Creating Large Semantic Lexical Resources

for the Finnish Language

Laura Löfberg

Lancaster University

Department of Linguistics and English Language

March 29, 2017

Abstract

Finnish belongs into the Finno-Ugric language family, and it is spoken by the vast majority of the people living in Finland. The motivation for this thesis is to contribute to the development of a semantic tagger for Finnish. This tool is a parallel of the English Semantic Tagger which has been developed at the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University since the beginning of the 1990s and which has over the years proven to be a very powerful tool in automatic semantic analysis of English spoken and written data. The English Semantic Tagger has various successful applications in the fields of natural language processing and corpus linguistics, and new application areas emerge all the time. The semantic lexical resources that I have created in this thesis provide the knowledge base for the Finnish Semantic Tagger. My main contributions are the lexical resources themselves, along with a set of methods and guidelines for their creation and expansion as a general language resource and as tailored for domain-specific applications. Furthermore, I propose and carry out several methods for evaluating semantic lexical resources. In addition to the English Semantic Tagger, which was developed first, and the Finnish Semantic Tagger second, equivalent semantic taggers have now been developed for Czech, Chinese, Dutch, French, Italian, Malay, Portuguese, Russian, Spanish, Urdu, and Welsh. All these semantic taggers taken together form a program framework called the UCREL Semantic Analysis System (USAS) which enables the development of not only monolingual but also various types of multilingual applications.

Large-scale semantic lexical resources designed for Finnish using semantic fields as the organizing principle have not been attempted previously. Thus, the Finnish semantic lexicons created in this thesis are a unique and novel resource. The lexical coverage on the test corpora

containing general modern standard Finnish, which has been the focus of the lexicon development, ranges from 94.58% to 97.91%. However, the results are also very promising in the analysis of domain-specific text (95.36%), older Finnish text (92.11–93.05%), and Internet discussions (91.97–94.14%). The results of the evaluation of lexical coverage are comparable to the results obtained with the English equivalents and thus indicate that the Finnish semantic lexical resources indeed cover the majority of core Finnish vocabulary.

Declaration

I declare that this thesis is my own work. I also declare that it has not been submitted in substantially the same form for the award of a higher degree elsewhere. My contributions to the following academic papers has arisen directly as a result of my research for this PhD thesis.

Related Publications

Löfberg, L., Archer, D., Piao, S., Rayson, P., McEnery, A. M., Varantola, K., & Juntunen, J-P. (2003). Porting an English semantic tagger to the Finnish language. In D. Archer, P. Rayson, A. Wilson, & A. M. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (Vol. 16 , pp. 457–464). Lancaster: Centre for Computer Corpus Research on Language Technical Papers, University of Lancaster.

Löfberg, L., Juntunen, J-P., Nykänen, A., Varantola, K., Rayson, P., & Archer, D. (2004). Using a semantic tagger as dictionary search tool. In Williams, G., & Vessier, S. (Eds.), *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)* (pp. 127–134). Lorient: Université de Bretagne Sud.

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P., Nykänen, A., & Varantola, K. (2005). A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics 2005 Conference*. Proceedings from the Corpus Linguistics Conference Series on-line e-journal.

Mudraya, O., Piao, S., Löfberg, L., Rayson, P., & Archer, D. (2005). English-Russian-Finnish cross-language comparison of phrasal verb translation equivalents. In *Proceedings of the Phraseology 2005 Conference* (pp. 277–281). Louvain-la-Neuve.

Mudraya, O., Piao, S. S., Rayson, P., Sharoff, S., Babych, B., & Löfberg, L. (2008). Automatic extraction of translation equivalents of phrasal and light verbs in English and Russian. In Granger, S., & Meunier, F. (Eds.), *Phraseology: An interdisciplinary* perspective, (293–309). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Phoey, L.T., & Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In Calzolari et al. (Eds.), *10th edition of the Language Resources and Evaluation Conference (LREC2016)* (pp. 2614–2619). European Language Resources Association (ELRA).

Acknowledgements

On reaching the end of the tunnel ...

First and foremost I would like to thank Dr Paul Rayson. I first got to know him as a colleague in 2002, and he later became the supervisor of my thesis. My deepest gratitude goes to him for his guidance, expertise, inspiration, and support through all these years; my work would have been impossible without his input. I am also extremely grateful to my second supervisor, Dr Andrew Wilson, whom I got to know two years ago, for his feedback and insightful comments which helped to improve my work substantially. In addition, Chancellor Krista Varantola of the University of Tampere contributed to my work early on, for which I am exceedingly grateful. Furthermore, I am deeply indebted to my examiners, Dr Michael Oakes and Dr Claire Hardaker, for their excellent comments and feedback on my work and also for making the viva a very pleasant experience as well as an interesting discussion.

It was a wonderful opportunity for me to have been able to participate in the Benedict project and to work on such an interesting topic. I felt privileged to become a member of this multi-disciplinary group of skilled specialists and wish to thank all of you at Lancaster University, the University of Tampere, Kielikone, Gummerus, HarperCollins Publishers, and Nokia.

In addition, my warmest appreciation goes to:

- Dr Carl Wieck for assiduous language checking and good-humoured support,
- Dr Jaana Suviniitty and Sarri Öörni for much-needed peer support,
- Dr Kimmo Kettunen for valuable insights and co-operation,

- Dr Mahmoud El-Haj for creating the test environment for the semantic labeling experiment and for processing the results, and
- The Lancaster University Library and the Hyvinkää City Library (Sari Piironen in particular) for excellent service.

Last but not least, my truly heartfelt thanks go to my family, relatives, and friends for their unfailing support and patience throughout this process!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACAMRIT	Automatic Content Analysis of Interview Transcripts
ACASD	Automatic Content Analysis of Spoken Discourse
ASSIST	Automatic Semantic Assist for Translators
CASS	ESRC Centre for Corpus Approaches to Social Science
CLAWS	Constituent Likelihood Automatic Word-Tagging System
CSC	IT Center for Science
DIMAP	Dictionary Maintenance Program
EAGLES	Expert Advisory Group on Language Engineering Standards
EST	English Semantic Tagger
FinnONTO	National Semantic Web Ontology Project in Finland
FiWN	Finnish WordNet
FST	Finnish Semantic Tagger
HTOED	Historical Thesaurus of English
ICSI	International Computer Science Institute
ICT	information and communications technology
LLOCE	Longman Lexicon of Contemporary English
LWIC	Linguistic Inquiry and Word Count
MCCA	Minnesota Contextual Content Analysis
MWE	multiword expression
NLP	natural language processing
POS	part-of-speech

REVERE	Requirements Reverse Engineering to Support Business Process Change
RST	Russian Semantic Tagger
SAMUELS	Semantic Annotation and Mark-Up for Enhancing Lexical Searches
SeCo	Semantic Computing Research Group
synset	synonym set
UCREL	University Centre for Computer Corpus Research on Language
URI	uniform resource identifier
USAS	UCREL Semantic Analysis System
WSD	word sense disambiguation
XML	Extensible Markup Language
YSA	Yleinen suomalainen asiasanasto ("General Finnish Thesaurus")
YSO	Yleinen suomalainen ontologia ("Finnish General Upper Ontology")

1 Introduction

This thesis describes the theory, motivation, development, and evaluation of large semantic lexical resources for the Finnish language. These resources provide the knowledge base for a Finnish Semantic Tagger (FST), in other words, they are dictionaries which are used by a computer, not by a human being. I describe and evaluate these resources, outline their further development, and suggest new applications for them. The thesis places this work in the context of a new tool for the development of various types of natural language processing (NLP) and corpus linguistics applications involving the Finnish language.

1.1 Context

Semantic tagging can be briefly defined as a dictionary-based process of identifying and labelling the meaning of words in a given text. It has received increasing attention during recent years, and various programs which carry out this task automatically—that is, semantic taggers— have been developed for this purpose for different languages. Semantic tagging has been found very useful in many NLP applications, for example, in the fields of terminology extraction, machine translation, bilingual and multilingual extraction of multi-word expressions, monolingual and cross-lingual information extraction, as well as in automatic generation, interpretation, and classification of language. Semantic tagging has also been successfully utilized in the field of corpus linguistics, for example, for content analysis, analysis of online language, training chatbots, ontology learning, corpus stylistics, discourse analysis, phraseology, analysis of interview transcripts, and key domain analysis. New application areas emerge constantly.

I became acquainted with semantic tagging in my work as a researcher in a language technology project called Benedict—The New Intelligent Dictionary[1]. This project was funded by the European Community under the Information Society Technologies Programme, and it lasted from 1 March 2002 to 28 February 2005. The consortium was formed by Lancaster University, University of Tampere (a Finnish university), Kielikone (a Finnish language technology company), HarperCollins Publishers, Gummerus Publishers (a Finnish publishing house), and Nokia (a Finnish multinational communications and information technology company). Together we set out to:

> [...] combine[s] forces from language technology providers, the academia, the dictionary publishing world, and user organizations to discover the best way to cater for the needs of dictionary users by combining state-of-the-art language technology with research results on user needs and on the potential of future dictionaries (University Centre for Computer Corpus Research on Language, n.d.-a).

In practice, we catered for these user needs by generating various novel solutions. For example, the resulting intelligent dictionary software allowed the user to gain access to corpus information via dictionary entries with the aid of "Semantic Corpus Look-Up", and "Search Improvers" were helpful for finding the correct spelling for the search word when performing dictionary look-ups. Furthermore, a dictionary editing tool named DixEdit was created as well as a user log analyser and online dictionary feedback system for development, updating, and marketing purposes. The most innovative aim of our project, however, was to develop a context-sensitive dictionary search tool. Our new intelligent dictionary solution would not only help the user to find the correct main dictionary entry, as many of its electronic

[1] The project reference is IST-2001-34237. For more information, see ftp://ftp.cordis.europa.eu/pub/ist/docs/ic/benedict-ist-results_en.pdf.

counterparts did, but, as a novel feature, it would also point him/her[2] to the correct sense of the word in case the word has more senses than one, in other words, to the very sense he is looking up.

The core component of the context-sensitive dictionary search tool is based on semantic taggers for English and Finnish. These semantic taggers provide the necessary semantic information for the looked-up words, and the tool is applicable both to consulting a bilingual Finnish-English or English-Finnish dictionary and to consulting a monolingual dictionary of English or Finnish. During the Benedict project, we improved the existing English Semantic Tagger (EST) which had been developed at Lancaster University and which had already been found very useful in the fields of NLP and corpus linguistics. In addition, we developed an equivalent tool for Finnish: the FST. The context-sensitive dictionary search tool utilizes clues which are provided by the context of the looked-up word to be able to detect the relevant sense. By way of illustration, a person might be reading a website without understanding what the word "game" means in the sentence "This dressing is especially good with a salad of crisp vegetables and smoked poultry or game." To solve this problem, he could simply click on the word "game" in the website, and the Benedict dictionary solution would not only open the entry for this word in a separate pop-up window, but it would also highlight the correct sense in that very context, in this case, the sense "flesh of wild animals when used as food". The search mechanism is described in more detail in Löfberg et al. (2004).

