data for obtaining reliable results in the experiments.

Especially, the category of children aged over 6 years seems to have very few stories, only 443. But even if the stories in this category are somewhat scarce, it should be noted that the stories by older children are typically much longer than those of the children in the other age categories. Thus, the length of the stories told by children aged over 6 years compensates for the fact that there are not so many of them. In fact, when the amount of data in this age category is counted in words instead of stories, it actually surpasses the word count of the category of the youngest children, aged from 1 to 4.

4.3 Preprocessing

The data was received in `rtf` format, each story in its own `rtf` file. Consequently, the first step in preprocessing the data was converting the files into plain text files that are much easier to process automatically[2].

Next, the actual preprocessing scripts were created in Perl programming language. The basic script for processing directories of story files takes as input a directory of plain text files, applying two other scripts on the input files. The first one of these scripts attempts to remove any metadata from the story files, and the second script is the actual preprocessing tool. These two phases will be described in more detail in the following two Sections.

4.3.1 Removing metadata

The task of the first script is to strip the story files of any headers or metadata they might contain. Originally, the metadata of each story file was meant to be encoded only in the name of the file (and in the first line of the file, where the file name is repeated), and this is indeed the case in many story files. However, there were almost equally many files in the data set that also contained some additional metadata inside the actual file. Such internal metadata could include for example the name and age of the child who told the story, the name of the person who wrote it down, the name of the daycare center the story was told at, the time, space or situation the story was told in, and so on. Since the stories have been collected from so many different people, there seems to have been no standard way for including this possible additional metadata in the files and for presenting the story itself. Consequently, because of this unstructuredness of the files, automatical separation of the actual story texts from the metadata turned out to be far from a trivial task.

For this purpose, a rather straightforward snippet of a script had to be created, capable of deciding which parts of the files are probably just metadata

[2] Many thanks to Petteri Räisänen for helping with the conversion of the files.

and which parts make up the actual story. The idea of this story-digging script is that it tries to find from the file a continuous sequence of text, using the following criteria:

1. The sequence of text must be continuous. It is either just one chunk of text with no line breaks at all, or it consists of two or more chunks of text that are separated by at most two line breaks.
2. The sequence of text must be long enough, i.e. its length should exceed a certain length treshold. Short sequences of text are probably just metadata.
3. The sequence of text must not contain any of the "illegal words" specified in a special list. This list includes for example some expressions of age, different forms of the word "kirjata" or "kirjaaja" (Finnish for 'write down' or 'person who writes down'), or other such words that clearly seem to indicate that the text in question is metadata rather than story text.

It should be noted that the story-digging script is by no means perfect. Quite obviously, it does make mistakes, sometimes deciding that some of the story text looks like metadata and thus leaving it out, or, worse still, sometimes classifying metadata as story text. However, developing a really good script for this task would have taken a considerable amount of time, and since this task wasn't really one of the central aims of this Thesis, it was decided at some point to freeze the development process and just leave the script as it was then. For this reason, the performance of the script can be said to be just acceptable enough that it should not make too many classification errors, and at least it should not have very much effect on the outcome of the actual experiments performed in this Thesis.

4.3.2 Preprocessing stories

After story text has been separated from metadata, the actual preprocessing script comes into picture. It is a very basic preprocessing tool that converts all words into lower case, removes all characters other than letters, numbers and punctuation marks, replaces all numbers with the sequence "NUM" and punctuation marks with "PUNCT", and converts the normal input text into a one-word-per-line format.

Finally, there is yet one preprocessing step which involves the selection of words into the sets ot training words of the word SOMs constructed in this Thesis. Originally, it was decided that the 200 most frequent word forms in the whole children's stories corpus would be used as the training words. However, based on the examination of some early experimental word SOMs, three word forms, namely “olipa”, “kerran” and “pituinen”, were soon decided to be put on a so-called stopword list as far as the selection of training words is concerned. This was due to the fact that these three words seemed to have such distinctive feature representations, completely different from any of the other words in the training word set, that they unnecessarily decreased the representative capacity of the early resulting SOMs. Also, the three words weren’t even of particularly much interest for the kind of word analysis performed in this Thesis, since they are special story words that occur very frequently in the traditional starting phrase (“*Olipa kerran...*”; ’Once upon a time...’) and ending phrase (“Sen *pituinen* se.”; ’And that’s how the story ends.’) of a story. Thus, they ended up on a stopword list which prevents them from being selected to any set of training words of a word SOM. However, these words are still allowed to occur in the sets of feature words/morphs or as the context words of some other training words.

4.4 Morphological analysis

In order to be able to use morphological information of context words as features of a word SOM, a morphological analysis of the children’s stories corpus had to be obtained first. Thus, after freezing the development of the preprocessing scripts, three different versions of Morfessor morphological extraction method were applied to the preprocessed data[3]. The versions used were the Baseline method, the Categories-ML variant and the Categories-MAP variant (see Chapter 3 for explanation on all three variants).

The precision and recall of all three methods were calculated against the *Hutmegs* (Creutz and Lindén, 2004), the Helsinki University of Technology Morphological Evaluation Gold Standard package[4]. The Hutmegs package contains gold-standard morphological segmentations for 1.4 million Finnish words, performed by the two-level morphological analyzer (Koskenniemi, 1983) for Finnish (FINTWOL). The results of the three Morfessor variants, calculated for the words in the children’s stories corpus that could also be found in the Hutmegs, can be seen in table 4.2.

As the table indicates, the Morfessor variant that yielded the best results on the children’s stories corpus was the Categories-ML method. The performance of the Categories-MAP model on the children’s stories corpus was ac-

[3] I thank Mathias Creutz for performing the actual morphological analyses and for providing the comparisons with the Hutmegs gold standard.

[4] The Hutmegs 1.0 evaluation package for Finnish and English is available at `http://www.cis.hut.fi/projects/morpho/`

Morfessor variant	Precision	Recall
Baseline	61.5%	58.5%
Categories-ML	72.4%	60.9%
Categories-MAP	66.6%	55.2%

Table 4.2: The performance of three Morfessor variants on the children's stories corpus. Notice that the Categories-ML method surpasses the other two in both precision and recall.

tually worse than in previous experiments on Finnish data (Creutz and Lagus, 2005a). Thus, even though the Categories-ML method is slightly older than the Categories-MAP variant and even though its language model isn't considered as elegant and intuitive as that of the latter, the Categories-ML method was chosen as the Morfessor variant for obtaining a morphological analysis of this data. All the word SOMs in this Thesis that use morphological information as features were hence constructed using the morphs extracted by the Categories-ML variant of Morfessor.

4.5 Suggested standardized format for stories

In order to avoid further trouble in the automatical separation of the actual stories and metadata in story files, a simple XML-based format for recording future stories in a more standardized and structured way was suggested to the Children are Telling group of researchers in one of the meetings.

In addition to improving the machine-readability of the stories, this format would also have the benefit of moving metadata from the name of the file to the inside of it, where it is easier to handle. Of course, there is no need to completely trash the idea of storing metadata in the name of the story file even if a standardized format like this was used; the two ways of representing metadata could perfectly well be used together.

The suggested XML-based template for recording new stories can be found in figure 4.3.[5] For demonstrative purposes, only an empty example XML template for writing down a story was created. In actual use, however, a formal document type definition (DTD) would naturally be needed to accompany the template.

As can be seen, the template consists of XML-conforming pairs of beginning and closing tags. At the main level, there are three pairs of tags: the <meta> tags for the metadata of the story, the optional <otsikko> tags for a possible title of the story, and finally the <satu> tags which should contain the story

[5] At the moment, I only have a Finnish version of the template available.

```
<!-- sadun metatiedot -->
<meta>

  <!-- maan tunnus: fi = Suomi, se = Ruotsi -->
  <maa> </maa>
  <!-- kielen tunnus: fi = suomenkielinen, se = ruotsinkielinen -->
  <kieli> </kieli>
  <!-- x = poika, y = tyttö, xx = poikaryhmä, yy = tyttör., xy = sekar. -->
  <sukupuoli> </sukupuoli>
  <!-- kertojien lkm -->
  <kertojia> </kertojia>

  <!-- paikkakuntakoodi -->
  <paikkakunta> </paikkakunta>
  <!-- toimipaikkanumero -->
  <toimipaikka> </toimipaikka>
  <!-- formaatti: vvvvkkpp -->
  <pvm> </pvm>
  <!-- saduttajan nimikirj. TAI: f = kotona/vanhempi, i = lapsi itse -->
  <saduttaja> </saduttaja>
  <!-- ketjukirjeen numero (0 = ryhmä tai ei tietoa) -->
  <ketjukirje> </ketjukirje>
  <!-- monesko satu (0 = ryhmä tai ei tietoa) -->
  <satunro> </satunro>
  <!-- s = satu, a = jokin muu kuin satu (esim. teatteri, leikki) -->
  <tyyppi> </tyyppi>

  <!-- lapsia voi olla 1...* kpl, kullekin oma tietueensa -->
  <lapsi>
    <ika> </ika>
    <etunimi> </etunimi>
  </lapsi>

</meta>

<!-- sadun otsikko, ei pakollinen -->
<otsikko>

</otsikko>

<!-- itse satuteksti -->
<teksti>

</teksti>
```

Figure 4.3: A suggested XML-based template for recording stories.

text itself. The <meta> tags can contain many kinds of optional and obligatory meta information, the most important one being perhaps the <lapsi> entry or entries which in turn have the tags for recording the age and first name of the child or children who told the story. Also, since so many of the people who collected the stories seemed to have a tendency of adding some non-required extra metadata into the story files, it might be a good idea to include an additional field in the template that they could utilize for such free-form extra information.

The possibility of converting the existing story files into this XML-based format was also studied. As a result, a simple script for performing most of the conversion work was created. Utilizing the information packed into the name of the story file and the story-digging script described in Section 4.3.1, the script tries to find and assign the correct information to each pair of tags in the template. As output, it creates a filled-in XML template for the input story, with an additional pair of tags labeled <unclassified> which contain the information in the story file that the script was not able to extract and fill in to some other tags.

As the performance of the story-digging script is far from perfect, the amount of information dumped into the <unclassified> category can sometimes be rather large. But even with its faults, this rough XML conversion script does perform quite nicely the most tedious part of the conversion work. For total conversion into the XML-based format, it is left for manual effort to just go through the pre-converted files and sweep up after the automatical conversion script, keeping especially an eye on the content of the <unclassified> tags. This should save a considerable amount of time compared to having to perform the whole task manually.

Chapter 5

Experiments on feature selection

In this Chapter, the selection of morph features for a word SOM is examined in more detail. An evaluation measure is presented for enabling automatical evaluation and comparison of many word SOMs at a time, and several word SOM variants are evaluated using the method. Studying the evaluation results, the task of choosing optimal sets of morph features for a word SOM is considered, and, finally, some observations are made concerning the possible phenomena underlying the evaluation performance of the different word SOM variants.

5.1 Word SOM parameters

As explained in Chapter 3, there are several parameters to be decided when constructing a word SOM, be it a traditional word SOM with whole context words as features or one with morph features. The parameter values chosen for the word SOMs used in the experiments of this Thesis are described in the following.

All word SOMs were trained using the SOM Toolbox package (Vesanto et al., 1999) for MATLAB, using the batch version of the training algorithm. The neighborhood function used was Gaussian, and typical learning rate and neighborhood radius parameter values were used.

The sets of training words of the word SOMs evaluated in this Chapter consist of the 200 most frequent word forms in the whole children's stories corpus. All word maps presented in this work are of a size of 14 × 10 units, which means that there were slightly more than one training word per each SOM node. This gives the word SOM enough resolution for a comfortable analysis of the map; with a smaller map size, too many word forms would have been mapped to single nodes, making the manual examination of the map more difficult.