The development of the FST and the semantic lexicons it is based on has been a challenging and long-running task. The work started in November 2002 when I visited the University Centre for Computer Corpus Research on Language (UCREL) team of Lancaster University. During this visit, I familiarized myself with the existing EST and the language resources incorporated within it, and together we drafted the initial guidelines for the

[2] Henceforth in this thesis, only the masculine pronoun will be used for the sake of simplicity.

development of the FST and of the Finnish semantic lexicons. We started playing philosophically with the idea of universalism: we wanted to experiment if a system originally created for the semantic tagging of English would work for the semantic tagging of Finnish as well. Our first experiences showed that it did work, even better than hoped. However, the task of such bridging of two languages which are very different from each other was neither fast nor easy. The software, which was created for the analysis of English, needed some modification to be able to process Finnish. In addition, the Finnish semantic lexical resources were created from scratch. They consist of a single word lexicon and a multiword expression lexicon (see chapter 3).

The more I learned about semantic tagging, the more it fascinated me. Eventually, I decided to continue improving the semantic lexical resources and studying the topic after the project ended. Thus, this thesis contains both the work I did in the Benedict project and the work I have been doing after its termination up to the present day. My main focus is on the semantic lexical resources which constitute the knowledge base for the FST and are my most important contribution to it.

1.2 Problem and Significance

So far relatively little work and research has been reported on the development of large machine-readable semantic lexical resources for Finnish (see section 2.5.2); most of the work in the field has been done for English. Furthermore, large semantic lexical resources based on a semantic field classification have not been attempted before for Finnish, but the existing semantic lexicons are based on different approaches. This thesis addresses this gap in the research by presenting Finnish semantic lexicons which use semantic fields as the organizing principle and are thus a unique resource created for Finnish.

In addition to describing and evaluating the Finnish semantic lexical resources, I will also outline their further development and suggest new applications for them. Even though they were developed in the Benedict project originally for the context-sensitive dictionary search mechanism, they can also be very practically applied in many other Finnish NLP and corpus linguistics applications and tailored for various purposes, as will become evident in this thesis.

The FST was the first non-English semantic tagger in the UCREL Semantic Analysis System (USAS; see section 2.4) framework. At present, there are equivalent semantic taggers based on equivalent semantic lexicons available for twelve languages, and the framework is continuously being expanded to cover new languages. The findings of this thesis, in regard to both the lexicon development and the software development of the FST, will benefit this work, especially when the USAS framework is extended to languages which, like Finnish, are highly inflectional. Moreover, now that there are equivalent semantic taggers available for many languages, this opens up exciting possibilities for the development of various multilingual applications in addition to monolingual Finnish applications.

1.3 Objective and Research Questions

The overall objective of this thesis is to contribute to the development of Finnish semantic lexical resources by investigating whether and how it is possible to create semantic lexical resources for Finnish which are compatible with the existing English semantic lexical resources while addressing the differences between these two languages. To fulfill this overall objective, I will address the following research questions:

- **RQ1: What do the Finnish semantic lexical resources consist of, what type of principles and practices have been followed in their creation, and how do these resources differ from their English counterparts both in terms of content and construction?**
- **RQ2: How extensive is the Finnish single word lexicon in terms of lexical coverage?**
- **RQ3: How suitable is the Finnish single word lexicon for use in the semantic analysis of Finnish in the FST software?**
- **RQ4: What resources and methods can be useful for the further development of the Finnish semantic lexical resources, firstly, as a general language resource, and, secondly, when they are applied to new domains?**

1.4 Organization of the Thesis

This thesis is organized in the following way. Chapter one is a general introduction. Chapter two establishes the background. To start with, I outline the general framework for the field of semantic tagging by introducing the most important related concepts. Thereafter, I provide an overview of some semantic ontologies and related systems which have been developed for various purposes. I then proceed to describing the UCREL Semantic Analysis System (USAS), another semantic ontology, which is based on the idea of semantic fields. The first semantic tagger developed in the USAS framework, the EST, and the semantic lexical resources which it relies on have functioned as a model for the development of the Finnish counterparts. I conclude the chapter by giving a brief account of the Finnish language. I particularly concentrate on those typical features of Finnish which have had an effect on the development of the FST software and its semantic lexicons in order to provide sufficient

background for non-Finnish speakers to understand the discussion about the grammar and structure of Finnish in the subsequent chapters. In addition, I briefly summarize some related lexical resources created for Finnish.

With these preliminaries dealt with, chapter three describes the Finnish semantic lexical resources. I begin by looking at the initial phases of the research and development process, and, subsequently, I provide a brief summary of the development and the structure of the software component in order to place the work in its immediate context. Although the FST software is not the main focus of this thesis, it is essential to start from it, since it is not possible to develop semantic lexical resources such as ours in isolation, but the software in which they will be applied needs to be taken into account in many respects throughout the development process. Thereafter, I provide a detailed description of the principles and practices which I have followed when creating the Finnish semantic lexical resources as well as of their contents. I also look at the similarities and differences between the Finnish and English semantic lexical resources and illustrate the output provided by the FST. This chapter answers **RQ1 (What do the Finnish semantic lexical resources consist of, what type of principles and practices have been followed in their creation, and how do these resources differ from their English counterparts both in terms of content and construction?)**.

Chapter four reports the results of four evaluations of the Finnish semantic lexical resources which were presented in chapter three. I start by briefly summarizing the results of the "formative evaluation". This evaluation was carried out at the end of the Benedict project during which the prototype of the FST was developed. Subsequently, I present the results obtained in the later experiments which I carried out after extending and improving the semantic lexical resources. The first set of these experiments, the "final evaluation", measures the lexical coverage by indicating the number of words which are covered by the single word

lexicon. This answers **RQ2 (How extensive is the Finnish single word lexicon in terms of lexical coverage?)**. The second set of experiments, the "application-based evaluation", measures the accuracy by indicating how well the single word lexicon performs when it is applied in the FST software. This answers **RQ3 (How suitable is the Finnish single word lexicon for use in the semantic analysis of Finnish in the FST software?)**. I analyze the errors which occurred in the application-based evaluation, and based on this analysis, I suggest ideas for improving both the semantic lexicons and the FST software. Finally, the fourth evaluation, the "semantic labeling experiment", measures how general native users of Finnish are able to replicate the USAS categorisation used in the Finnish semantic lexical resources.

Chapter five contains the discussion. Based on the findings from the evaluations presented in chapter four, I first draft guidelines for the continued development of the Finnish semantic lexicons as a general language resource similar to the English counterpart. Subsequently, I investigate the possibility of tailoring the Finnish semantic lexical resources for domain-specific applications. This chapter answers **RQ4 (What resources and methods can be useful for the further development of the Finnish semantic lexical resources, firstly, as a general language resource, and, secondly, when they are applied to new domains?)**.

Chapter six provides the conclusions. The first section comprises a summary of the thesis. Thereafter, I revisit the research questions and consider the limitations of the work as well as the novel contributions which the thesis makes to the field. I conclude the chapter by suggesting further work on the semantic lexical resources and also envisage new applications for them as well for the FST.

2 Background and Related Work

2.1 Introduction

The second chapter establishes the background for this thesis in order to place the work on the Finnish semantic lexical resources in context. I will begin by defining the most important related concepts which are: corpus linguistics, corpus annotation, linguistic annotation, part-of-speech (POS) tagging, parsing, and semantic tagging. Semantic tagging belongs to the field of corpus linguistics and represents one of the different types of corpus annotation. Linguistic annotation, which is one of the methods used in corpus annotation, includes, for example, POS tagging and parsing in addition to semantic tagging. POS tagging and parsing lay the basis for semantic tagging. Thereafter, I will review some semantic ontologies which have been developed for various purposes and which are relevant for the topic of this thesis. I will also briefly introduce some other, less related systems relying on semantic ontologies. Subsequently, I will describe the UCREL Semantic Analysis System (USAS), another semantic ontology, which has been created at the UCREL research centre at Lancaster University. The first semantic tagger developed in the USAS framework, the EST together with its semantic lexical resources, was used as the model to create an equivalent semantic tagger for Finnish, the FST, as well as equivalent Finnish semantic lexical resources. Finally, I will provide a brief overview of the Finnish language and of some of its specific grammatical features which have had an effect on the development of the FST, both in terms of the semantic lexical resources and the software, as well as a summary of related lexical resources created for Finnish.

2.2 Central Concepts

This section introduces the most important concepts related to semantic tagging[3]. I will start from the most general concept, which is corpus linguistics, and will then move on to successively specialized ones.

2.2.1 Corpus linguistics

McEnery and Wilson (2001, pp. 1–2) define corpus linguistics as the study of language that is based on examples of "real life" language use. They note that corpus linguistics as a discipline is not a branch of linguistics in the same sense as, for example, syntax, semantics, and sociolinguistics which concentrate on describing or explaining some aspect of language use. Rather, it is a methodology that can be applied to various aspects of linguistic study. McEnery and Hardie (2011, p. 1) describe corpus linguistics as an area which focuses upon a set of procedures for studying language, and that given these procedures, it is possible to take a corpus-based approach to many areas of linguistics. According to them, corpus linguistics has the potential to reorient our entire approach to the study of language by refining and redefining a range of theories of language. Furthermore, they postulate that corpus linguistics may enable us to use theories of language which were, at best, difficult to explore before the development of corpora of suitable size and of computers of sufficient power to exploit them and that corpus linguistics has also facilitated the exploration of new theories of language which are based on attested language use and on the findings drawn from this. Knowles (1996, p. 49) points out that "[t]he use of corpora brings back into linguistics the text about

[3] Semantic tagging can also be related to lexicography, but the approach in this thesis is connected with corpus linguistics. The SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches) project (University of Glasgow, n.d.-b) is an example of lexicographical applications of semantic tagging.

which dictionaries and grammars make generalizations"; the texts studied are not products of linguists' imagination, but they are records of actual events which can be printed out, picked up, and examined.

Corpora vary a great deal both in terms of size and content. For example, the size of a corpus can be anything from one single text document or book to the entire World Wide Web. In addition, a corpus can consist simply of written text, or it can also contain different kinds of multimedia sources, such as spoken discourses, video clips, pictures, and sound. The size and content of corpora depend on the aims and tasks for which they are collected.

Corpus linguistics has a long history, with the earliest corpus-based language studies being traced back to the late 19th century. Only after the early 1980s, however, did corpus linguistics become a prevalent and generally applied methodology in language studies. While in the beginning corpus linguistics was a marginalized approach used mainly by linguists studying the English language, it has subsequently been applied worldwide and multilingually. (McEnery & Wilson, 2001, pp. 1, 3, 24–25)

Information technology has changed the nature of corpus linguistics in a revolutionary manner. In the early days of corpus linguistics, before computerization, the collection and study of corpora was a manual task. This was naturally very time-consuming, expensive, and error-prone. Now the term "corpus" is almost synonymous with the term "machine-readable corpus". Computers have enabled corpus linguists to carry out the processes of searching, retrieving, sorting, and calculating linguistic data rapidly, accurately, and also at low cost. (McEnery & Wilson, 2001, p. 17) Quoting Sinclair:

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular. (1991, p. 1)

The development of the first machine-readable corpora with computerized search tools began in the 1940s (McEnery & Wilson, 2001, p. 20), and Leech (1997, p. 1) describes the creation of the one-million-word Brown Corpus, which was started in 1961, as the first true milestone in the field. With the currently available tools, corpus collection has thus become faster and easier, and this has resulted in a change from small corpora to increasingly large ones.

Machine-readable corpora are very flexible in the sense that it is easy to supplement them with additional information. One way of supplementing information is simply by growing the size of a corpus through the addition of new text and/or other material. Another way of adding information to a corpus is through corpus annotation.