As for the size of the context window, in this Thesis a window of a length of one word into both directions was adopted. In other words, only the word forms occurring immediately before and after the training word under consideration are examined for feature elements. The feature sets of the word SOMs vary from experiment to experiment, as the main purpose of this Chapter is to evaluate word SOMs with different types of features and to find the optimal sets of features.

5.2 Evaluation measure

For this Thesis, an evaluation measure for automatically evaluating the quality of word SOMs was developed. Since it wasn't obvious which combinations of feature morph types would yield the best resulting word maps, such an evaluation measure was needed in order to save the labor of manually comparing several SOM variants and to get reliable information about the best morph feature sets for constructing morph-featured word SOMs.

5.2.1 Manual baseline categorization of word forms

The idea of the evaluation measure that was adopted is to use part-of-speech information of the 200 most frequent word forms in the whole children's stories data set. These same 200 most frequent words are also used in training the word SOMs based on the whole data set and projected on the resulting maps. Part-of-speech information of the word forms was chosen as the basis of the evaluation measure because part-of-speech classes are the traditional way of categorizing words. Even if the word categorizations emerging from the word SOMs trained in this Thesis do not necessarily correspond to any established linguistic theory on word categorization, it is nevertheless difficult to imagine that there could be such emergent word categorizations that are good conceptual representations of the data but that have nothing to do at all with the established theory. The traditional part-of-speech classification of words is thus regarded here as a kind of a minimum requirement for the emergent word SOM categorizations: good emergent word categorizations probably have at least something in common with the established part-of-speech -based theory of organizing words into classes.

The 200 most frequent word forms of the data set were thus manually analyzed, and each word form was given a list of all possible parts-of-speech that it could belong to, according to a recently published new descriptive Finnish grammar called "Iso suomen kielioppi" (Hakulinen et al., 2004). Notice that these part-of-speech lists of each word form may also include such parts-of-speech or senses that the word in question did not occur in in this particular data set, but that could be valid parts-of-speech for the word form given some other data set. The part-of-speech lists were thus intended to be as comprehensive and exhausting as possible. An example excerpt from the list of the manually classified 200 most frequent word forms can be found in figure 5.1.

It should also be noted that the grammar that was used as a source for finding the parts-of-speech is of a descriptive nature rather than normative, and, following the descriptive tradition, many words could not be strictly classified as belonging to just one or the other part-of-speech. Rather, the resulting classification was soft, meaning that a word could belong to more than one part-of-speech, and in some cases, one part-of-speech classification in the list of a given word could contain more than one distinct part-of-speech. For example, the word “toinen” received a list of four different classifications, of which the first one was “adjective/numeral” (meaning in English: ordinal number ’(the) second’). Finnish ordinal numbers are used in such an adjectival way that it is difficult to say whether they should be considered as belonging to a separate class of numerals at all or just a special case of adjectives. Thus, it is justified to give the word form “toinen” used in this ordinal number sense a classification as something that is both an adjective and a numeral, or as something being in between these two parts-of-speech; hence the compromising classification “adjective/numeral”. Other part-of-speech classifications of the word form “toinen” included meanings like “pronoun” (reciprocal pronoun ’(each) other’) and “adjective/pronoun” (quantitative, indefinite or comparative prounoun ’other’).

Finally, it is worth noticing that since the data set consists of stories told by small children, the youngest of them being only one year old, many words and word forms in the data set contain spoken language and aberrations from the commonly accepted Finnish orthography that adults would probably consider as “errors” or at least slangy use of language. A traditional normative grammar would normally dismiss many of them as just non-orthographical or ungrammatical word forms. Luckily, the descriptive Finnish grammar used as the classification source for the words was constructed using real text corpora, both in written and in spoken language, and thus it contains many examples of spoken or otherwise slangy language use as well. During the manual classification of the 200 most frequent word forms, in the cases where even the descriptive grammar failed to present the needed examples of such non-orthographical language, word forms were classified like they were their normal orthographical versions instead (but labeled as being “slang”, like all non-orthographical word forms encountered in this set of 200 words).

```
1098 lähti
1. VERBI_ITR : dynaaminen, konkreettinen, siirtymis- tai asettumisverbi
977 mutta
1. PARTIKKELI : 5-2-A yksiosainen rinnastuskonjunktio
916 siellä
1. ADVERBI/PRONOMINI : 7-A/2-A-1 lokatiivinen demonstratiivinen proadverbi
889 kotiin
1. SUBSTANTIIVI : 1-A jaoton yleisnimi
2. ADVERBI : 1-C paikan adverbi, muu sijainti
782 näki
1. VERBI_TR : dynaaminen, mentaalinen havaintoverbi
741 äiti
1. SUBSTANTIIVI : 1-A jaoton yleisnimi
717 loppu
1. SUBSTANTIIVI : 1-A jaoton yleisnimi
2. VERBI_ITR (slangia) : dynaaminen, konkreettinen tilanmuutosverbi
631 joka
1. PRONOMINI : 6 relatiivipronomini
2. PRONOMINI : 7-B-2 distributiivinen universaalinen kvanttoripronomini
599 pois
1. ADVERBI : 1-C paikan adverbi, muu sijainti
582 sitä
1. PRONOMINI : 2 demonstratiivipronomini
2. PRONOMINI (slangia) : 1-B anaforis-deiktinen persoonapronomini
537 menivät
1. VERBI_ITR : dynaaminen, konkreettinen siirtymis- tai asettumisverbi
531 kaikki
1. PRONOMINI : 7-B-1 yleiskäyttöinen universaalinen kvanttoripronomini
508 kissa
1. SUBSTANTIIVI : 1-A jaoton yleisnimi
486 sinne
1. ADVERBI/PRONOMINI : 7-A/2-A-1 lokatiivinen demonstratiivinen proadverbi
481 sieltä
1. ADVERBI/PRONOMINI : 7-A/2-A-1 lokatiivinen demonstratiivinen proadverbi
470 koira
1. SUBSTANTIIVI : 1-A jaoton yleisnimi
467 minä
1. PRONOMINI : 1-A puheaktin persoonapronomini
438 söi
1. VERBI_TR : dynaaminen, konkreettinen nauttimisverbi
415 pikku
1. ADJEKTIIVI : 4 taipumaton adjektiivi
```

Figure 5.1: An excerpt from the list of the manually classified 200 most frequent word forms in the whole children's stories corpus. The numbers on the left of the word forms are their frequencies in the corpus. Notice that a single word form can have more than one part-of-speech classifications, and that some classifications are compromises between two separate parts-of-speech.

5.2.2 Evaluation algorithm

With the manual classification of word forms done, the base for the evaluation measure was ready. Intuitively, the idea of the evaluation measure is to find out how well or tightly word forms are clustered on each particular word SOM according to the part-of-speech classifications. Ideally, word forms close to each other on the map should have some parts-of-speech in common; for example, nouns would form a tight group with each other, separated from all other groups of word forms on the SOM, as would also verbs, adjectives etc.

The evaluation measure calculates for each word form the percentage of the words in the same or the immediately neighboring map nodes that had one or more parts-of-speech in common with the word-form in question. After comparing the part-of-speech lists of each word form with the lists of word forms in the neighboring map nodes, an average percentage for the whole SOM is calculated over the results of each individual word form. Also, in order to rule out the possibility of chance, the final results for each type of word SOM are calculated over the individual results of 100 randomly initialized word maps of that type.

In more detail, the evaluation loop for one word SOM type proceeds as follows:

1. Train a word SOM that is to be evaluated.[1]

2. For each word form of the SOM, find the best matching unit (BMU), e.g. the map node the word was mapped to.

3. Find the immediately neighboring nodes of this best matching unit.

4. Find the words that were mapped to either the same map node as the word form under examination or to one of the immediately neighboring map nodes.

5. Of these words, find the ones that have at least one part-of-speech classification in common with the word form under consideration, and calculate their percentage of all the word forms mapped to these nodes. If none of the neighbor words have any part-of-speech classifications in common with the word form, the percentage is zero.

6. Calculate an evaluation result for this word SOM by taking the average percentage over the results of all individual word forms.

[1] It should be noted that all word SOM variants in the evaluation loop are initialized using the same random seed, in order to eliminate the possibility of some maps getting a better random initialization than others and thus faring better in the evaluation process.

7. Calculate an evaluation result for this word SOM type by taking the average percentage over the results of 100 randomly initialized word maps of this type.

It is possible that some of the word SOM variants that are being evaluated were not trained exactly on the same 200 most frequent word forms of the whole data set that were manually classified and used as the basis of evaluation. For example, a word SOM may have been trained on only a portion of the whole data set instead of all of the story data, e.g. on the portion of stories told by children aged from 1 to 4, and thus the 200 most frequent word forms in this portion of the data set may differ from those in the whole data set. For this reason, the evaluation algorithm checks the word lists, using for evaluation only those word forms in the training word list of the word SOM being examined that can also be found on the original list of manually classified 200 most frequent word forms in the whole story corpus.

Of course, if the training word list of the particular word SOM being evaluated should differ greatly from the original list, the evaluation results will naturally be affected and become less reliable. To cover more of the list of the most frequent word forms in the data and in its subsets, more words would have to be manually classified. For the purposes of this Thesis, however, this did not seem necessary. Almost in all evaluation cases the training word lists of particular word SOMs were identical or at least very close to the original list of manually classified examples, so the results can be considered quite reliable.

The evaluation measure was implemented as a Matlab program that runs the evaluation algorithm on several different types of word SOMs with morphs as features, and also on a couple of traditional word SOMs which use whole context words as features. The experiments will be described in more detail in the following Section.

5.3 Feature selection experiments

The evaluation algorithm was run on quite a few different word SOM variants which were all constructed on the whole children's stories data. These variants included:

1. Two word SOMs using as features the first 200 or 100 morphs from the frequency list of all morphs in the whole data set.

2. Two word SOMs using as features the first 200 or 100 morphs from the frequency list of root morphs (morphs that Morfessor labeled with "STM") in the data set.

3. Two word SOMs using as features the 200 or 100 most frequent root morphs plus the 20 or 10 most frequent suffix morphs (morphs that Morfessor labeled with "SUF") in the data set.

4. A word SOM using as features the 200 most frequent root morphs plus 23 hand-picked suffix morphs.

5. Two traditional word SOMs using as features the 200 or 100 most frequent whole context words in the data set.

6. A word SOM using as features the 80 most frequent suffix morphs in the data set.

Notice that in all the word SOMs that utilize morph features, the morphs were extracted by the Categories-ML variant of Morfessor, described in Chapter 3.

Of most of the SOM variants, there were thus two versions in the evaluation: one with around 200 features and another with about 100. The feature set of the first SOM is composed of the 200 most frequent morphs from the frequency list of all Morfessor-extracted morphs of the whole data set. In detail, this list of 200 morphs consists of 145 root morphs, 52 suffixes and 3 prefixes. Two word SOM variants had only root morphs for features, and the next three experiments are kind of compromises between using just root morphs and root morphs together with suffixes. After noticing that the evaluation results actually showed a slight declination when suffixes were introducecd to the feature set, a variant with 200 root morphs and 23 hand-picked suffixes was also added to the evaluation process. Here, hand-picking simply means that from the suffixes frequency list, only a handful that looked like especially good and natural Finnish suffixes were chosen for features[2]. Finally, to compare the evaluation results of word SOMs with Morfessor-extracted morphs as features to traditional word SOMs, a couple of word SOMs with whole context words as features were added to the evaluation process.

5.3.1 Experiment results

The evaluation results for these SOM variants can be found in table 5.1. As explained previously, the results are average percentages over 100 randomnly initialized word SOMs of the type. In turn, the result percentage of one individual word map indicates the average portion of words in the immediate

[2] Studying the effect of prefixes as features of word SOMs was not considered to be of importance, since the number of the prefixes that Morfessor extracted from the data was very small (only 19 prefixes were found) and many of them were highly infrequent. This goes along with the fact that Finnish morphology is extremely suffix-centered of nature.

Word SOM variant	Eval. result	S.d. of results
200 most frequent morphs from ALL-list	60.41%	1.48%
100 most frequent morphs from ALL-list	60.75%	1.55%
200 most frequent morphs from STM-list	62.91%	1.63%
100 most frequent morphs from STM-list	63.11%	1.56%
200 most frequent STM-morphs + 20 most frequent SUF-morphs	61.29%	1.60%
100 most frequent STM-morphs + 10 most frequent SUF-morphs	61.05%	1.40%
200 most frequent STM-morphs + 23 hand-picked SUF-morphs	62.25%	1.44%
200 most frequent whole context words	54.76%	1.63%
100 most frequent whole context words	54.43%	1.57%
80 most frequent morphs from SUF-list	43.75%	1.43%
Baseline similarity	22.51%	-

Table 5.1: The results of the word map quality evaluation measure for different word SOM variants, calculated as average percentages over 100 randomly initialized word SOMs of the type. Notice that all word SOMs clearly outperform the baseline similarity, and that all but one of the SOMs that had morph features fared better than traditional word SOMs with whole context words as features. "S.d." denotes the standard deviation of the evaluation result percentages.

neighborhood of a given word form that have at least one part-of-speech in common with the word form in question. For comparison, the table also includes a so-called baseline similarity. This baseline similarity counts for every training word the percentage of other training words sharing at least one part-of-speech with it, and again the result is averaged over all the words in the training set. In practice, this corresponds roughly to the idea of a SOM organized in a completely random fashion.

As can be seen from table 5.1, all word SOMs clearly outperformed the rather crude baseline similarity measure, whether they had morphs or whole context words as features. This indicates that all the word SOMs that were evaluated succeeded in creating categorizations which surpass in quality an entirely random organization of the data.

The best result was yielded by a word SOM with only root morphs as features. When also suffixes were added to features, the evaluation results seemed to slightly decline, but if the added suffixes were hand-picked, the result was very close to the best SOM variants that used only root morphs. If the feature morphs were chosen from the list of all morphs, the evaluation results again would deteriorate. However, the evaluation process clearly shows that all but one word SOMs that used morphs as features fared notably better in the evaluation than the traditional whole context words -based word SOMs.

So why did the results decline when also suffix morphs were included in the feature set, compared to using just root morphs? And why did the word SOM with suffix morphs alone as features fare as badly as it did? One answer may lie in the nature of the evaluation measure itself. As explained previously, the evaluation algorithm uses as criterion the similarity of part-of-speech classes of words. If part-of-speech classes are used as the basis of evaluation, and if the evaluation results get worse when suffix morphs are introduced to the set of features, it may be that suffix morphs encode some entirely different characteristic of words than their part-of-speech class, making this particular type of evaluation measure a poor choice for evaluating word SOMs with plenty of suffixes as features. It would indeed seem natural that an evaluation measure based on parts-of-speech would favour word SOMs with root morph features and penalize the use of suffixes, since usually it is the case that roots of words carry the basic information on their part-of-speech classes whereas prefixes and suffixes are mostly involved in producing different inflected forms of the words. Of course, there are some affixes that are also used in deriving new word forms and they can also change the part-of-speech of the original word, so this distinction of roots marking part-of-speech information and affixes marking the inflection of a word is clearly an oversimplification of the situation.

Another explanation may be that not all of the Morfessor-extracted suffix morphs that were used are of such great quality, even if they were in the top 20 on the suffixes' frequency list. For example, some of the top suffixes were very short, composed of only one letter. Even if one-letter suffixes are by no means rare in the Finnish language – and many of the high-frequency one-letter suffixes were indeed completely acceptable Finnish suffixes – the seemingly high frequencies of these suffixes do not necessarily reflect their actual frequencies in the data set. This is due to the fact that while the Morfessor segmentation tool generally does a great job on Finnish morphology, it can sometimes get a bit carried away with oversegmentation, splitting also correct but less frequent longer morphs into highly frequent one-letter morphs. This explanation seems to be supported by the observation that hand-picking to the feature set some longer morphs that were unlikely to have been involved in occurrences of oversegmentation seemed to yield almost as good results as using just root morphs as features.

Interestingly, the evaluation results also seem to verify the hypothesis that using morphological information of context words could indeed result in improvement of quality of word SOMs. Compared to the evaluation results of

54.76% and 54.43% of the traditional word SOMs based on whole context words, all morph-featured word SOMs (except for the suffix-features-only experiment) fared distinctly better in the evaluation.

It is especially interesting to compare the two traditional whole-context-words-as-features word SOMs with those that had the 200 or 100 most frequent root morphs for features, since word roots are intuitively much more closely related to whole words than are for example suffixes. Using word roots as features instead of whole words does indeed seem to greatly improve the quality of a word SOM, at least according to this part-of-speech -based evaluation measure. This is obviously due to the fact that when word forms are morphologically segmented into roots and affixes, many inflected word forms that were previously counted separately now fall under the same word root, prefixes and suffixes having been clipped off. If, for example, the singular and plural forms of a noun are now reduced to the same root, they no longer affect the training of a SOM as two separate features but rather as just one (their common root)[3]. This means that, the redundant singular and plural endings having been removed, the inflected forms of this particular noun now take up less space in the feature set, resulting in the feature set being more "packed" with information. This, in turn, leads to better clustering of word forms into part-of-speech -based groups on the word SOM.

Finally, when comparing word SOM pairs having feature sets of different sizes but with the same types of features, there did not seem to be much differences in the evaluation results. In all four cases where there was a pair of experiments on a certain feature type combination (one with a set of around 200 features and the other with around 100 features), the difference between the evaluation results of the pair was not of great importance. Looking at the standard deviations of the results of the SOM pairs, it can be seen that the seeming superiority of the one or the other feature set size is probably just a coincidence.

5.4 Conclusions

An evaluation measure was developed for automatically evaluating word SOMs during the task of finding the best morph feature sets for constructing word SOMs with morph features. The method is based on comparing the part-of-

[3] The suffix indicating the plural will of course also be recognized as a morph, and if suffixes are accepted into the feature set then it would probably be frequent enough to be chosen for feature, too. But instead of counting separately singular and plural forms of high-frequency nouns in the feature set, we would now have just the roots of these frequent nouns and one or two (due to morphological variation in Finnish plural endings) features for plural suffixes.

speech information of the training words of a word SOM with the parts-of-speech of the words in the same or the immediately neighboring map nodes, and it uses a manually classified list of the 200 most frequent word forms in the whole children's stories corpus as a basis.

As the results of the evaluation experiments described in this Chapter indicate, utilizing Morfessor-extracted morphological information as features for a word SOM does indeed seem to improve the quality of the resulting word SOM. All morph-featured word SOMs fared clearly better in the evaluation than the traditional whole context word -based SOM.

However, not all morph-featured word SOMs scored equally good results. Thus, when constructing word SOMs with morphological information as features, it is important to consider the type of the morphs that are chosen to the feature set, and to try to find a combination of the different morph types (roots, suffixes, prefixes) that is optimal for the task at hand. It seems that for the word SOMs of this Thesis, trained on the children's stories corpus, choosing only root morphs to the feature set yielded the best evaluation results. When also affixes were included in the set of features, the evaluation results displayed a slight declination. This may be explained by the nature of the evaluation measure that was adopted, or perhaps by the quality of the Morfessor-extracted morphs that were used as features.

Chapter 6

Data analysis using the SOM

In this Chapter, the children's stories corpus is analyzed from a few different points of view by using word SOMs with morph features. First, a word SOM trained on the whole corpus is examined for emergent word categorizations. Some of the word categories that emerged from the word SOM on the whole story data are studied in more detail, together with component plane images of some features dominant for the categories. Then, word SOMs are trained on the stories of each of the three age categories, explained in Chapter 4, and they are analyzed and compared to the word SOM on the whole corpus. The objective of this comparison is to study the differences and similarities in the use of language of the stories told by children from different age categories.

6.1 Analysis of children's stories data

For analyzing the whole children's stories corpus and its emergent word categorizations, a word SOM was trained with 220 Morfessor-extracted morph features[1]. The feature set included the 200 most frequent root morphs in the whole data set (labeled "STM" by Morfessor), as well as the 20 most frequent suffixes (labeled "SUF"). From among the different feature set variants evaluated in Chapter 5, this particular combination of root morphs and suffixes was chosen for the final analysis because it fared quite well in the evaluation, and also because a word SOM analysis with also suffix morphs as features was considered more interesting than one with just roots.

As for the set of training words, the 200 most frequent word forms in the whole corpus were chosen. The word SOM was thus trained on 200 words, each represented as a feature vector with 440 features (the 220 feature morphs in

[1] As in the evaluation phase, the morphs were extracted with the Categories-ML variant of Morfessor, which yielded the best precision and recall on the children's stories corpus when compared to the Hutmegs Gold standard (see Section 4.4).

both the left and the right context of the training words). As in the evaluation phase (see Chapter 5), the map was constructed using the SOM Toolbox package (Vesanto et al., 1999) for Matlab.

Figure 6.1 shows the U-matrix display of the resulting word SOM, with the 200 training words projected on it. A black-and-white representation was preferred over a more colourful one to make more visible the words that the map was labeled with. In the figure, some clusters of words on the word map have been highlighted by manually drawing circles with different colours around them. These six clusters will be examined in more detail in the remainder of this Section.

Close to the lower edge of the map, towards the left corner, there is a cluster of words highlighted with **red**. This group of words contains exclusively nouns: "kettu" ('fox'), "äiti" ('mother'), "prinssi" ('prince'), "hevonen" ('horse'), "tyttö" ('girl') and so on. More specifically, the cluster is composed of nouns which are all in the nominative case and which are probably typical characters in the stories of the corpus. The hypothesis is that these story character nouns, or *agent* nouns, are probably used in the syntactic role of subjects in the sentences of the stories. The term "agent" is used here to refer to entities, usually animate, that are capable of initiating or performing an action of some kind.

To get a more detailed view on the linguistic contexts of these agent nouns, the component plane images of the most frequent feature morphs were manually compared to the U-matrix. A number of features emerged that seemed to be particular to the word forms mapped to this area in the lower left corner of the SOM. The component plane images of some of such features can be found in figure 6.2. The colours used in the images scale from dark blue, denoting values that are close to zero, to red, marking high values.

For example, it seems that the nouns in the agent cluster were often preceded by words that contained the Morfessor-extracted root morphs "iso/STM" or "yksi/STM". "Iso" is a common Finnish adjective meaning 'big', and "yksi" is a numeral meaning 'one'. In the children's stories corpus, and in any use of language of Finnish children and youth, the word "yksi" is also often used adjectivally as a kind of an indefinite article (which the Finnish language officially lacks), with very similar semantics as the English indefinite article 'a' or 'an'. Both of these feature morphs are thus typical adjectival attributes of a noun. There were also many other noun attribute feature morphs found in the left context of the words in the agent cluster, for example "pien/STM" ("pieni" or 'small'), "pikku/STM" (also 'small') and "toin/STM" ("toinen" or 'other').