2.2.2 Corpus annotation

Corpus annotation is the process of building interpretative information into corpora. Unannotated corpora consist of raw or plain texts that can be used as a basis for linguistic study, but they become far more useful if they are further refined and developed into annotated corpora. In this case, they are enriched with different kinds of linguistic information referred to as "annotations" to enable the manipulation of the data contained in the corpus in more diverse ways (McEnery & Wilson, 2001, p. 32). The term annotation refers both to the task of adding annotations to the text and to the actual linguistic symbols which are added (Leech, 1997, p. 2). Corpus annotation has been utilized extensively in corpus-based language study and NLP over the last several decades, and various annotation schemes and tools have been developed. The main focus has generally been on the English language, but since the turn of the millennium similar tools for other languages have become increasingly common.

Corpus annotation offers many advantages. According to Leech (1997, pp. 4–6), the first advantage is that it is easier to extract information from a corpus which is enriched with

annotations. Secondly, an annotated corpus can constitute a valuable resource that can be reused by other members of the research community. Thirdly, annotations are multi-functional; there are different levels of annotation, and one level prepares the way for the following level. For example, POS tagging can be seen as the first step towards more challenging levels of annotation, such as syntactic and semantic annotation; these will be discussed in more detail in the following sections.

Leech (1997, p. 6) remarks that during the history of corpus annotation some of the various annotation types that have been employed have been found to be difficult or even impossible to use by other members of the research community. To overcome this problem, Leech (1997, p. 6–8) drafts some practical guidelines for successful annotation of corpora:

1) The raw corpus should be recoverable, in other words, it should be easy to delete the annotations, if necessary.
2) Correspondingly, it should be easy to remove the annotations from the corpus and store them independently, if necessary.
3) An annotated corpus should come with appropriate documentation including information about the annotation scheme itself and of how, where, and by whom the annotations have been applied. Furthermore, there should be some account of the quality of annotations.
4) No annotation scheme should claim to represent "God’s Truth". The people who use readily annotated corpora use them simply for practical reasons. They consider it a much wiser choice than to start compiling their own corpora from scratch and inventing and using their own annotations.

5) The annotation schemes used should be based as far as possible on consensual or theory-neutral analyses of the data to avoid misunderstandings and misapplications.
6) No annotation scheme should claim to represent the absolute standard. The nature of the corpus as well as the particular needs of the task at hand have a decisive effect on what kind of annotation scheme is considered to be the most useful and sensible.

Leech (1997, pp. 7–8) raises two good points supporting the idea of a certain degree of unification in corpus annotation practices. The first advantage to be gained is in saving time and effort. It is clearly sensible to adhere to an annotation scheme that one is already familiar with and that has been found to be effective and useful. The second advantage is related to the reusability factor indicated above. If researchers wished to interchange data and resources, this would obviously be easier if the corpora were made compatible by following the same standards and guidelines worldwide. In fact, there was an attempt to standardize corpus annotation practices in 1990s, when a large community of language engineers set out to propose standards, guidelines, and recommendations for good practice in the core areas of the field. These were named the "EAGLES (Expert Advisory Group on Language Engineering Standards) Guidelines", and they included computational lexicons, text corpora, computational linguistic formalisms, spoken language resources, as well as assessment and evaluation (Institute for Computational Linguistics "A. Zampolli", n.d.).

Though corpus annotation clearly offers advantages, not everyone has fully supported its use. Sinclair (2004, pp. 190–191), for instance, admits that corpus annotation can be a helpful procedure, but he strongly cautions against its overuse. In his opinion, it allows the handling of documents without engaging in the interpretation of the language they contain. As long as a

text is marked up with annotations, the computer works with the annotations and ignores the language resulting in a study of the annotations, as opposed to a study of the language used. Sinclair also points out that if corpus data is observed through annotations, anything the annotations are not sensitive to will be missed. Hunston (2002, p. 93) has made similar observations. She suggests that while annotations add to the usefulness of corpora, they also make them less readily updated, expanded, or discarded. Furthermore, since the categories used for the annotation are typically determined before any actual annotation work has been carried out, this, in her opinion, limits the type of research questions that can be made.

There are several different types of corpus annotation. While this thesis deals with linguistic annotation, which will be discussed in the following subsections, other types include textual and extra-textual annotation, orthographical annotation, prosodic annotation, and phonetic transcription.

2.2.3 Linguistic annotation

Linguistic annotation refers to the process of enriching corpora through various types of linguistic information. The type of linguistic annotation in which special codes are attached to words in order to indicate particular features is often referred to as "tagging" rather than "annotation", and the codes which are assigned in this process are called "tags" (McEnery & Wilson, 2001, p. 46). The types of linguistic annotation that are relevant to this thesis are POS tagging, parsing, and semantic tagging. These will be discussed in the following subsections.

2.2.3.1 Part-of-speech tagging

The most basic type of linguistic annotation is POS tagging which is also known as grammatical tagging or morphosyntactic annotation. The annotation program automatically assigns each lexical unit in a text a tag that indicates its part of speech. The information about the part of speech is valuable, for instance, for corpus queries. Furthermore, POS tagging forms an essential foundation for further, more challenging levels of annotation, such as parsing and semantic tagging. (McEnery & Wilson, 2001, p. 46)

Nonetheless, POS tagging is not as uncomplicated as it may at first seem. In the previous paragraph was the sentence "[t]he annotation program automatically assigns each lexical unit in a text a tag that indicates its part of speech". Defining a lexical unit, however, is not always straightforward. In the simplest case, it is one single orthographic word preceded and followed by a space, for example, *kirjasto* ("library") or *lippalakki* ("cap"). However, a lexical unit can also be a unit of thought consisting of two or more separate orthographic words with an intervening space. These are referred to as "multiword expressions" (MWEs). In the Finnish language, MWEs can be considered to include, for example, proper names (e.g. *Englannin kanaali* ("English Channel"), *Buenos Aires*, *Euroopan Unioni* ("European Union"), *Hennes & Mauritz*), noun phrases (e.g. *biologinen kello* ("biological clock"), *musta pörssi* ("black market")), verb phrases (e.g. *avata tuli* ("to begin shooting"), *kirjoittaa ylös* ("write down")), idioms (*kuin seipään niellyt* ("as stiff as a ramrod"), *päätä pahkaa* ("headlong")). By comparison, MWEs are not as common in Finnish as they are in English, since, in addition to multiword proper names (e.g. "United Kingdom"), idioms (e.g. "out of this world"), and verb phrases (e.g. "die out"), English also contains an abundance of noun phrases the equivalent of which in Finnish would be written as single orthographic compound words (e.g. *yleisopinnot*

"general studies")[4]**.** The prevalence of MWEs in running text was calculated at 16% for English, in other words, 16 out of every 100 words in text participate in MWEs (Rayson 2005, p. 4). Unfortunately, corresponding information is not available for Finnish. It would be necessary that POS taggers as well as other types of linguistic annotation programs could recognize MWEs in text and tag them as one entity.

There are various POS taggers which are developed for processing different languages and which employ different types of tagging schemes. One such tool is Morfo, a POS tagger of Finnish (Jäppinen & Ylilammi, 1986). Morfo extracts all morpho-syntactic information from single words and MWEs and returns the candidate base forms with syntactic categories. By way of illustration, Morfo produced the following output for the sentence *Eilen oli varsin aurinkoista ja kesäistä, ja moni istahti puiston penkille nauttimaan lämpimästä säästä*.[5]:

EILEN ADVERB

OLLA VERB

VARSIN ADVERB

AURINKOINEN ADJECTIVE

JA CONJUNCTION

KES─INEN ADJECTIVE

---

[4] In section 2.5, I will present the Finnish language briefly and summarize some specific grammatical features, such as compounding, which have had an effect on the development of the FST software and its semantic lexicons.

[5]Translation: "It was very warm and summery yesterday, and many people sat on a park bench to enjoy the warm weather." The Morfo output was provided by Jukka-Pekka Juntunen from Kielikone. The character *Ä* is replaced by a dash in the output.

, ABBREV

JA CONJUNCTION

MONI PRONOUN

ISTAHTAA VERB

PUISTO NOUN

PENKKI NOUN

NAUTTIA VERB

L─MMIN ADJECTIVE

S── NOUN

S──ST── VERB

Note that Morfo gave two different interpretations of the last word, *säästä*. This word could be either the elative singular of the noun *sää* ("weather") or the second person singular of the imperative form of the verb *säästää* ("to save"). Morfo also generates inflectional information, which can be utilized for parsing presented in the following subsection, but this information does not show in the output.

Another example of a POS tagger is the Constituent Likelihood Automatic Word-Tagging System (CLAWS) (Garside & Smith, 1997, pp. 102–121) which the EST uses as a preprocessing component for semantic tagging. CLAWS has been continuously developed at

Lancaster University since the early 1980s. A more recent production at Lancaster University is a POS tagger created for the analysis of morphosyntactic categories in Urdu (Hardie, 2004).

2.2.3.2 Parsing

Another commonly used type of linguistic annotation is syntactic annotation, also referred to as "parsing". Parsing is often seen as the first stage of more comprehensive linguistic annotation, and programs that have been developed for this purpose are called "parsers". A parser assigns markers to each sentence in a corpus to indicate dependency relationships between words, for instance, predicates and objects. The DC Parser, which is a parser for Finnish developed by Kielikone Ltd, produced the following output for the example sentence in the previous subsection (*Eilen oli varsin aurinkoista ja kesäistä, ja moni istahti puiston penkille nauttimaan lämpimästä säästä*[6]):

```
==>Vaihe: DP<==
SLex=Eilen,SForm=Eilen,SCat=Adverb,SRel=Adverbial,SPosition=1,SInitCase=Upper
SLex=varsin,SForm=varsin,SCat=Adverb,SRel=IntensAttr,SPosition=3

SLex=aurinkoinen,SForm=aurinkoista,SCat=Adjective,SRel=ConjPreComp,SCase=Part,SNumber=SG,SPo
sition=4
SLex=ja,SForm=ja,SCat=Conjunction,SRel=CoordPreDep,SPosition=5
SLex=_COMMA,SForm=\,,SCat=Delimiter,SRel=Connector,SPosition=7,SAttached=AtLeft
SLex=kesäinen,SForm=kesäistä,SCat=Adjective,SRel=Subject,SCase=Part,SNumber=SG,SPosition=6,SAtt
ached=AtRight
```

[6] DC Parser output was provided by Jukka-Pekka Juntunen from Kielikone. Abbreviation SCase stands for Source Case, SCat for Source Category, SForm for Source Form, SLex for Source Lexeme, SRel for Source Relation, SNumber for Source Number, SPosition for Source Position, SSub Cat for Source Subcategory, STense for Source Tense, SVoice for Source Voice, etc.

SLex=moni,SForm=moni,SCat=Pronoun,SRel=Subject,SCase=Nom,SNumber=SG,SPosition=9,SSubCat=QuantPron

SLex=puisto,SForm=puiston,SCat=Noun,SRel=GenAttr,SCase=Gen,SNumber=SG,SPosition=11

SLex=penkki,SForm=penkille,SCat=Noun,SRel=Adverbial,SCase=All,SNumber=SG,SPosition=12

SLex=lämmin,SForm=lämpimästä,SCat=Adjective,SRel=AdjAttr,SCase=El,SNumber=SG,SPosition=14

SLex=sää,SForm=säästä,SCat=Noun,SRel=Adverbial,SCase=El,SNumber=SG,SPosition=15,SAttached=AtRight

SLex=nauttia,SForm=nauttimaan,SCat=Verb,SRel=Adverbial,SCase=Ill,SVoice=Act,SModal=IIIinf,SNumber=SG,SPosition=13

SLex=istahtaa,SForm=istahti,SCat=Verb,SRel=ConjPostComp,STense=Imp,SVoice=Act,SModal=Ind,SPersonN=S,SPersonP=3P,SPosition=10,SSubCat=Intr

SLex=ja,SForm=ja,SCat=Conjunction,SRel=CoordPostDep,SPosition=8

SLex=_PERIOD,SForm=.,SCat=Delimiter,SRel=Separator,SPosition=16,SAttached=AtLeft

SLex=olla,SForm=oli,SCat=Verb,SRel=Head,STense=Imp,SVoice=Act,SModal=Ind,SPersonN=S,SPersonP=3P,SPosition=2

In addition to the two grammatical annotation schemes described above, semantic tagging represents a further step of the levels of linguistic annotation that allows a more in-depth analysis of texts.

2.2.3.3 Semantic tagging

The primary focus of this thesis is on creating linguistic resources for semantic tagging (also referred to as "semantic annotation") which can be defined as the dictionary-based process of identifying and labelling the meaning of words in a given text. According to

Garside and Rayson (1997, p. 188), this process parallels that of grammatical tagging except that it is more abstract and more difficult to achieve. Semantic tagging has received increasing attention during recent years, and various automated tools which carry out this task—semantic taggers—have been developed for this purpose for different languages. Semantic tagging has been found to be very useful in diverse fields, such as terminology extraction, machine translation, bilingual and multilingual MWE extraction, monolingual and cross-lingual information extraction, as well as in automatic generation, interpretation, and classification of language. There is a variety of ways to carry out semantic tagging. While the approach dealt with in this thesis is based on the idea of semantic fields, there are also numerous other techniques which are called semantic annotation or semantic tagging, but they are based on different approaches. Examples of these different approaches are: semantic role labeling (e.g. Carreras & Màrquez, 2005; Gildea & Jurafsky, 2002), word sense disambiguation (e.g. Ide & Véronis, 1998; Stevenson & Wilks, 2003), named entity recognition (e.g. Tjong & De Meulder, 2003; Nadeau & Sekine, 2007), sentiment analysis (e.g. Pang & Lee, 2008; Wilson, Wiebe, & Hoffmann, 2005), and content analysis (e.g. Krippendorff, 2012; Shieh & Shannon, 2005).

Semantic tagging is certainly an effective method, but it also faces the difficulty that the same object or concept can be referred to in a number of ways; the identification of the meaning of a word is not necessarily an easy task, as Wilson and Thomas (1997, pp. 53–54) point out. By way of illustration, the animal *kissa* ("cat") can also be called *katti*, *mirri*, and *kisu*. This phenomenon is related to synonymy. On the other hand, one single word can refer to a number of concepts. For instance, the polysemous[7] noun *hiiri* ("mouse") can refer both to a rodent and to a pointing device for the computer. Equally, the homonymous[8] word *kuusi* can refer to the noun "spruce" as well as to the numeral "six", and it can even mean "your moon"

[7] A word is polysemous when it has two or more related meanings.

[8] Two or more words are homonymous if they have the same form but different unrelated meanings.

(although this expression is highly unlikely to appear very often in corpora). This kind of ambiguity can often cause confusion, because, even though in most cases human beings can differentiate between these different senses with the aid of their knowledge of the world, computer programs do not possess this knowledge and thus may be unable to choose the correct sense for a word in the context at hand. By way of illustration, if a person is using a search engine to find information about a certain word and enters into the search field a word that has multiple senses, he may end up with considerable amounts of unnecessary information in the search results, such as many hits on the number six when he actually wants to learn more about spruce trees. The task of automatically selecting the relevant sense for a word from a set of possibilities is referred to as "word sense disambiguation" (WSD) (Preiss & Stevenson, 2004, p. 201). Semantic tagging provides one method for carrying out this task. Likewise, if a person wished to search for the word *takki* ("coat") with a search engine, he would only achieve hits with the word *takki* in them, and the program would ignore websites containing, for example, the words *ulsteri* ("ulster"), *jakku* ("jacket"), *bleiseri* ("blazer"), and *anorakki* ("parka"). In such cases, semantic tagging can also be very useful by helping to find all the relevant information—and the relevant information only.

The type of semantic tagging which is discussed in this thesis is based on the idea of semantic fields. Wilson and Thomas (1997, p. 54) define a semantic field as "a theoretical construct which groups together words that are related by virtue of their being connected—at some level of generality—with the same mental concept". Words which belong to the same semantic field can be synonyms, antonyms, hyponyms, meronyms, or expressions that are associated with each other in one way or another. Synonymy (e.g. *lähellä* and *lähettyvillä* (both of these words denote "near")) and antonymy (*lähellä* ("near") and *kaukana* ("far")) are relations which exist between two words. The relations can also be hierarchical, as in case of hyponymy and meronymy, in which some words have a more general meaning whereas some

have a more specific meaning, when they are referring to the same entity. Hyponymy is the "kind of" relation. The most general term (e.g. *vaate* ("garment")) is on the top level of this hierarchy, and it is referred to as the "hypernym", and the more specific terms (e.g. *takki* ("coat")) on the level below are referred to as the "hyponyms". The second level terms, in turn, are hypernyms of even more specific terms (e.g. *anorakki* ("parka")) on the third level. By comparison, meronymy is the "part of" relation, where phenomena are analyzed into parts. Here the superordinate term (e.g. *paita* ("shirt")) refers to the complete entity, whereas the terms on the lower levels represent its parts (e.g. *hiha* ("sleeve") on the following level and then *kalvosin* ("cuff") on the subsequent level). Consequently, the words (*vaate* ("garment"), *takki* ("coat"), *anorakki* ("parka"), *paita* ("shirt), *hiha* ("sleeve"), and *kalvosin* ("cuff") as well as, for instance, the words *asu* ("attire"), *helma* ("hem"), *housut* ("trousers"), *riisuutua* ("undress"), *pukeissa* ("dressed"), *ilki alaston* ("stark naked"), and *haute couture* could all be considered to belong to the same semantic field. If we attach a semantic tag, a "label", to every word in a text indicating the semantic field into which each falls, we will then be able to extract all the related words from a text by querying on the specific semantic field. There is a problem, however, in the classification of words, since not all of them always fall conveniently into the predefined semantic fields, as Wilson and Thomas (1997, pp. 58–59) point out with the example word "sportswear". This word could be classified in the semantic field of clothing equally well as in the semantic field of sports. Such "fuzzy sets" will be discussed in more detail in section 2.4.1.1.

A collection of words classified into semantic fields can be designated as a "semantic annotation scheme" or a "semantic annotation system". According to Wilson and Thomas (1997, pp. 54–55), semantic annotation systems are something of a compromise between, on the one hand, attempting to mirror how words are believed to be organized into relationships in the human mind, and on the other hand, the need for usable annotated corpora and

reference works by linguists and other scholars. At present, we have only limited knowledge of the content and the form of the mental lexicon, but future discoveries may give us more insight into these issues. Wilson and Thomas (1997, p. 57) further observe that the majority of existing semantic annotation systems consist of very similar basic categories, but they differ from each other in terms of hierarchy (in other words, the structure of the categories) and in terms of granularity (in other words, the level of detail; how many categories the system distinguishes). Different types of semantic ontologies will be discussed in section 2.3.

Moreover, Wilson and Thomas (1997, p. 55) remark that there is no "ideal" semantic annotation system. Nevertheless, they suggest taking the following features into consideration when choosing which system to use or when developing a new system (Wilson & Thomas, 1997, pp. 55–57):

1) The system should be comprised of a linguistically or psycholinguistically consistent categorization.
2) The whole vocabulary in the corpus should be included in the system. A limited vocabulary is sufficient for some purposes in the field of content analysis but not for more general corpus annotation tasks.
3) The system should be adaptable to possible amendments which are necessary for treating a different period, language, register, or textbase.
4) Related to point 3, the system should operate at an appropriate level of granularity. This means that the annotation system should contain conceptually related words at varying levels of generality. There is no absolute in terms of granularity, but the correct level depends at least partly on the aims of the end user.
5) Related to point 4, a hierarchical structure would be an advantage for being able to adjust the granularity to the aims of the end user. If a system had a hierarchical

structure based on increasingly general levels of related words, it would be possible to identify all these different levels without having to try to decide which is the level the end user wishes to employ. Indeed, it would be easy for the end user to look at all the different levels by simply moving up or down to the next level of granularity in the hierarchy.

6) The system should conform to a standard if there is one. The existence of a standard would make it easier to accumulate and compare research results for, for instance, different languages, periods, and genres.

Rayson and Stevenson (2008, pp. 568–571) distinguish between four types of semantic field annotation. The first approaches were based on artificial intelligence. Thereafter, in the 1980s, knowledge-based approaches were developed utilizing the abundance of information which was contained in readily available machine-readable dictionaries. The third approach was corpus-based where machine-readable corpora offered large lexical resources that could be exploited for the purpose. The fourth approach were hybrid methods which are a combination of the previously mentioned methods. The type of semantic field annotation which is described in this thesis belongs to the fourth approach. Our approach is a hybrid one, because it combines knowledge-based and corpus-based approaches.

Texts can be annotated with semantic field information in three different ways depending on the level of automation (Wilson & Thomas, 1997, p. 62). The first option is to attach all annotations in the text manually. The second option, computer-assisted tagging, represents a semi-automatic form of manual tagging which is supported by a computer-readable lexicon containing possible semantic fields for given words. Such systems may also contain a limited amount of automatic WSD mechanisms. In this case, the computer is used to assign candidate semantic field tags to all the words in a text on which there is already information, and it

leaves for manual treatment only those words that it does not recognize or which remain ambiguous after the application of disambiguation methods. The third option is a fully automatic semantic tagger. This is a program which assigns the correct semantic fields automatically to all the known words in a text without any manual intervention and without leaving any words ambiguous. The semantic tagging approach dealt with in this thesis utilizes the third option.

A major advance in the development and evaluation of semantic tagging systems has been the introduction of the SensEval evaluation exercises which provide a uniform framework for comparing the performance of existing systems (Rayson & Stevenson, 2008, p. 575). SensEval is an international organization which has operated since 1997 and whose goal is to further our understanding of lexical semantics and polysemy. They organize and run evaluation and related activities to test the strengths and weaknesses of WSD systems with respect to different words, different aspects of language, and different languages (Rada Mihalcea, n.d.). SensEval later evolved into SemEval, and their ninth workshop on semantic evaluation was held in 2015.

2.2.3.4 Linguistic annotation summary

In the previous subsections, I have looked at different types of linguistic annotations on different levels. The most basic type, POS tagging, lays the basis for parsing which, in turn, prepares the way for semantic tagging that permits a yet deeper analysis of text. Thus, POS tagging and parsing are both necessary steps in successful implementation of semantic tagging.

Yet another, relatively recent step that represents movement toward deeper analysis of text is what is known as "pragmatic annotation". Whereas syntax involves a mono relationship (a

relationship between linguistic forms) and semantics involves a dyadic relationship (a relationship between linguistic forms and world entities), the relationship in pragmatics is triadic, involving not only linguistic forms and world entities but also the language user. Thus, pragmatics focuses on language together with its contexts, such as the speaker's intentions, the hearer's understanding, as well as the social and physical contexts. Since these contexts differ according to the task at hand, pragmatic annotation cannot be fully automated in the same way as grammatical annotation. Nevertheless, the computer can be of valuable assistance in the tagging process. (Archer, Culpeper, & Davies, 2008, pp. 615, 637)

Finally, it must be pointed out that no computer program written for any type of automatic linguistic analysis or manipulation of text is one hundred per cent reliable. The reliability and the accuracy depend to a large extent on the language resources which are included in the system.