When looking at the features that dominated the right context of the agent nouns, it becomes obvious that these nouns really are used as subjects in the stories. The feature morphs found in the right context included many verb

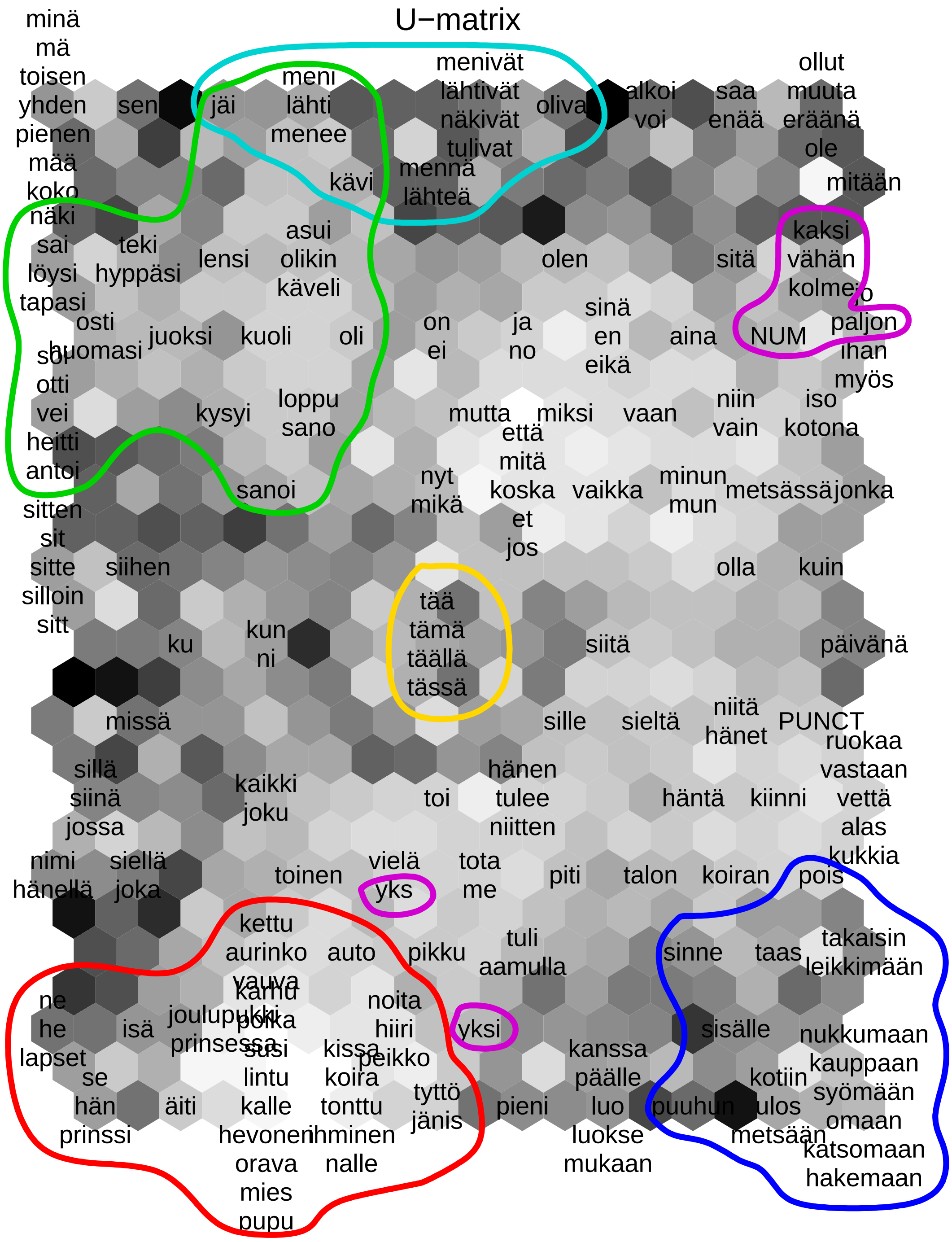

Figure 6.1: The U-matrix representation of the word SOM on the whole children's stories corpus. Some clusters of words on the map have been manually highlighted. See text for more discussion on the clusters.

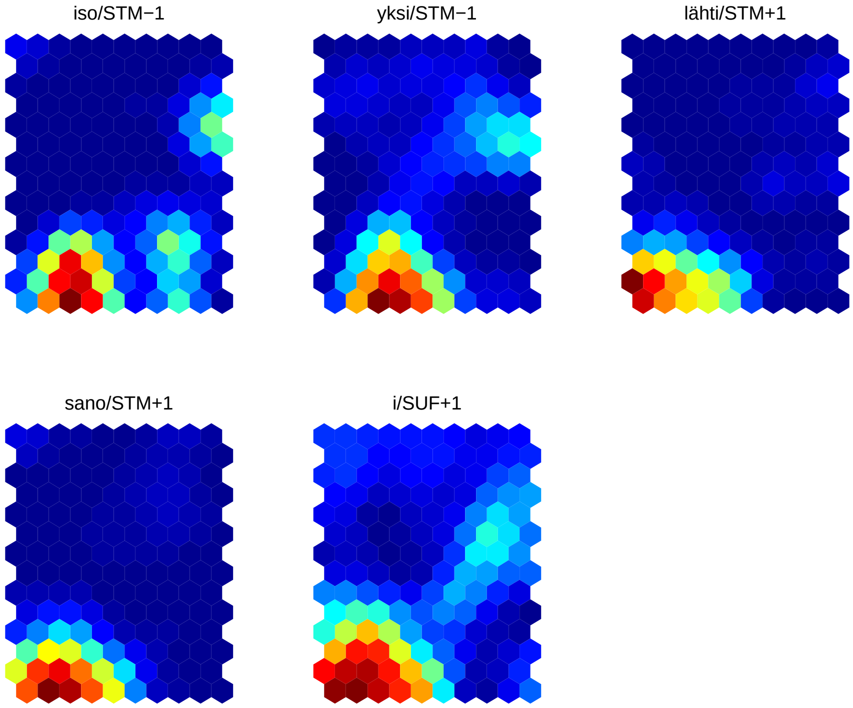

Figure 6.2: Component plane images of some feature morphs active in the agent noun cluster. Features found in the left context of the word forms are marked with "-1", and features in the right side context of the word forms have a "+1" attached to the feature name.

roots, for example the "lähti/STM" ('he/she left') and "sano/STM" ("sanoi", or 'he/she said') displayed in figure 6.2, as well as some conjugational endings Morfessor extracted from verbs, like the imperfect tense suffix "i/SUF".

In the upper left quarter of the map in figure 6.1, there is a group of words highlighted with a **green** circle. These word forms are all verbs in the 3rd person singular imperfect tense: "löysi" ('he/she found'), "otti" ('he/she took'), "juoksi" ('he/she ran'), "sanoi" ('he/she said'), etc. There also seems to be two subgroups inside this cluster: on the leftmost edge of the cluster, the verbs seem to be transitive, i.e. they usually take a direct object of some kind (for example "tapasi" or 'he/she met'), whereas the rest of the verbs seem to be more or less intransitive, i.e. verbs that do not normally take direct objects (for example "kuoli" or 'he/she died').

As with the agent nouns, this cluster of imperfect tense verbs was also studied more closely by examining the component plane images of some features that were active in this area of the map. Some of such features can be found in figure 6.3. In the left side context of these imperfect tense verbs, there seemed to be mainly nouns, for example the root "karhu/STM" ('bear') displayed in the figure, and also words that resemble and replace nouns, i.e. pronouns like

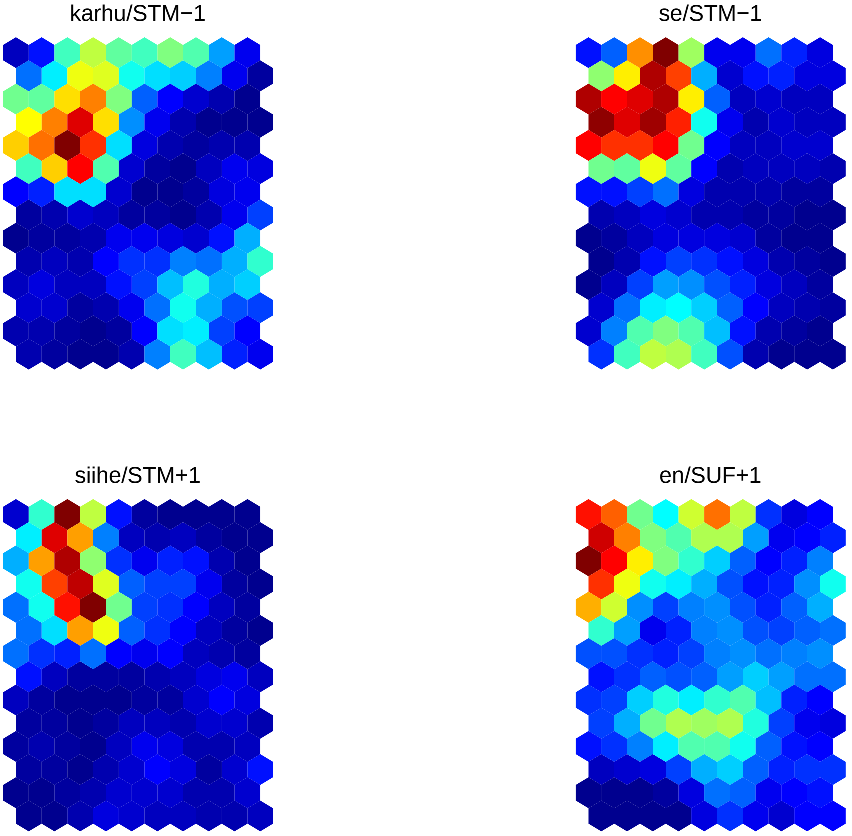

Figure 6.3: Component plane images of some feature morphs active in the imperfect tense verb cluster. Features found in the left context of the word forms are marked with "-1", and features in the right side context of the word forms have a "+1" attached to the feature name.

the feature "se/STM" ('it', or a slangy way of expressing 'he/she'). These nouns and pronouns are obviously words that were used as the subjects of the verbs in this verb cluster. Other features active in the left context of imperfect verbs included for example the root morphs "äiti/STM" ('mother'), "kissa/STM" ('cat'), "koir/STM" ("koira" or 'dog'), "pupu/STM" ('bunny'), "tyttö/STM" ('girl'), "noita/STM" ('witch'), and so on.

If the left context was mainly dominated by nouns, the right side context of the imperfect tense verbs seemed to have more variation in features. A couple of these features active in the right side context are presented in figure 6.3: namely the root morph "siihe/STM" ('(to) it' or '(to) there') and the suffix "en/SUF", a Morfessor-extracted morph marking the genitive case. In fact, the features in the right side context of the verbs seem to be what caused the emergence of the two subclusters of transitive and intransitive verbs inside the larger imperfect tense verb cluster; for example, these two example features have quite distinct distributions as far as the imperfect tense verb cluster is concerned. The feature “siihe/STM” seems to be active in the area of the intransitive verbs, probably because of expressions like “lensi siihe+n”, “kuoli siihe+n” or “sanoi siihe+n”, quite frequent in the children’s stories. The suffix feature “en/SUF”, on the other hand, seems to be especially active among the transitive verbs, obviously marking the genitive case in the direct objects of these verbs; “tapasi pien+en”, “vei hevo+s+en”, “söi yhd+en”, etc.