2.3 Semantic Ontologies

Computerized tools for assisting text analysis have existed for decades. These tools are based on classifying words according to their meaning in semantic ontologies, and there are different ways of carrying out this task. Schmidt (1986, p. 780) has suggested one way of dividing the approaches used. His basic types are: 1) the conceptual analysis method, 2) the content analysis method, and 3) the collocation or co-occurrence method. The semantic lexical resources dealt with in this thesis utilize the conceptual analysis method. In this subsection, I will present some examples of the first two of these approaches, both of which are relevant for this thesis, and I also briefly discuss some other, less related lexical resources. The third approach, the collocation or co-occurrence method, which has been widely used by psycholinguists to reveal certain regularities of the occurrence of connotative and associative

content in a given text (Schmidt, 1986, p. 787), is beyond the scope of this thesis[9]. There are also other ways of dividing the approaches, for example, to dictionary-based approaches and to collocation-based approaches, but the division suggested by Schmidt was the most practical division in the context of this thesis.

2.3.1 Conceptual analysis method

The conceptual analysis method refers to complete conceptual systems represented in thesauri (Schmidt, 1986, p. 787). Thesauri are dictionaries, but they differ from "traditional", alphabetically organized dictionaries in the sense that the semantic macrostructure takes the place of the alphabet. According to Hüllen (2006, p. 13), the term "thesaurus" was made popular by *Roget's Thesaurus* (see section 2.3.1.1.1) and has over the years become a generic noun, while other terms used for this type of reference works are "thematic", "topical", "conceptual", "ideographical", and "onomasiological" dictionaries. Hüllen himself uses the term "topical dictionary", whereas, for example, McArthur (see section 2.3.1.1.2) uses the term "thematic dictionary".

Thesauri are based on some systematic arrangement of topics derived from some scientific system or semantic classification which is expected to be generally understood by a non-expert user (Hüllen, 2006, p. 14–15). Thesauri can contain, for example, synonyms, antonyms, hyponyms, hypernyms, definitions, paraphrases, quotations, and pictures. They are typically arranged in two parts: a systematic part and an alphabetical index. A good analogy was provided by Schmidt (1986, p. 788) who compared the two different parts to a telephone directory. The yellow pages represent the arrangement of the lexical material along the conceptual system, whereas the white pages represent the alphabetical arrangement of the

[9] However, there have been experiments using this approach on automated thesaurus extraction to assist conceptual analysis (e.g. Schütze & Pedersen, 1997; Curran & Moens, 2002).

material with complete references to the conceptual fields of the yellow pages. McArthur justifies the advantages of thesauri in the following, very apt way in his preface to his *Longman Lexicon of Contemporary English* (see section 2.3.1.1.2):

> The alphabet, with all its virtues, places animals and zoos, uncles and aunts far apart in its scheme of things, whereas in the human mind such words go closely together. The alphabetical dictionary has a logic, but it is not the logic of everyday life. In principle, one feels, words should be defined in the company they usually keep. (Mc Arthur, 1981, p. vi)

In fact, the tradition of organizing things thematically is much older than the tradition of organizing things alphabetically. The former was the dominant practice in information organization beginning in ancient times (for instance, scribes in Mesopotamia learned their cuneiform signs in thematic groups drawn from the everyday world), whereas the latter only became an established tool in the world of reference more than one hundred years after the advent of printing (McArthur, 1986, pp. 74–77). Hüllen (2009, p. 124) remarks that alphabetical writing systems are perhaps the only linguistic convention which is universally accepted and which has never been contested in its history.

In the following subsections, I will present some thesaurus-based systems which I have divided into two different groups for practical reasons, based on the level of hierarchy. The first group contains deep hierarchies which are built on three or more levels, and the second group contains shallow hierarchies with one or two levels.

2.3.1.1 Deep hierarchies

The following thesauri are built on three or more levels.

2.3.1.1.1 Roget's Thesaurus

I begin my review of semantic ontologies with *Roget's Thesaurus*. It is one of the most successful dictionaries of the English language ever created, a true milestone in the history of topical lexicography in particular, and an example followed by many other dictionary compilers for over 160 years now.

Since the beginning of the 19th century and all through his professional career, Peter Mark Roget, a physician and a scientist, collected words, phrases, and other forms of expression in various orders in a notebook (Davidson, 2002, p. viii–xiv). His original intention was to use his findings to aid him in expressing himself as a writer and a lecturer, but later he came to realize that the findings might be useful for other people as well. Hence, when he had retired from work, he spent the first four years further expanding the material which he had collected and organized it into a coherent system. The first edition of *Roget's Thesaurus* was published in 1852. Over the years, the thesaurus has been expanded and updated many times. Roget collected new words and expressions for the thesaurus until his death in 1869, after which his work was continued by, among others, his son, John Lewis Roget, and his grandson, Samuel Romilly Roget (Davidson, 2002, p. xv–xvi). The latest edition is named the "150th Anniversary Edition". It was edited by George Davidson and published in 2002.

When devising his system of classification of "the ideas which are expressible by language" (Roget, 1852/2002, p. xxii), Roget's aim was first and foremost practical. In his introduction to the first edition, he wrote:

> I have accordingly adopted such principles of arrangement as appeared to me to be the simplest and most natural, and which would not require, either for their comprehension or application, any disciplined acumen, or depth of metaphysical or antiquarian lore. (Roget, 1852/2002, p. xxii)

The top level, as presented in the "plan of classification" of the first edition, consists of the following six "classes" which are further subdivided into "sections" (Davidson, 2002, p. xxxiiii):

1. Abstract Relations
    - 1.1 Existence
    - 1.2 Relation
    - 1.3 Quantity
    - 1.4 Order
    - 1.5 Number
    - 1.6 Time
    - 1.7 Change
    - 1.8 Causation
2. Space
    - 2.1 Generally
    - 2.2 Dimensions
    - 2.3 Form
    - 2.4 Motion
3. Matter
    - 3.1 Generally
    - 3.2 Inorganic
    - 3.3 Organic
4. Intellect
    - 4.1 Formation of Ideas
    - 4.2 Communication of Ideas

5. Volition
    - 5.1 Individual
    - 5.2 Intersocial
6. Emotion, Religion, and Morality
    - 6.1 Generally
    - 6.2 Personal
    - 6.3 Sympathetic
    - 6.4 Moral
    - 6.5 Religious

According to the instructions in the 150th Anniversary Edition (Davidson, 2002, p. xxxviii), the logical progression from abstract concepts through the material universe to mankind itself culminates in morality and religion which Roget considered mankind's highest achievements.

The sections, in turn, are further subdivided into subcategories referred to as "heads" which are the basic units of *Roget's Thesaurus* and under which the words and phrases are arranged in paragraphs according to their parts of speech (Davidson, 2002, pp. xxxviii–xxxix). By way of illustration, the heads "Existence", "Nonexistence", "Substantiality", "Insubstantiality", "Intrinsicality", "Extrinsicality", "State", and "Circumstance" are included in the section "Existence" in the class "Abstract Relations". Hüllen (2004, p. 339) lists three functions for Roget's heads. Firstly, they serve as flags for each article and are thus a semantic companion to the numbers which accompany the heads. Secondly, they are a point of reference for the synonyms to follow. Thirdly, they are also a point of reference for the possible antonym as the headword of the corresponding article. In the first edition, there are 1,000 heads, whereas in the 150th Anniversary Edition, there are 990 heads. However, the

classes and the sections have remained exactly the same over the years; they all follow Roget's original plan of classification. (Davidson, 2002, pp. xxxviii–xxxix)

In his introduction to the first edition, Roget (1852/2002, pp. xxix–xxx) writes that a work constructed on his plan of classification could be "of great value, in tending to limit the fluctuations to which language has always been subject, by establishing an authoritative standard for its regulation", and he also suggested that the principles of its construction could be universally applicable to all languages. Furthermore, he envisaged bi- and even multilingual thesauri based on his plan of classification, and indeed, the classification used in *Roget's Thesaurus* has been transferred to other languages. Hüllen (2009, pp. 60–91) reports two adaptations: Théodore Robertson's *Le Dictionnaire Idéologique* (1859) in French and Daniel Sanders' *Sprachschatz* (1873) in German[10]. The basic structures of these two dictionaries and the English original are very similar. Sanders had increased the number of classes from six to seven, but the new class resulted simply from a division of class six, "Emotion, Religion and Morality", into two separate classes. In addition, he had reduced the number of heads from 1,000 to 688, since he had considered the original number of heads artificially ambitious. However, there is no conceptual modification behind these changes. Robertson, in turn, used identical classes and sections to Roget's. Furthermore, a parallel special field thesaurus was subsequently created; Day applied the same classification in his *Roget's Thesaurus of the Bible* (Day, 1992; Roget's Thesaurus of the Bible, n.d.).

[10] *Deutscher Wortschatz*, a German thesaurus reworked first by Hugo Wehrle and subsequently by Hans Eggers, was also quite similar, since it was designed to be aligned with Roget's categories. This thesaurus was originally compiled by Anton Schlessing under the title *Deutscher Wortschatz oder Der passende Ausdruck*, and it was published in 1881. However, during the publishing history Schlessing's name was omitted. (Wehrle & Eggers 1961, p. v; Zillig 2014, p. 1)

2.3.1.1.2 McArthur's Longman Lexicon of Contemporary English

The *Longman Lexicon of Contemporary English* (*LLOCE*) is a thesaurus created by Tom McArthur and published in 1981. It aimed at covering the core vocabulary of the English language arranged according to a hierarchical structure of related meanings (McArthur, 1981, p. vi). The written material is supplemented by pictures. According to McArthur (personal communication, June 8, 2007), his classification was not created in isolation, but it arose from the centuries-old tradition of dividing up the world into constituent elements, most famously represented by the structuring of *Roget's Thesaurus*, even though the categories the two contain are quite different from each other. Jackson and Zé Amvela (2000, pp. 112–113) consider the *LLOCE*, with its semantic field arrangement, a more interesting and more revealing account of English than the accounts presented in alphabetically organized dictionaries, even though the *LLOCE* is neither very extensive in scope nor up-to-date. This thesaurus is of particular interest to this thesis, since the initial tagset of the USAS framework (see section 2.4), to which the FST belongs, was based on the classification of the *LLOCE*.

The hierarchy contains 14 "semantic fields" which are identified by upper case letters running from "A" to "N" (McArthur, 1981, pp. vi–vii):

A. Life and Living Things
B. The Body: Its Functions and Welfare
C. People and the Family
D. Buildings, Houses, the Home, Clothes, Belongings, and Personal Care
E. Food, Drink, and Farming
F. Feelings, Emotions, Attitudes, and Sensations
G. Thought and Communication, Language, and Grammar

H. Substances, Materials, Objects, and Equipment

I. Arts and Crafts, Science and Technology, Industry, and Education

J. Numbers, Measurement, Money, and Commerce

K. Entertainment, Sports, and Games

L. Space and Time

M. Movement, Location, Travel, and Transport

N. General and Abstract Terms

These semantic fields of a pragmatic, everyday nature are further divided into 127 "set titles" of related words. In turn, the set titles further expand into 2,441 "sets" which are identified by reference letters and numbers. For example, the word "cottage" is identified by D4 and can thus be found in the set "Smaller Houses" together with the words "hut", "shack", "hovel", "shanty", "cabin", and "chalet".

2.3.1.1.3 Historical Thesaurus of English

The *Historical Thesaurus of English* (*HTOED*) is a unique resource which does not exist for any other language. It provides a detailed record of English vocabulary from the earliest times up to the present day. It includes current meanings of words, words that have become obsolete, and obsolete meanings of words which still exist, and it presents them arranged in semantic categories, together with information of the dates of currency for all meanings of each word. This vast undertaking was initiated in 1965 by Michael Samuels, a professor of English Language at the University of Glasgow, and it was finalized in 2009. (Kay, Roberts, Samuels, & Wotherspoon, 2009, pp. xiii–xiv; Kay & Alexander, 2010, pp. 107, 109) The main source for the *HTOED* was formed by data from the *Oxford English Dictionary*

(Simpson & Weiner, 1989) which contains full coverage from the year 1150 all the way up to the present day. The coverage for the Old English Period (700–1150) is more selective, and it has been supplemented with material from *A Thesaurus of Old English* (Roberts, Kay, & Grundy, 1995) which was a spin-off of the *HTOED* project. (Kay et al., 2009, p. xvi)

Not only is the amount of data vast, but the classification used in the *HTOED* is also very detailed, much more so than in the other conceptual analysis systems discussed here. At the beginning of the classification work, the categories of *Roget's Thesaurus* were used as a preliminary filing system, but many of them were later abandoned (Kay et al., 2009, p. xiv). The top level of the classification includes three categories: 01) The External World, 02) The Mental World, and 03) The Social World (Historical Thesaurus of English, n.d.). On the second level, they subdivide as follows:

01 The External World (The World)

- 01.01 The Earth
- 01.02 Life
- 01.03 Health and Disease
- 01.04 People
- 01.05 Animals
- 01.06 Plants
- 01.07 Food and Drink
- 01.08 Textiles and Clothing
- 01.09 Physical Sensation
- 01.10 Matter
- 01.11 Existence and Causation
- 01.12 Space

- 01.13 Time
- 01.14 Movement
- 01.15 Action
- 01.16 Relative Properties
- 01.17 The Supernatural

02 The Mental World (The Mind)

- 02.01 Mental Capacity
- 02.02 Attention and Judgement
- 02.03 Goodness and Badness
- 02.04 Emotion
- 02.05 Will
- 02.06 Possession
- 02.07 Language

03 The Social World (Society)

- 03.01 Society and the Community
- 03.02 Inhabiting and Dwelling
- 03.03 Armed Hostility
- 03.04 Authority
- 03.05 Law
- 03.06 Morality
- 03.07 Education
- 03.08 Faith
- 03.09 Communication
- 03.10 Travel and Travelling
- 03.11 Occupation and Work

03.12 Trade and Finance

03.13 Leisure

The following, third level includes 377 categories in all. The system is expanded even further, having provision for seven main category levels and five subcategories (Kay et al., 2009, p. xviii). The total size of the category set is presently 225,131 categories (University of Glasgow, n.d.-a).

By far the most commonly used organizing principle in the *HTOED* has been synonymy, whereas antonymy was considered less suitable for the purposes of this project, since oppositions vary both in content and nature. For instance, the categories "Love/Hate" and "Pain/Pleasure" would generally be placed together, since the opposition is obvious. It would not, however, be equally obvious with categories like "Truth", because there can be a progression of meaning which covers several oppositions, for instance, in the case of "Truth" moving from "Validity" through "Truth", "Sincerity", "Falsehood", and "Error" to "Deceit". Where appropriate, the *Oxford English Dictionary* style labels have been added to give further information, for example, to indicate slang, irony, or dialectal use (Kay et al., 2009, p. xix).

Recently, the *HTOED* has been applied in an extension of the USAS system to create a historical semantic tagger for the English language. It complements the semantic tags used in the FST (see section 2.4.1.1) by offering finer-grained meaning distinctions for use in WSD. (Alexander, Dallachy, Piao, Baron, & Rayson, 2015, p, i16)

2.3.1.1.4 Hallig and von Wartburg's Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas

The *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas* ("A Concept System as a Basis for Lexicography: An Attempt at an Organizational Model") is a thesaurus of French[11] compiled by Rudolf Hallig and Walther von Wartburg. In their thesaurus, they attempted to combine both a conceptual and an alphabetical approach by setting up a system of concepts which were supposed to be independent of language, despite the fact that people can think of concepts only with the help of words (Hüllen, 1990, p. 134). The first edition was first published in 1952. It enjoyed wide recognition, and it was praised as an important lexicographical achievement and as a masterplan for future lexicographical work. (Hüllen, 1990, pp. 129–132) The system was initially applied to the analysis of mid- to late 20th century vocabulary (Wilson, 2002, p. 417). Similarly to Roget, Hallig and von Wartburg envisaged that their dictionary could provide the foundation for thesauri of all languages, although they admitted that their system might be more easily adapted to Indo-European languages and might also have to be adapted according to the needs and cultural shape of some languages (Hüllen, 2006, p. 19; 1990, p. 136). The system was later elaborated by Klaus Schmidt in developing his series of conceptual glossaries for the medieval German epic and yet further by Andrew Wilson for building conceptual glossaries for the Latin Vulgate Bible. Schmidt modified the system to achieve a better treatment of the world of the medieval German epic. In contrast, Wilson modified the system primarily to better meet the needs of analyzing biblical text by amending the parts of Schmidt's conceptual system which were culture-specific to fit the context of the medieval epic. (Wilson, 2002, pp. 417–418)

[11] This is a thesaurus of the French language, but the metatext is mostly written in German.

The hierarchy in the Hallig and von Wartburg's system is built on three to six levels. Its top level contains three categories which, on the second level, expand into ten subcategories (Hallig & von Wartburg, 1963, p. 101–112[12]):

A. The Universe
   1. Sky and Atmosphere
   2. The Earth
   3. Plants
   4. Animals

B. Man
   1. Physical Being
   2. Mind and Soul
   3. Man as Social Being
   4. Social Structure

C. Man and the Environment
   1. A Priori
   2. Science, Learning, and Technology

Category A contains items which are related to nature but exclude the human being. Category B contains items which are related to the human being, both in terms of physiology, illness, life death, sex, nutrition, and clothing, as well as psychological and social processes. Category C includes not only items related to learning and technology but also a subcategory named "A Priori". This subcategory expands into a wide variety of lower level categories which are

[12] The English translations have been taken from University Centre for Computer Corpus Research on Language (n.d.-b) which displays all the categories of this system.

related to different fields, such as existence, conditions, order, numbers, quantity, time, causality, change, and motion. The total size of the category set is 402.

2.3.1.2 Shallow hierarchies

The following thesauri are built on one or two levels.

2.3.1.2.1 Laffal's Concept Dictionary of English

Julius Laffal first published his *Concept Dictionary of English* in 1973 for the purposes of automatic content analysis. He was a psychologist by background, and he had studied word association behaviour and methods for isolating the natural cognitive sets to which words seem to belong (Huntsman, 1975, p. 46). Laffal originally created his system for analyzing the content of psychiatric materials, but he also applied it, for instance, to the analysis of literature (e.g. Laffal, 1995) as well as to free speech and conversations (e.g. Laffal, 1967).

In the preface to his dictionary, Laffal (1973, p. x) mentions that the dictionary follows the tradition demonstrated by Roget and Dornseiff. However, Laffal's approach differs from the other thesauri discussed here in that his dictionary could be described as a thesaurus in reverse. While a thesaurus traditionally starts from categories of different types under which words are grouped, Laffal's concept dictionary lists the words first and then after each word includes one to five categories[13] which, in his view, are related to the word in question (Laffal, 1995, p. 339). After the alphabetical listing, however, the dictionary also contains a listing of all words by category.

[13] Laffal himself uses the term "concept" instead of "category".

Laffal's categories are not arranged in a hierarchy but on one level only. In the first edition in 1973, he used 114 categories (Laffal, 1970, p. 175), while in the 1990 edition the number had been increased to 168 (Laffal, 1995, p. 339). The categories are identified by two- to four-character mnemonic names and are presented in alphabetical order. The following list displays the first 20 categories, along with two prototypical example words of each category (Laffal, 1995, p. 350):

- AFAR Distant, Strange
- AGEN Repeat, Again
- AGGR Aggression, Anger
- AGRE Concur, Agree
- AID Help, Support
- ANAL Anality, Excrement
- ANGL Angle, Bend
- ANML Animal, Dog
- ARM Arm, Elbow
- ART Art, Sculpture
- ASTR Astronomy, Sky
- BACK Rear, Behind
- BAD Evil, Bad
- BGIN Start, Commence
- BIND Constrain, Tie
- BIRD Bird, Eagle
- BLOK Prevent, Stop
- BLUR Vague, Dubious

- BODY Body, Torso
- BONE Bone, Tooth

Note that the system is not entirely theory-neutral. For instance, the category "ANAL" is influenced by psychoanalysis, referring to the anal stage which is, according to Sigmund Freud, the second stage in the human psychosexual development.

2.3.1.2.2 Dornseiff's Wortschatz nach Sachgruppen

Der deutsche Wortschatz nach Sachgruppen, created by Franz Dornseiff and first published in 1933, is another noteworthy thesaurus of German. Dornseiff as well had aspirations related to universality, and in the preface to the first edition he wrote that the dictionary had linguistic goals which could also be useful for languages other than German (Dornseiff, 1970, p. 5). In fact, he had originally proposed a conceptually organized dictionary for Old Greek, but eventually this was realized for German (Hüllen, 1990, p. 156).

The hierarchy in Dornseiff's classification consists of two levels. The top level is formed by 20 categories which all expand further into 14–121 subcategories. The top level category with the largest number of subcategories is "Society and Community", whereas the top level category with the smallest number of subcategories is "Literature. Science". All in all, there are 910 subcategories. The following lists the categories on the top level:

1. Inorganic World, Matter
2. Plants, Animals, Man (Physically)
3. Space, Location, Form
4. Size, Amount, Number, Degree

5. Nature, Relationship, Event
6. Time
7. Visibility, Light, Colours, Sound, Temperature, Weight General Situation of the Whole, Odor, Taste
8. Change of Location
9. Wanting and Acting
10. Feelings of the Senses
11. Feeling. Emotional State. Personality Traits
12. Thinking
13. Signs, Communication, Languages
14. Literature, Science
15. Art
16. Social Relationships
17. Equipment, Technology
18. Economy
19. Justice, Ethics
20. Religion, The Supernatural.

Dornseiff's system was later adopted by Dietmar Najock for creating a Latin vocabulary of *The Eclogues* by Vergil, arranged according to conceptual categories (Najock, 2004). Najock had first considered the use of the system developed by Hallig and von Wartburg, but he had decided to use Dornseiff's system instead, because it offered a finer analysis (Kytzler, 2005).