A third interesting cluster of words, highlighted with the **magenta** colour, is located near the upper right corner of the map. These words all express some kind of quantity: “kaksi” (’two’), “kolme” (’three’), “vähän” (’little’ or ’a little’), “paljon” (’a lot’) and “NUM”, denoting any numerals marked with numbers (’23’) instead of letters (’twenty-three’) in the children’s stories corpus. Notice that there are also two other quantity words, namely “yks” and “yksi” (’one’), located at the borders of the agent noun cluster. The quantity words inside the main cluster seem to be characterized by left context features like linking verbs connecting a subject with its predicate (for example the copula verb feature “on/STM” or ’is’, displayed in figure 6.4), and right context features that involve nouns in partitive case, for example noun roots “karhu/STM” (’bear’) (see figure 6.4), “tyttö/STM” (’girl’), “hevo/STM” (“hevonen”, or ’horse’) and “tonttu/STM” (’elf’) as well as partitive-marking suffixes like “a/SUF” (see figure 6.4), “ä/SUF” and “ta/SUF”.

The two quantity words “yks” and “yksi” located outside the main quantity word cluster, on the other hand, seem to have quite different feature distributions. Even if these two words do seem to display activity for noun roots in the right side context like the other quantity words (see for example the feature “karhu/STM” in figure 6.4), they lack the presence of other features characteristic of the words inside the main quantity word cluster. This is apparently due to the fact that in the Finnish language, the syntactic agreement between the numeral “yksi” (’one’) and the word it is associated with is completely different from the agreement of other numerals. Like in English, ’one bear’ would be “yksi karhu”, but in the case of ’two bear+s’, the noun takes a singular partitive case instead of plural nominative, “kaksi karhu+a”. Also, when further inflection is involved due to for example the fact that the phrase is used as an object of some verb, e.g. “näin kaksi karhu+a” (’I saw two bears’), the distinct behaviour of the numeral “yksi” takes it even further away from the usual linguistic context of other quantity words: “näin yhd+en karhu+n” (’I saw one bear’), with a genitive case instead of partitive or nominative.

Also, as was noted in the analysis of some of the features active in the agent noun cluster, the use of the word “yksi” has evolved into something resembling a Finnish indefinite article, at least in the more informal use of language. Used

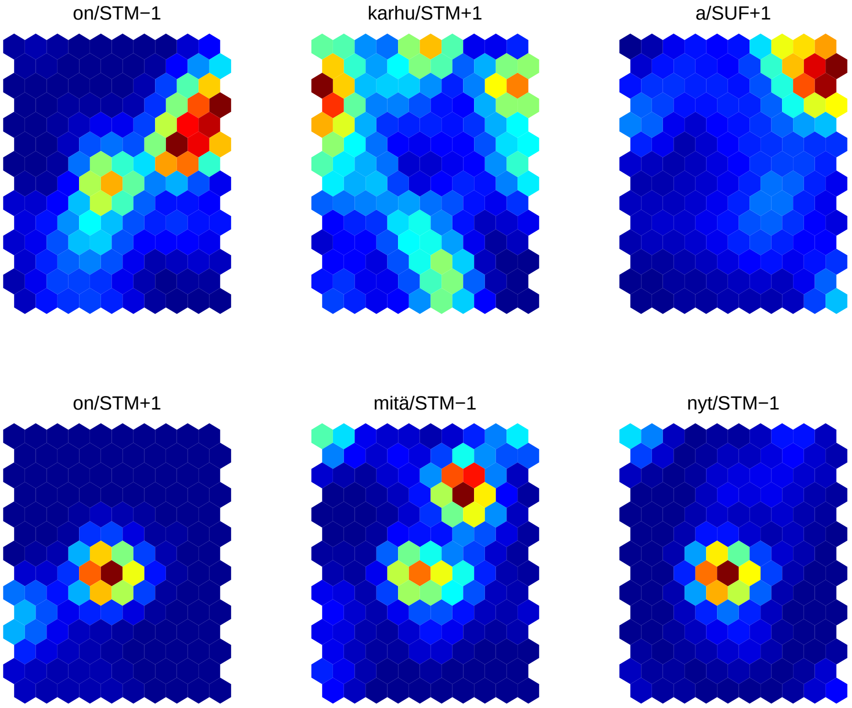

Figure 6.4: Component plane images of some feature morphs active in the quantity word cluster or in the this-cluster. Features found in the left context of the word forms are marked with "-1", and features in the right side context of the word forms have a "+1" attached to the feature name.

in this sense instead of its original numeral meaning of 'one', the word "yksi" is semantically close to the Finnish indefinite existential quantitative pronoun "eräs". This other meaning of 'yksi", quite common in spoken Finnish, naturally contributes to setting it apart from the other quantitative words found in the children's stories corpus, explaining their different locations on the word map.

Another interesting cluster seems to have emerged right in the middle of the map. This cluster, highlighted with **yellow** and clearly separated from all the word forms of the rest of the word map, contains inflected or slangy forms of the word "tämä" ('this'). The component plane images (see figure 6.4) reveal that in the children's stories corpus, these word forms have been extensively collocated with words containing particular morph features: "on/STM" ('is') both in the left and the right side context, and "mitä/STM" ('what') and "nyt/STM" ('now') in the left side context. These features seem to imply that in the children's stories corpus, there are some expressions involving the different forms of the word "tämä" that are typical for children telling a story. For example, especially with the younger children, expressions like "toi on isäpilvi ja tää on pikkupilvi" ('that one is a father cloud and this one is a little cloud'), "täällä on lohikäärme" ('there is a dragon here'), "ja sitte tässä on sateenkaari" ('and then, here there is a rainbow'), "tämä meni kertomaan isälle" ('this one went to tell dad') or "mitä tämä on?" ('what is this?') are rather frequent[2].

Expressions involving these kinds of deictic references to extralinguistic objects and circumstances suggest that children feel that the characters and objects in their stories are very close to themselves, almost as if they were present in the situation where the story was told. This resembles the way young children communicate their actions and feelings when they are playing together, for example when they are playing house or assuming a role as one of their dolls or action figures. The children seem to identify with the characters of their stories much in the same way that they identify with their characters and dolls during play.

Finally, there are two separate but semantically connected clusters of words on the map, marked with **blue** and **turquoise blue**. The words in the cluster highlighted with blue are typical location words that the children used in their stories. The group contains both nouns, for example "kotiin" ('(to) home'), "kauppaan" ('to store') or "metsään" ('to forest'), and also MA-infinitive forms of verbs which were used in a way very similar to the location nouns, like "nukkumaan" ('to sleep'), "syömään" ('to eat') or "katsomaan" ('to look'). The cluster even has a few locational adverbs, for example "takaisin" ('back'), "ulos" ('outdoors'), "sisälle" ('inside') and "pois" ('away').

The other cluster, not quite as coherent on the map as the other clusters but nevertheless discernible, is highlighted with turquoise blue, and the leftmost part of it overlaps a bit with the cluster of imperfect tense verbs. The cluster contains inflected forms of verbs that express mainly the actions of going, coming and being somewhere: for example "jäi" ('he/she stayed'), "menee" ('he/she goes'), "tulivat" ('they came'), "lähti" ('he/she left') and "olivat" ('they were'). These movement verbs are exactly the kinds of verbs that one would expect to find in the left side context of the location words of the cluster marked with blue. Also, vice versa, the right side context of these verbs probably should contain many of those location words.

Looking again at the component plane images of some features active in these areas of the map, this seems indeed to be the case. The left side context of location words seems to be dominated by different movement verb root features, for example the features "lähti/STM" ('he/she left') and "meni/STM" ('he/she went') displayed in figure 6.5, and also by conjugational verb endings like "vät/SUF" or "vat/SUF". On the other hand, the features active in the map area of movement verbs include agent nouns like "karhu/STM" ('bear') or pronouns like "ne/STM" ('those', or slang for 'they') in the left context, and,

[2] These phrases are authentic examples from the children's stories corpus.

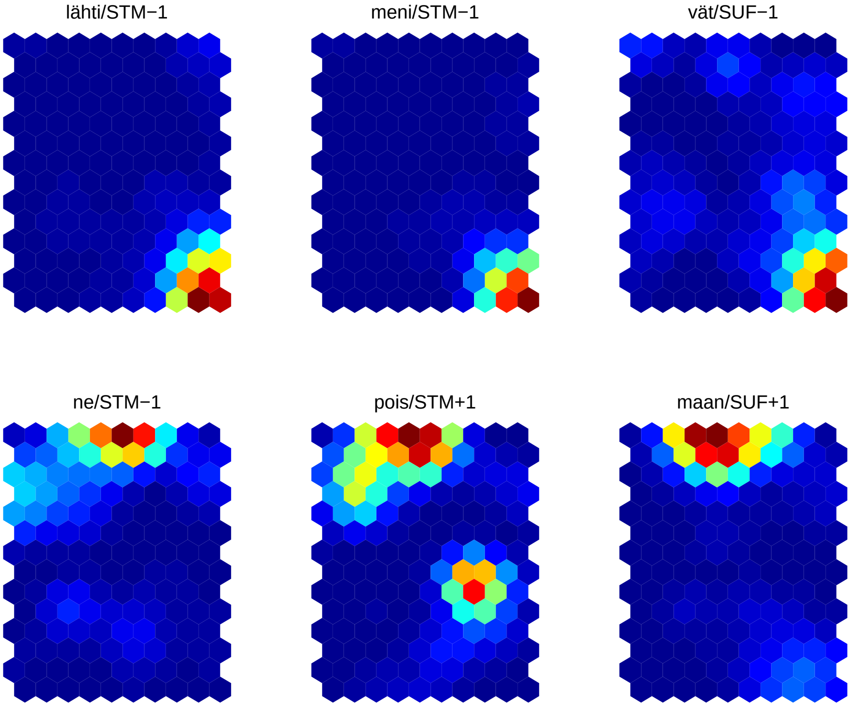

Figure 6.5: Component plane images of some feature morphs active in the location word cluster or in the movement verb cluster. Features found in the left context of the word forms are marked with "-1", and features in the right side context of the word forms have a "+1" attached to the feature name.

as expected, location words like the "pois/STM" ('away') or verb MA-infinitive endings like "maan/SUF" in the right side context (also displayed in figure 6.5).

6.2 Comparison: Different age categories

As explained in Section 4.2, the children's stories corpus was divided into three age categories: stories from the 1 to 4 year-old children, from the 5 to 6 year-olds and from the over 6 year-olds. On the data in each age category, a word SOM was trained for comparison between the different age groups and also with the word SOM trained on the whole corpus. As before, all three maps were constructed using the SOM Toolbox package (Vesanto et al., 1999) for Matlab.

For these age category maps, a slightly smaller set of 170 features was adopted, due to the smaller amount of data in each age category (see Section 4.2). The feature set for each of the three word SOMs thus included the 150 most frequent root morphs and the 20 most frequent suffixes in the stories of the particular age category. Consequently, each word form in the training

Age category	Feature morph set	Size of training word set
1 to 4 year-olds	150 root morphs + 20 suffixes	158 word forms
5 to 6 year-olds	150 root morphs + 20 suffixes	200 word forms
Over 6 year-olds	150 root morphs + 20 suffixes	184 word forms

Table 6.1: The types and sizes of the feature morph and training word sets for word SOMs of each age category. The cutoff value for the acceptance of a word form into the training word set was fixed at a minimum frequency of 30 occurrences.

word set was represented as a feature vector with 340 features (the 170 feature morphs in both the left and the right side context of the training words).

As for the set of training words, the number of training samples depended on the age category. Since the word form frequency counts of each of the three age categories varied, it was decided to impose a frequency limit of a minimum of 30 occurrences as a cutoff value for the acceptance of a word form into the set of training words. Thus, heeding this criterion, the training word set for the story data in the category of from 1 to 4 year-olds included 158 word forms, the category of from 5 to 6 year-olds had 200 training words[3], and the category of over 6 year-olds had 184. The sizes of the training word sets and the types and sizes of the sets of feature morphs for each age category are summarized in table 6.1.

The U-matrix representations for each of the three age category word SOMs can be found in figures 6.6, 6.7 and 6.8 respectively. Again, the word clusters discovered and presented in the previous Section have been manually highlighted in the age category word maps, using the same colours as before. As for an overall view on the three age category word SOMs, none of them seems to have a cluster structure quite as clear and distinctive as the word SOM on the whole story corpus. This is probably due to the fact that with the corpus divided into age categories, the sets of training data in each category have become much smaller, and there is perhaps not quite enough data for this kind of an analysis. Nevertheless, the same clusters of words that were observed in the word SOM on the whole corpus seemed to emerge from the age category maps as well, even if they were not as coherent and clear as before.

Looking at the agent noun clusters (highlighted with **red**) that emerged in each of the three age category word SOMs, their contents seem to be roughly the same. As in the first word map, the most important family members are present: "äiti" ('mother'), "isä" ('father'), "vauva" ('baby'), "tyttö" ('girl') and "poika" ('boy'). Also, there are several typical fairytale characters, like "prin-

[3] In this age category, there were more than 200 word forms with the minimum frequency of 30 occurrences, so only the first 200 were included.

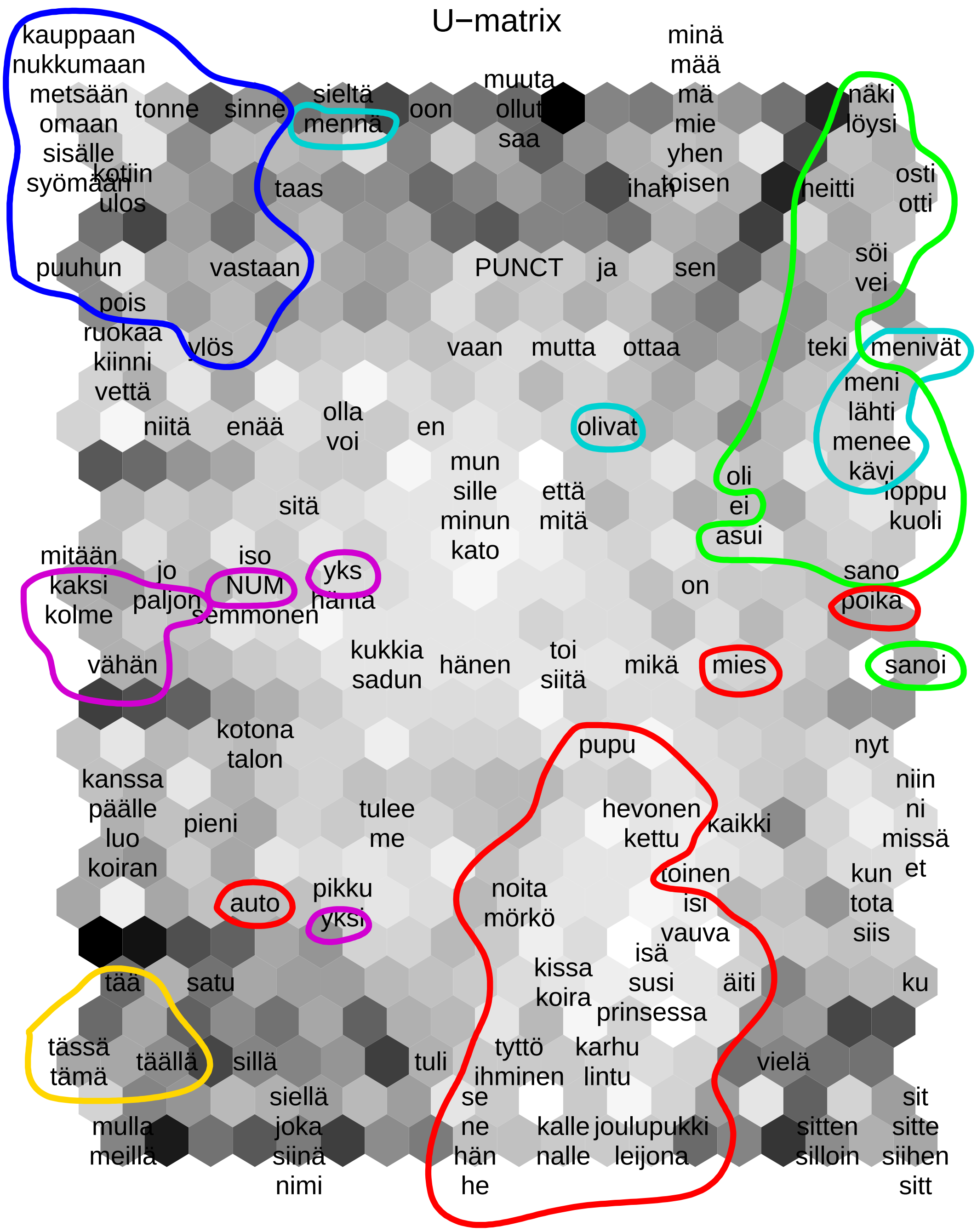

Figure 6.6: The U-matrix representation of the age category word SOM on the stories of children aged from 1 to 4. Some clusters of words on the map have been manually highlighted. See text for more discussion on the clusters.

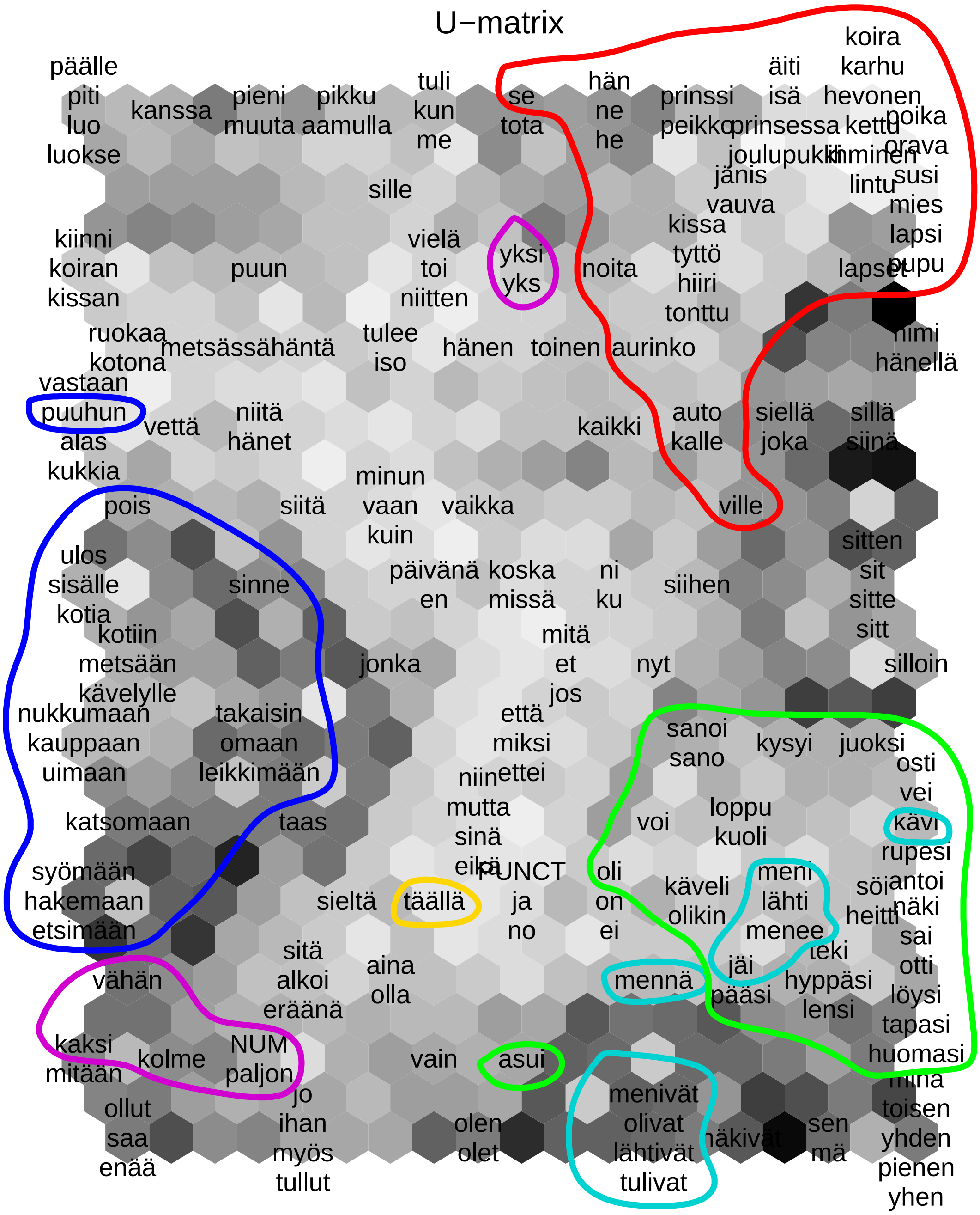

Figure 6.7: The U-matrix representation of the age category word SOM on the stories of children aged from 5 to 6. Some clusters of words on the map have been manually highlighted. See text for more discussion on the clusters.

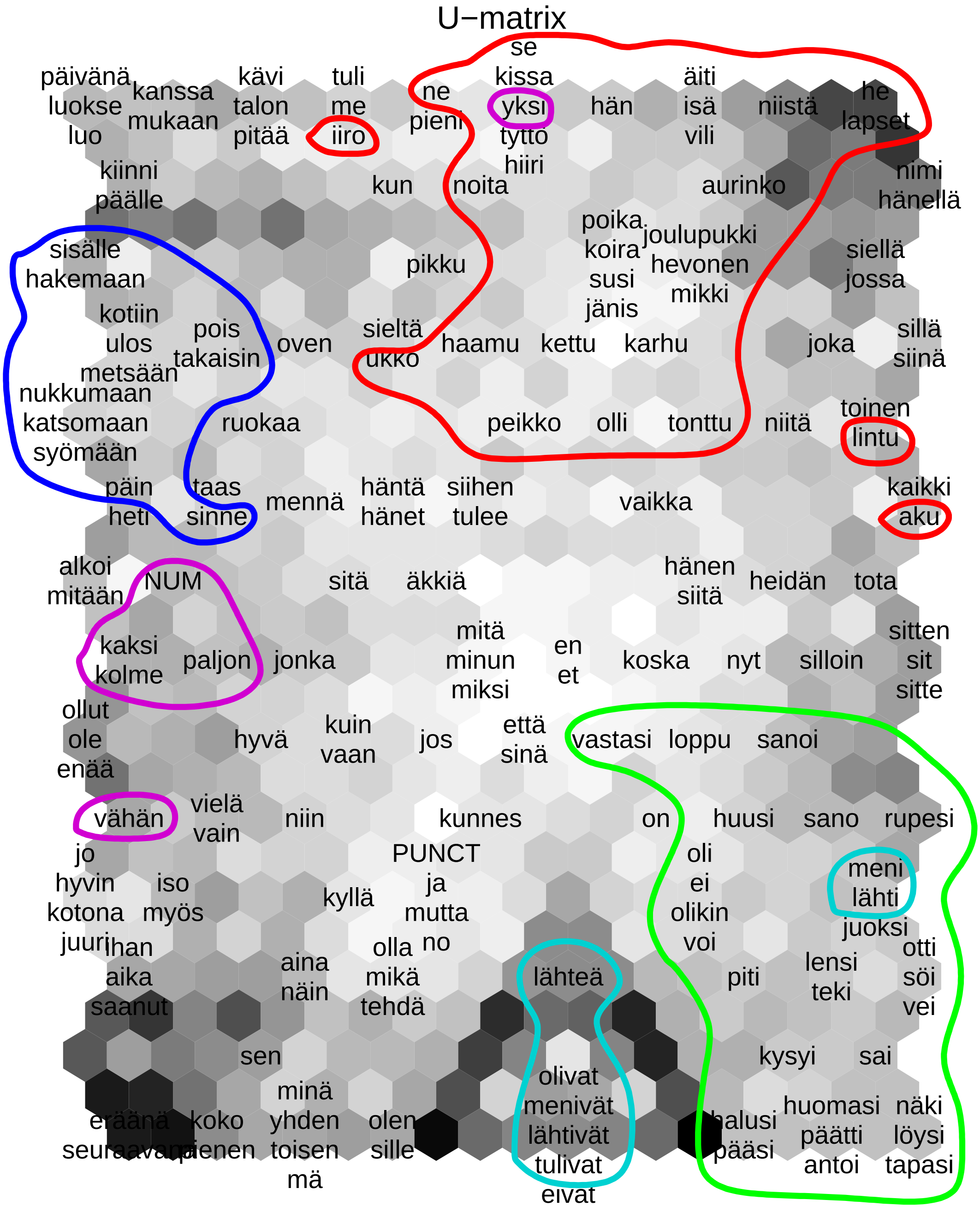

Figure 6.8: The U-matrix representation of the age category word SOM on the stories of children aged over 6. Some clusters of words on the map have been manually highlighted. See text for more discussion on the clusters.