2.3.1.2.3 Louw and Nida's Greek-English Lexicon of the New Testament Based on Semantic Domains

Johannes P. Louw and Eugene A. Nida published their *Greek-English Lexicon of the New Testament Based on Semantic Domains* in 1988. As is evident from the title, this thesaurus differs from the other thesauri dealt with here in that it is bilingual. Nida was a renowned linguist and semanticist by background and published widely on various topics, such as componential analysis (e.g. Nida, 1975), translation (e.g. Nida, 1969), and morphology (e.g. Nida, 1949), whereas Louw's main interest lay in the study of New Testament Greek (e.g. Louw, 1973, 1982). The authors targeted the work primarily at translators of the New Testament into various languages, but they anticipated that it might also be of interest to biblical scholars, pastors, and theological students, as well as to linguists and lexicographers (Louw & Nida, 1988, p. iv).

The *Greek-English Lexicon of the New Testament* contains the entire vocabulary of the third edition of *The Greek New Testament* (Aland, Black, Martini, Metzger, & Wikgren 1975). The lexicon is arranged in a hierarchy which consists of 93 top level categories. Lexical items related to objects and entities are grouped in categories 1–12, lexical items related to events are grouped in categories 13–57, and lexical items related to abstracts and relationals are grouped in categories 58–91. (Louw & Nida, 1988, p. vi) The following lists show the first eight categories of each group (for the complete hierarchy, see University Centre for Computer Corpus Research on Language, n.d.-c):

1. Geographical Objects and Features
2. Natural Substances
3. Plants

4. Animals
5. Foods and Condiments
6. Artefacts
7. Constructions
8. Body, Body Parts, and Body Products

13. Be, Become, Exist, Happen
14. Physical Events and States
15. Linear Movement
16. Non-Linear Movement
17. Stances and Events Related to Stances
18. Attachment
19. Physical Impact
20. Violence, Harm, Destroy, Kill

58. Nature, Class, Example
59. Quantity
60. Number
61. Sequence
62. Arrange, Organise
63. Whole, Unite, Part, Divide
64. Comparison
65. Value

Category number 92, "Discourse Referentials", consists of pronominal and deictic expressions. The final category "Names of Persons and Places", number 93, is somewhat different from categories in the other semantic ontologies mentioned here, in that it contains proper names. (Louw & Nida, 1988, p. vi) All in all, 22 top level categories contain only one level, whereas 70 top level categories expand into a second level. Only the top level category "Time" expands into three levels, so for this reason I have included this category system among shallow hierarchies rather than among deep hierarchies. The total size of the category set is 587.

2.3.1.3 Differences and similarities in conceptual analysis method ontologies

There are various differences and similarities between the ontologies which represent the conceptual analysis method. The editors of the *A Thesaurus of Old English*, in the introduction to the work (Roberts et al., 1995, p. xxv), suggest that "Schemes of classification have no inherent truth, but represent the best attempts of the compilers to present their materials within a coherent and illuminating framework." Fischer (2004, p. 49, 54–55) postulates that a truly universalist scheme cannot even exist, firstly, because there is no general consensus about what is a "natural" or "logical" order of things, and secondly, because any scheme will be coloured by the culture from which it originates and by the language in which it is written. Moreover, Fischer points out that classifications also differ for the reason that categorization is a multidimensional operation in that human beings will see most concepts as belonging to several categories. Kay et al. (2009, pp. xix) propose very similar thinking in their introduction to the *HTOED*. They remark that no semantic category is likely to be wholly clear-cut and cite the example of the categories of "Music" or "Religion". Their content is typically well-defined, but this brings into question as to how religious music

should be categorized. A corresponding issue arose in section 2.2.3.3: Wilson and Thomas (1997, pp. 58–59) point out that words do not always fall conveniently into predefined semantic fields and cite as an example the word "sportswear" which could be classified both in the semantic field of clothing and in the semantic field of sports. This view is also shared by Hüllen (2006, pp. 14–15; 2009, p. 60) and Schmidt (1986, p. 788). Schmidt compares the chase after the purely "objective" system to the chase after the same illusion as Kant's "pure object" and states:

> The true test for any pre-established conceptual system can only be to what extent it is acceptable to as many human minds as possible beyond the boundaries of individual languages and cultures. That means the higher the degree of abstraction the greater is the likelihood of universal acceptance. As long as we cannot reach general understanding at this higher level of abstraction we cannot possibly find it on the level of specific meaning. This does not mean that there should be only one system, it just means that the basic ingredients of each system should be the same, while there could be many different degrees of differentiation as well as differences in hierarchical order. (Schmidt 1986, p. 788–789)

In fact, Archer, Rayson, Piao, & McEnery (2004, p. 817) have observed that even though many semantic category systems are different in terms of their structure and granularity, they often agree to a greater or lesser extent on the basic major categories they contain. Schmidt (1986, p. 788) even suggests that a simple conversion program could rearrange any conceptual dictionary from one conceptual system to another.

The similarity is clearly evident in the conceptual systems presented above, despite the fact that they are arranged differently. They vary a great deal as to the depth of the hierarchy and the number of categories they include, but they comprise the same "basic ingredients". In

addition to the differences discussed earlier, there are some other disparities as well. McArthur's categorization appears the most practical, including several categories for concrete entities, whereas, for example, the majority of Roget's and Dornseiff's categories are more abstract by nature. Interestingly, McArthur's 20 top level categories also include "Entertainment, Sports, and Events". A category covering those topics does not exist on such a high level in any of the other deep hierarchies. In the *HTOED*, there is a second level category "Leisure" which subdivides on the third level into the categories "Entertainment", "Social", "The Arts", "Sport", and "Dancing". In Roget's plan of classification, the category "Leisure" is placed on the third level. In Hallig and von Wartburg's system, there is a fifth level category "Celebrations, Games, Amusements" which further expands into the subcategories of "Festivals/Festivities", "Games/Diversions", "Sport", and "Traditions/Customs".

Another interesting difference is the treatment of religious and supernatural issues. With regard to the deep hierarchies, Roget, McArthur, and Hallig and von Wartburg place them in the same top level category, whereas in the *HTOED*, they are placed in separate top level categories. Within the *HTOED*, there is a second level category called "Faith" under the top level category "The Social World", whereas the second level category "Supernatural" can be found under the top level category "The External World". With regard to the shallow hierarchies, Laffal's one level system contains the categories "HOLY" (referring to religious figures, activities, and objects) and "MYTH" (referring to the supernatural, the mythical and the magical), and, similarly, Louw and Nida's system contains the top level category "Supernatural Beings and Powers" as well as the top level category "Religious Activities". In addition, Louw and Nida's system has a second level category "Be a Believer, Christian Faith" under the top level category "Hold a View, Believe, Trust". In comparison, Dornseiff's system, which contains fewer top level categories than Laffal's and Louw and Nida's systems,

includes topics related to both religion and the supernatural in the same top level category. Most of the category systems discussed here represent a modern view of the world and may not be altogether applicable to historical texts[14], and the division between religion and the supernatural is a good example of such a case. In the modern world, they are considered two different issues, while, in earlier times, they were not necessarily separate from each other.

Looking back at the features 3–5 which Wilson and Thomas (1997, pp. 55–57; see section 2.2.3.3) suggest to be taken into consideration when choosing which system to use or when developing a new system:

3) The system should be adaptable to possible amendments which are necessary for treating a different period, language, register, or textbase.
4) Related to point 3, the system should operate at an appropriate level of granularity. This means that the annotation system should contain conceptually related words at varying levels of generality. There is no absolute in terms of granularity, but the correct level depends at least partly on the aims of the end user.
5) Related to point 4, a hierarchical structure would be an advantage for being able to adjust the granularity to the aims of the end user. If a system had a hierarchical structure based on increasingly general levels of related words, it would be possible to identify all these different levels without having to try to decide which is the level the end user wishes to employ. Indeed, it would be easy for the end user to look at all the different levels by simply moving up or down to the next level of granularity in the hierarchy.

On the basis of these guidelines, systems which are built as deep hierarchies would be a more practical choice than systems which are built as shallow hierarchies. Indeed, a deep hierarchy

[14] An exception to this is the *HTOED* which is intended to cover Early Modern English and Old English.

is more flexible, since the end user can determine the appropriate level of granularity depending on the task at hand. Furthermore, it would be easier to amend a deep hierarchy if a particular task requires.

Even though most of the work discussed here has concentrated on the English language, there are many conceptual systems for other languages as well, in addition to the systems mentioned above. One such is Paul Fortier's (1989) ontology for his computer-aided analysis system which he used for the analysis of French prose fiction. Moreover, Julio Casares (1942) compiled the *Diccionario ideológico* which is the only existing large thesaurus for Spanish and was later made available in electronic format (Valderrábanos, Dïaz, & Pérez, 1994). To date, no thesauri have been compiled for Finnish. There are two synonym dictionaries, the *Synonyymisanakirja* ("Synonym Dictionary") (Jäppinen, 1989) and the *Synonyymisanasto* ("Synonym Lexicon") (Leino & Leino, 1990), but these are simply synonym finders which do not utilize any categorization but only list total and partial synonyms in their entries.

2.3.2 Content analysis method

While exhaustive, semantically categorized thesauri provide the foundation for the conceptual analysis method discussed above, the content analysis method uses a selection of pre-established specialized dictionaries as a basis against which texts are compared, resulting in different types of statistical analyses. In this method, the emphasis is on the general content and on the distribution of lexical items within corpora to examine which of the materials are relevant from a given viewpoint or for a specific purpose. This approach has been used, for example, in psychology, in social and behavioural sciences, and in literary research. (Schmidt, 1986, pp. 780, 786) Thus, unlike conceptual analysis systems, content analysis systems are not aiming at full coverage, but they concentrate on those categories which are relevant for

the above fields. In the following three subsections, I will present three examples of such systems.

2.3.2.1 General Inquirer

The General Inquirer system has been developed at Harvard University since 1961 for the purposes of applying various computer-assisted content analysis procedures to the field of social science. The core of the system is formed by two thesauri which are merged together. The first of them is the Harvard III Psychosocial Dictionary. The developers call it a psychosocial dictionary, since it was aimed at investigators with psychological and sociological objectives and theories (Stone, Dunphy, Smith, & Ogilvie, 1966, p. 171). Nowadays, the updated version is available as the Harvard IV-4 dictionary. The second thesaurus is the Lasswell Value Dictionary. Together they contain a framework of 182 categories which are represented by tags (e.g. "Weak" representing the category related to weakness, "Legal" representing the category related to legal, judicial, and police matters, and "NonAdlt" representing the category related to infants and adolescents) (Harvard University, n.d.-a). The Harvard IV-4 dictionary includes categories belonging to the following major groups:

1. "Osgood"[15] three semantic dimensions (positive words, negative words, words implying strength, weakness, active orientation, and passive orientation),
2. words of pleasure, pain, virtue, and vice,
3. words indicating overstatement and understatement, often reflecting presence or lack of emotional expressiveness,

[15] These are categories which reflect psychologist Charles Osgood's semantic differential findings regarding basic language universals (Harvard University, n.d.-a; Brooke, 2001, p. 3).

4. words reflecting the language of a particular "institution",
5. words referring to roles, collectivities, rituals, and forms of interpersonal relations, often within one of these institutional contexts,
6. ascriptive social categories as well as general references to people and animals,
7. references to places, locations, and routes between them,
8. references to objects,
9. processes of communicating,
10. motivation-related words,
11. other process or change words,
12. cognitive orientation (knowing, assessment, and problem solving),
13. pronouns reflecting an "i" vs. "we" vs. "you" orientation as well as names,
14. "yes", "no", negation, and interjections,
15. verb types, and
16. adjective types.