sessa” (’princess’), “prinssi” (’prince’), “noita” (’witch’) and “peikko” (’troll’), as well as a whole host of animals, common in fairytales that many Finnish children are told in their early years: “karhu” (’bear’), “susi” (’wolf’), “kettu” (’fox’), “jänis” (’rabbit’) and so on. In the word SOMs of the two elder groups of children, the sets of training data also seemed to include some names: Ville, Kalle, Vili, Iiro, Aku (Finnish for ’Donald [Duck]’), Mikki (Finnish for ’Mickey [Mouse]’) and Olli[4]. Taking a closer look at the frequency lists of the three age categories, it seems indeed that the youngest children tended mostly not to give names to the characters of their stories, whereas the two groups of older children did increasingly name their characters. Looking at the word SOMs on the stories of the older children, the word map on the stories of the children aged from 5 to 6 years has two names, and the word map of the over 6 year-olds has a total of five.

An examination of the 3rd person singular imperfect tense verb clusters (marked with **green**) also reveals interesting differences between the age groups. The word SOMs on the stories of the two older groups of children both contain roughly 30 verbs in this cluster, but the youngest children get along with just 17 verbs. It should be kept in mind, of course, that the set of training data in the category of the youngest children contained only 158 word forms, which is less than for the other two categories. However, it seems that even if more word forms had been included, the youngest children would still have had considerably less verbs in this cluster than the older children. Thus, it can be concluded that the variety of verbs, or at least of verbs in 3rd person singular imperfect form, that children use in their stories seems to correlate with the age of the children.

Looking at the remaining highlighted clusters of words, they seem to be much the same in each of the three word SOMs. All maps have a location word cluster (marked with **blue**) with more or less the same types of location words, and a quantity word cluster (coloured **magenta**) with the word forms “yksi” and “yks” separated from the main cluster. As for the movement verb cluster (**turquoise blue**), its location on the map seems to vary depending on the age group: in the word SOM of the youngest children, the movement verbs are almost completely integrated inside the main verb cluster of imperfect tense verbs, but with the older children, the movement verbs that are not in 3rd person singular imperfect form seem to become more and more separated from the main imperfect tense verb cluster. In the word map of the over 6 year-olds, only the two movement verbs that actually share the number and

[4] The rather high frequency of the name “Olli” in the corpus is probably explained by the fact that a portion of the stories were collected from children at the Museum of Contemporary Art Kiasma, where they had just visited an exhibition by Olli Lyytikäinen.

tense of the verbs in this main imperfect tense verb cluster, namely “meni” (’he/she went’) and “lähti” (’he/she left’), are located inside the cluster, and the other five forms of movement verbs are clearly separated from the first two. This seems to indicate that as the children grow older, they start to pay an increasing amount of attention to the correct agreement between Finnish subjects and verbs. With young children, expressions with a singular verb following a plural subject such as “ne meni kauppaan” (’they went to the store’) are frequent, but the older children seem to have a tendency of using more often the orthographically correct plural form, i.e. “ne menivät kauppaan” or even the fully orthographical norm -conforming “he menivät kauppaan”.

Finally, there is yet another interesting observation concerning the two latter word SOMs trained on the stories of the older children. The cluster of different forms of the word “tämä” (’this’), highlighted with **yellow**, is very distinctive in the word SOM trained on the whole story corpus and also in the word map on the stories of the youngest children, but the word SOMs of the two older age categories seem to lack this cluster entirely. The middle SOM on the stories by children aged from 5 to 6 has only the word form “täällä” (’here’) from this cluster, and the last one does not seem to have any of them. The fact that these deictic, situation-dependent words did not make it into the sets of training words in the two categories of older children suggests that the older children seem to rely less on such deictic expressions. It may be that they have adopted a more abstract approach to story-telling, which resembles perhaps more the fairytale books that they have been read to by adults than the very situational, concrete identification with the story characters that was observed in the stories told by the youngest children.

6.3 Conclusions

In this Chapter, the children’s stories corpus was analyzed by using word SOMs with morph features. First, a word SOM on the whole story corpus was constructed, and some of the word categories that emerged in this SOM were analyzed in more detail by looking at the component plane images of some feature morphs from the feature set. These emergent word categories included for example clusters of agent nouns, 3rd person singular imperfect tense verbs, quantity words, location words and movement verbs, and also a cluster of different forms of the deictic word “tämä” (’this’). The study of the component plane images revealed that these clusters usually had particular types of feature morphs active in their parts of the word map, and that morphs seem indeed to be very useful especially in this kind of an analysis of a text corpus which contains plenty of non-orthographical word forms.

Next, utilizing the division of the story data into three age categories, as explained in Section 4.2, three age category word SOMs were trained on the data in each separate subcategory. These three word SOMs were compared to each other and to the word SOM on the whole story corpus, and several interesting differences were observed. The observations made in this Section could be of use for the research on the emergence of human linguistic competence in small children, as many of them suggest that the way children tell stories and the expressions they use in their stories seem to change and evolve as the children grow older.

Chapter 7

Discussion

In this Chapter, the work performed in this Thesis is summarized, and a number of ideas for further development in the research area are suggested.

7.1 Conclusions

In this Thesis, a Finnish text corpus of children's stories, collected using a method called Storycrafting, was analyzed with self-organizing word maps. The main innovation of this work is the construction of word SOMs which utilize unsupervised morphological information as their features. The feature morphs used in this work were automatically extracted from the children's stories corpus with an unsupervised morphology induction method called Morfessor, making this the first completely unsupervised morphological information -based SOM categorization of Finnish words.

The resulting word SOMs with different combinations of morph types as features were evaluated on the children's stories data against each other and against two traditional word SOMs with whole context words as features. The evaluation measure developed for this task utilizes the part-of-speech information of 200 manually classified word forms from the children's stories corpus as a basis, calculating a kind of a density score for the word clusters of a particular word SOM. The evaluation results obtained by this measure showed that using unsupervised morphological information as features of a word SOM clearly improves the quality of the SOM, at least when quality is measured as the part-of-speech -based density of the emergent clusters. Also, of all the feature set variants with different morph type combinations (roots, prefixes and suffixes), word SOMs with only root morphs as features seemed to yield the best results.

Finally, the children's stories corpus, consisting of 2642 stories in Finnish told by children aged from 1 to 14, was analyzed from a couple of different points of view. For the word SOMs in these analyses, both root morphs and suffixes were chosen into the feature sets. The inclusion of also suffix morphs into the feature sets was considered to yield more interesting analyses than with just root morphs as features.

First, a word SOM with 200 root morphs and 20 suffixes as features was trained on the whole story corpus. Some of the word clusters that emerged from this analysis were examined in more detail, using the component plane images of typical feature morphs that were active in those areas of the word map. Then, based on the age category division of the story data, three word SOMs with 150 root morphs and 20 suffixes as features were trained on the stories in each separate age category. These word SOMs were compared both to each other and to the word SOM on the whole story corpus, and interesting differences between the maps of the three age categories emerged. For example, it was observed that the use of certain deictic expressions in the stories seems to decrease as the children grow older, and that the older children seem to pay an increasing amount of attention to the correct agreement between Finnish subjects and verbs. These kinds of observations on the stories of the different age categories relate to the research on the emergence of human language abilities in children.

In summary, this Thesis shows that it is possible to obtain good emergent categorizations of Finnish words using just unsupervised methods. This was achieved using self-organizing maps with feature representations obtained with the Morfessor morphology induction method. In addition, analyzing the children's stories corpus with the proposed method yielded interesting results on the use of language of small children.

7.2 Future work

In the course of this work, several ideas emerged for improving the self-organizing map -based analysis and the methods used in this Thesis. These observations may serve as a good basis for future research in this area.

First, from the point of view of the unique children's stories corpus, it would be interesting to implement the kind of analysis described in Chapter 6 for yet new subcategories of the data. For example, besides the existing age category division, the stories could be divided into categories according to gender (stories by boys and stories by girls), or into stories told by an individual child versus stories by groups of children. The analysis of the gender-divided data could help understand the differences or similarities between the stories, worlds and ways of thinking of young Finnish boys and girls, which could be of interest and benefit for the study area of child research. On the other hand, a separate analysis on the individual and group stories could reveal some interesting facts about group dynamics among small children, and also about the special characteristics of group stories when contrasted to stories told by individual children of roughly the same age.

Further, with regards to the evaluation measure developed in this work for the automatic evaluation of word SOMs, it can hardly be considered perfect. Currently, the measure just looks whether the words in the same or the neighboring map nodes have any common parts-of-speech with the word form under examination. The measure could be improved by for example making it reward cases where there are more than one common parts-of-speech, since this probably means that the word forms are more similar with each other than those which only share one common part-of-speech. Additionally, the evaluation measure could also reward cases where there were lots of word forms in the same or in the neighboring nodes (a big cluster of many word forms), and a large proportion of these word forms had at least one part-of-speech in common with the word form at hand. This is based on the observation that the formation of bigger clusters with several word forms having common parts-of-speech probably implies a map of better quality than one which has lots of small clusters with only a couple of word forms inside them, whether these mini-clusters share common parts-of-speech or not.

As for the construction of the new kinds of word SOMs with morph features, there are several improvement ideas that could be implemented and evaluated. First, the morph-featured word SOMs presented in this Thesis should perhaps be trained and evaluated on some larger sets of data. As fascinating as the children's stories corpus is, its size is not, at least for the time being, very large (only a total of 198 036 word forms in the stories in Finnish). The new word SOMs should therefore be tested also on some other corpora with millions of word forms, to see whether the evaluation results on the different feature set variants obtained in this Thesis still hold even for larger amounts of data, and for data of different types. With larger corpora, a bigger amount of words could also be analyzed; in this Thesis, only the 200 (or less) most frequent word forms were chosen for training samples and projection onto the resulting word SOMs, since the use of more infrequent word forms would probably have lead into the sparsity of data and the amount of noise in feature vectores becoming an issue.

Also, the size of the context window used for calculating the feature vectors could be extended. Instead of looking at just the immediately preceding and following context words, the context window could encompass an area of two, three or even more words into both directions, or into just one or the other direction (for example, a context window of three words from the left context but only two from the right side context). Evaluation tests could be run on morph-featured word SOMs with different sizes and types of context windows, in order to find the optimal context windows in general or for the task currently at hand.

Further, it could be interesting to experiment with the possibility of including morphological information of the training sample word itself in the feature vectors. Instead of searching only the words in the context for feature morphs, in this version the morphs present in the morphologically segmented training word itself would also count in its feature representation. It is difficult to predict the effect, if any, this kind of an approach would have on the evaluation results. It might be that searching only for a certain type of morphs in the segmented training word would prove to be useful. For example, it might be best to consider only the suffix morphs of the training word, given the fact that they usually have a much higher frequency than e.g. root morphs. This kind of an approach might improve the capability of a word SOM in discovering semantically similar word collocations like “the cat/STM purred” and “the cat/STM+s/SUF purred”, or it might even result in detecting a novel family of morph collocations occurring in consecutive words, like “talo/STM+n/SUF luo/STM+na/SUF” (’by the house’), “kaveri/STM+n/SUF luo/STM+kse/SUF” (’to a friend’s place’) and “auto/STM+n/SUF luo/STM” (’to the car’). Here, the genitive case suffix morph “n/SUF” in the preceding context word of different inflected forms of the postposition “luo” (’by’) constitutes a kind of a morph collocation with the root morph “luo/STM”, and the similarity of these cases would be recognized even if none of the word forms ever match completely (only some of their morphs do).

Yet another way of improving the distinguishing potential of the feature set is using sets of morphs as features instead of individual morphs. A large preliminary set of morphs could be first organized into a smaller number of morph subgroups, and each subgroup of morphs would then be used as an individual component in the feature vector. The presence of feature morphs in context words would thus be checked against a list of morphs in a certain subgroup, not against individual morphs each occupying their own component slots in the feature vector. This kind of grouping of similar morphs into just one feature would help relieve the problem of data sparsity and reduce the dimensionality of the feature set. A “compressed” feature set of this type would also enable the utilization as features of a much larger amount of individual morphs than before, as items in the subgroups of morphs.

The morph subgroups needed for the creation of this kind of feature sets could also be constructed in an unsupervised manner. Using the textual contexts of the morphs, a large amount of Morfessor-extracted morphs could be first organized automatically with a self-organizing map, a “morph SOM” with emergent clusters of morphs instead of words. These morph clusters could then be regarded as the subgroups of the morph set, ready to be used as components of a feature vector for a word SOM with morph features. It might be a good idea to train a separate morph SOM for each morph type: one for root morphs, one for suffixes and one for prefixes[1]. In this way, we would have subgroups of root morphs (for example, the roots of semantically similar movement verbs could again end up as one cluster) and subgroups of suffixes, all to be used together as the components of a word SOM feature vector. Of course, it should be noted that the selection of optimal feature sets and other parameters for these kinds of novel morph SOMs is a whole different story, deserving a thorough treatise of its own.

From the point of view of the Morfessor method, the SOM-based organization of suffixes into subgroups seems especially interesting. For the time being, Morfessor does not recognize allomorphic variation, meaning that it does not for example understand that two such verb endings like "-vat" and "-vät" (products of Finnish vowel harmony rules) could be just the realizations or allomorphs of a common morpheme "-vAt", dependent on the vowels of the word they occur in. Could morph SOMs have the power needed to link together Morfessor-extracted morphs that are in a complementary distribution, i.e. that are allomorphic variants of the same underlying morpheme? In fact, looking briefly at some of the component plane images of suffixes used in the data analysis of the whole children's stories corpus in Chapter 6, the component plane images of morph pairs that are probably involved in allomorphic variation do indeed display promising similarities. Even if these are just morph features of a word SOM and do not really have anything to do with morph SOMs in proper, these similarities do give some indications that the morphs might indeed be organized rather nicely into clusters on a morph SOM. The use of morph SOMs in improving the performance of Morfessor could thus be an innovation well worth further examination.

[1] Although prefixes are so rare, at least in Finnish, that they could probably just be completely excluded from the feature set

Bibliography

Lou Burnard. The Users Reference Guide for the British National Corpus, 1995. http://www.natcorp.ox.ac.uk/.

Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30, Philadelphia, Pennsylvania, July 2002.

Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2004.

Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 106–113, Espoo, June 2005a.

Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Computer and Information Science Report A81, Helsinki University of Technology, 2005b.

Mathias Creutz and Krister Lindén. Morpheme segmentation Gold Standards for Finnish and English. Computer and Information Science Report A77, Helsinki University of Technology, 2004.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science*, 41:391–407, 1990.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2001.

Hervé Déjean. Morphemes as necessary concept for structures discovery from untagged corpora. In *Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adelaide, January 1998.

W. Nelson Francis and Henry Kucera. Brown corpus manual: Manual of information to accompany a standard corpus of present day edited American English. Technical report, Brown University, Providence, Rhode Island, 1964.

Stephen I. Gallant. A practical approach for representing context and for performing word sense disambiguation using neural networks. *ACM SIGIR Forum*, 3:293–309, 1991.

John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

Kadri Hacioglu, Bryan Pellom, Tolga Ciloglu, Ozlem Ozturk, Mikko Kurimo, and Mathias Creutz. On lexicon creation for Turkish LVCSR. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, 2003.

Margaret Hafer and Stephen Weiss. Word segmentation by letter succession varieties. *Information Storage and Retrieval*, 10:371–385, 1974.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Heinonen, and Irja Alho. *Iso suomen kielioppi.* SKS:n toimituksia 950. Suomalaisen Kirjallisuuden Seura, Helsinki, 2004.

Zellig Harris. From phoneme to morpheme. *Language*, 31(2):190–222, 1955.

Timo Honkela and Juha Winter. Simulating language learning in community of agents using self-organizing maps. Computer and Information Science Report A71, Helsinki University of Technology, Helsinki, Finland, December 2003.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, volume 2, pages 3–7, Paris, 1995. EC2 et Cie.

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. WEBSOM – Self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps*, Finland, June 1997. Helsinki University of Technology.

Timo Honkela, Aapo Hyvärinen, and Jaakko Väyrynen. Emergence of linguistic features: Independent component analysis of contexts. In *Proceedings of the 9th Neural Computation and Psychology Workshop (NCPW9)*, Plymouth, 2005. In press.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Fred Karlsson. *Yleinen kielitiede*. Helsinki University Press, Helsinki, Finland, 1998.

Liisa Karlsson. *Lapsille puheenvuoro. Ammattikäytännöt murroksessa. Giving children the floor. Transition in the tradition of professional practice.* PhD thesis, Helsingin yliopiston käyttäytymispsykologian tutkimusyksikkö, Helsinki, Finland, 2000.

Liisa Karlsson. *Saduttamalla lasten kulttuuriin. Ammattilaisverkostolla työn kehittämiseen yhdessä lasten kanssa.* 1999.

Simon Kirby. Syntax without Natural Selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, editor, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pages 303–323. Cambridge University Press, 2000.

Simon Kirby. Spontaneous evolution of linguistic structure: An Iterated Learning Model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, 5:173–203, 2001.

Teuvo Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer, Berlin, 3rd edition, 2001.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

Teuvo Kohonen. The 'neural' phonetic typewriter. *Computer*, 21(3):11–22, 1988.

Teuvo Kohonen, Kai Mäkisara, and Tapio Saramäki. Phonotopic maps – insightful representation of phonological features for speech recognition. In *Proceedings of the Seventh International Conference on Pattern Recognition*, pages 182–185, Los Alamitos, California, 1984. IEEE Computer Society Press.

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela. Self-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.

Kimmo Koskenniemi. *Two-level morphology: A general computational model for word-form recognition and production.* PhD thesis, University of Helsinki, 1983.

Krista Lagus, Anu Airola, and Mathias 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 2002.

Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the WEBSOM method. *Information Sciences*, 163(1-3):135–156, 2004.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

Tiina Lindh-Knuutila. Simulating the emergence of a shared conceptual system in a multi-agent environment. Master's thesis, Helsinki University of Technology, Espoo, Finland, October 2005.

Brian MacWhinney and Catherine Snow. The child language data exchange system. *Journal of child language*, 12:271–296, 1985.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts, 1st edition, 1999.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Peter H. Matthews. *Morphology.* Cambridge Textbooks in Linguistics. Cambridge University Press, 2nd edition, 1991.

Marshall R. Mayberry and Risto Miikkulainen. Lexical disambiguation based on distributed representations of context frequency. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, number AI94-217, January 1994.

Risto Miikkulainen. *DISCERN: A Distributed Artificial Neural Network Model of Script Procecssing and Memory.* PhD thesis, Technical Report UCLA-AI-90-05, University of California, Los Angeles, 1990.

Risto Miikkulainen. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory.* MIT Press, 1993.

Risto Miikkulainen. Self-organizing feature map model of the lexicon. *Brain and Language*, 59:334–366, 1997.

Ville Pulkki. Data averaging inside categories with the self-organizing map. Computer and Information Science Report A27, Helsinki University of Technology, 1995.

Monika Riihelä. *The Storycrafting Method.* Stakes, Helsinki, Finland, 2001.

Monika Riihelä. *Aikakortit - tie lasten ajatteluun. (Timecard - the way to children's thinking.).* VAPK-kustannus, Helsinki, Finland, 1991.

Helge Ritter and Teuvo Kohonen. Self-Organizing Maps. *Biological Cybernetics*, pages 241–254, 1989.

Edward Sapir. *Language: An Introduction to the Study of Speech.* Harcourt Brace, New York, 1921.

Johannes C. Scholtes. Resolving linguistic ambiguities with a neural data-orientede parsing (DOP) system. In I. Aleksander and J. Taylor, editors, *Artificial Neural Networks*, volume 2, Amsterdam, 1992.

Patrick Schone and Daniel Jurafsky. Knowledge-free induction of morphology using Latent Semantic Analysis. In *Proceedings of the Fourth Conference on Computational Natural Language Leargning and of the Second Learning Languages in Logic Workshop, Lisbon 2000*, pages 67–72, Somerset, New Jersey, 2000. Association for Computational Linguistics.

Patrick Schone and Daniel Jurafsky. Knowledge-free induction of inflectional morphologies. In *Proceedings of the 2nd Conference of the North American chapter of the Association for Computational Linguistics (NAACL-2001)*, June 2001.

Vesa Siivola, Teemu Hirsimäki, Mathias Creutz, and Mikko Kurimo. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pages 2293–2296, Geneva, Switzerland, September 2003.

Kenny Smith, Henry Brighton, and Simon Kirby. Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4):537–558, December 2003a.

Kenny Smith, Simon Kirby, and Henry Brighton. Iterated Learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003b.

Ville Turunen. Spoken document retrieval in Finnish based on morpheme-like subword units. Master's thesis, Helsinki University of Technology, Espoo, Finland, November 2005.

Alfred Ultsch. Self-organizing neural networks for visualization and classfication. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 307–313. Springer-Verlag, Berlin, Germany, 1993.

Juha Vesanto, Esa Alhoniemi, Johan Himberg, Kimmo Kiviluoto, and Jukka Parviainen. Self-Organizing Map for Data Mining in MATLAB: The SOM Toolbox. *Simulation News Europe*, (25):54, March 1999.

Sami Virpioja. New methods for statistical natural language modeling. Master's thesis, Helsinki University of Technology, Espoo, Finland, November 2005.

Richard Wicentowski. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. PhD thesis, Johns Hopkins University, 2002.

David Yarowsky and Richard Wicentowski. Minimally supervised morphological analysis by multimodal alignment. In K. Vijay-Shanker and C-N. Huang, editors, *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pages 207–216, Hong Kong, October 2000.