In addition, there are two large valence categories: words of positive outlook and words of negative outlook. The Lasswell Dictionary, which complements the Harvard IV-4 Dictionary, in turn concentrates on issues dealing with value, and its categories are related to power, rectitude, respect, affection, wealth, well-being, enlightenment, and skill (Harvard University, n.d.-c). In both of these dictionaries, many of the categories further expand into subcategories. In addition, the General Inquirer system is enriched with syntactic marker categories (e.g. "Articles", "Genitives", "Prepositions", "Pronouns", "Conjunctions", "Endings", and "Punctuation") and semantic marker categories (e.g. "Animate", "Collective", "Time", "Distance", "Social Place", "Emotions", and "Degree Adverbs") as a resource for disambiguation (Harvard University, n.d.-d). Furthermore, users can develop their own

dictionaries and categories which can be made compatible with the General Inquirer system (Harvard University, n.d.-b).

The basic procedure in this type of content analysis is to identify the relevant "language signs" when and if they occur in text as instances of a particular semantic category, after which they are scored. However, it is very seldom that only one semantic category is used, but, in general, it is rather a selection of semantic categories, since a researcher usually wishes to investigate relationships between of a number of categories which are represented in a given text. Such a selection is referred to as a "content analysis dictionary". Thus, a content analysis dictionary is ideally compiled with a view to testing one or more theories. (Stone et al., 1961, pp. 134–135, 139) By way of illustration, the researchers who carried out a study about discriminating between genuine and simulated suicide notes used the categories: "Roles", "Objects", "Emotional States", "Actions", "Institutions", "Statuses", "Qualities", and "Symbolic Referents" (Ogilvie, Stone, & Schneidman 1966, p. 528). Consequently, results with the very same text may well be very different depending on the selection of categories (Schmidt, 1986, p. 786). However, it is often the case that a few narrow categories do not fully or adequately reveal the complexity of the relationships among content and non-content variables and thus may cause the generation of invalid or limited conclusions. This phenomenon is referred to as "Weber's Paradox". Instead, a broad category scheme should reveal relationships which might not be captured by fewer variables. (Botchway, 1989, p. 42)

2.3.2.2 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a word counting software program which references a dictionary of grammatical and content word categories. It is widely used for quantitative analysis of text in the field of social sciences for a variety of psychological states

and behaviours. Its development began in the 1980s from a series of studies carried out by James Pennebaker in which he examined health improvements resulting from writing about one's thoughts and feelings related to a traumatic or stressful event. In the beginning, Pennebaker used large groups of research assistants to conduct the analysis of the essays, but the method was soon discovered to be too complex, unreliable, and subjective. As a solution, he and his colleague, Martha Francis, developed a program which allowed the derivation of several word count categories relating to emotions and cognitive processes, and over the years, the program has been expanded and improved. In addition to the social sciences, LIWC has also been utilized in the fields of computational linguistics, forensics, marketing, and social computing, for example, to build a lie detector, a status encoder, and a social barometer. (Chung & Pennebaker, 2012, pp. 206–207; Pennebaker, 1993, p. 541)

The core component in the LIWC software is its dictionary, the most recent version of which is named the "LIWC 2015". It contains almost 6,400 words, word stems (such as "hungr*"), and select emoticons, each of them belonging to one or more word categories or subdictionaries (Pennebaker, Boyd, Jordan, & Blackburn, 2015, p. 2). The psychological and content word categories are as follows (Pennebaker et al., 2015, pp. 3–4):

Psychological Processes

- Affective processes
    - Positive Emotion
    - Negative Emotion
        - Anxiety
        - Anger
        - Sadness
- Social Processes

    - Family
    - Friends
    - Female References
    - Male References
- Cognitive Processes
    - Insight
    - Causation
    - Discrepancy
    - Tentative
    - Certainty
    - Differentiation
- Perceptual Processes
    - See
    - Hear
    - Feel
- Biological Processes
    - Body
    - Health
    - Sexual
    - Ingestion
- Drives
    - Affiliation
    - Achievement
    - Power
    - Reward

    - Risk
- Time Orientations
    - Past Focus
    - Present Focus
    - Future Focus
- Relativity
    - Motion
    - Space
    - Time
- Personal Concerns
    - Work
    - Leisure
    - Home
    - Money
    - Religion
    - Death

Furthermore, there are categories for linguistic and other types of information, for example, "Articles", "Auxiliary Verbs", "Negations", "Numbers", "Quantifiers", and "Informal Language" which includes the subcategories "Swear Words", "Netspeak", "Assent", "Nonfluencies", and "Fillers". Thus, one lexicon entry often belongs to more than one category. For example, the word "cried" belongs to the categories "Sadness", "Negative Emotion", "Overall Affect", "Verb", and "Verb Past Tense". Entries for the LIWC dictionary have been collected from various sources, one of them being *Roget's Thesaurus*. (Pennebaker et al., 2015, p. 2–3, 6)

The LIWC framework has been translated into other languages as well. These are: Arabic, Chinese, Dutch, French, German, Italian, Portuguese, Russian, Serbian, Spanish, and Turkish.

2.3.2.3 Minnesota Contextual Content Analysis

Minnesota Contextual Content Analysis (MCCA) is a program which analyzes textual material to discover patterns of emphasized ideas as well as the social context or underlying perspective reflected in texts. It has been used for various types of studies, and the textual material investigated includes, for instance, transcripts of conversations, written documents, such as diaries, organization reports, books, written or taped responses to open-ended questions, media recordings, and verbal descriptions of observations (McTavish & Pirro, 1990, p. 245).

The MCCA dictionary, which the system relies on, includes 116 "idea categories" which are grouped under the following 23 "supercategories"[16]:

- Auxiliary Verbs
- Connectives
- Pronouns
- Conditionals
- Relative Pronouns
- Physical Descriptions
- Becoming Aware
- Role
- Time

[16] The categories have been retrieved from the Help file of the MCCALite dictionary, downloaded from http://www.clres.com/.

- Traditional Nouns
- Pragmatic Nouns
- Emotional Nouns
- Analytic Nouns
- Positive Adjectives
- Negative Adjectives
- Other Adjectives
- Control Verbs
- Analysis Verbs
- Deviance Verbs
- Activity Verbs
- Pressure Verbs
- Positive Reaction Verbs
- Negative Reaction Verbs

The MCCA is one of the different modules incorporated into the DIMAP[17] dictionary creation and maintenance software which improves the MCCA dictionary and the function of the program by allowing the creation of sublexicons for individual categories. The sublexicons are based on various sources, such as the WordNet synonym sets (see section 2.3.3). (Litkowski, 1997) Other dictionaries of the DIMAP framework include, for example, the Alphabetic FrameNet Dictionary and the FrameNet Frame Element Dictionary (CL Research, n.d.).

[17] DIMAP is an abbreviation for Dictionary Maintenance Program.

2.3.3 Related notions

In addition to the above-mentioned systems representing the conceptual analysis method and content analysis method, there are also various other systems which rely on semantic ontologies. This subsection contains a brief overview of those systems which are the most interesting in regard to the topic of this thesis.

WordNet is a large semantic lexical database of English which is designed as a network. Its development at Princeton University was started in 1985 (Miller 1998, p. xv), and since then it has been a very popular source of semantic data among NLP researchers.

When developing the lexical resources, the WordNet project members collected words from various sources, and when the list had become sufficiently long, they started to structure it. The first division was carried out according to part of speech: nouns, adjectives, and verbs were divided in separate groups or "nets". Later, in 1992, a group for adverbs was also created. (Miller 1998, pp. xix) The top level of nouns consists of the following ten categories (Gangemi, Guarino, & Oltramari, 2001, p. 290):

- Abstraction
- Act, Human Action, Human Activity
- Entity
- Event
- Group, Grouping
- Location
- Phenomenon
- Possession
- Psychological Feature

- State

The words which are included in the groups divided by parts of speech are referred to as "synsets" (synonym sets); synsets consist of all the words by which a given concept can be expressed. In this sense, WordNet resembles a thesaurus. The synsets, in turn, are linked to each other by means of various other semantic relations, such as hyponymy, meronymy, and antonymy. However, WordNet differs from a thesaurus in the sense that the relations between concepts and words are coded systematically. In this way, the user can select the relation that guides him from one concept to the next and choose the direction in which he wishes to navigate in this "conceptual space". Furthermore, WordNet contains features of an alphabetical dictionary as well, since it gives definitions and sample sentences for most of its synsets, and it also provides information about morphologically related words. (Fellbaum, 1998, pp. 4, 7–9)

The original English WordNet has constantly grown and evolved. In addition, there are WordNets for dozens of other languages which all follow the Princeton WordNet design (The Global WordNet Association, n.d.). According to Fellbaum (1998, p. 8), the majority of lexicalized concepts are shared among languages, although some languages have words for certain concepts which may not be lexicalized in another language. The Finnish WordNet, which was first released in 2010, will be presented briefly in section 2.5.2.2.

Another important but quite different source of semantic data is FrameNet. This computational lexicography project was initiated for the English language at the International Computer Science Institute (ICSI) in Berkeley in 1997. The lexical database is built on the theory of meaning referred to as "frame semantics" which was developed by Charles J. Fillmore and his colleagues. The basic idea of frame semantics is that the word meanings can be best described and understood in relation to semantic frames. The inspiration for the name

FrameNet came from WordNet, a system also concerned with networks of meaning in which words participate (Fillmore, Johnson, & Petruck, 2003, p. 235). Semantic frames are different types of interactive situations which include all aspects of interaction that they may involve. The situation itself is called a "frame", and participants, props, and other aspects are referred to as "frame elements". Frames are linked to each other with different frame relations. (Baker, 2012, s. 270) An example of a frame is "Apply_heat" which includes the following frame elements (FrameNet, n.d.-a):

- Cook (the person who is doing the cooking)
- Food (the food that is being cooked)
- Container (the object inside or on which the food is cooked)
- Heating_instrument (the source of heat)

By way of illustration, the Cook could be Jamie Oliver, the Food could be crumble, the Container could be a baking dish, and the Heating_instrument could be an oven.

FrameNet has been developed by annotating sentences from real-life corpora which are expected to show how words are actually used. Thus far the annotation work has been carried out manually, which is very time-consuming. (FrameNet, n.d.-a) However, in a recent project, new software tools have been built to facilitate the development process. This project is named "Rapid Vanguarding", and the tools have been modelled on the Sketch Engine (Baker, 2012, p. 274). Sketch Engine is a lexical profiling program developed by Adam Kilgarriff and colleagues (Kilgarriff et al., 2014).

Boas (as cited in Baker, 2012, p. 279) points out that the theory of frame semantics has always presupposed that many frames should be more or less language-independent. Similarly to WordNet, FrameNet was also originally developed for English, but later the framework has

been extended to include parallel FrameNets in various other languages (FrameNet, n.d.-b). The Finnish FrameNet will be briefly presented in section 2.5.3.

Yet another valuable source incorporating machine-readable semantics is offered by the ontologies which have been developed within the framework of the Semantic Web project. This project was initialized by the World Wide Web Consortium which was founded by Tim Berners-Lee. The need for such an initiative arose from the fact that the content of the Internet, consisting of text and pictures, can be easily manipulated by human beings, but it is not easily accessible to computers. If semantic information was incorporated into the Internet, this would make its content more easily machine-processable. (Antoniou, 2012, pp. 1–3) The World Wide Web Consortium is now worldwide, and a Finnish Semantic Web project has also been undertaken. This project will be briefly presented in section 2.5.2.1.

The different sources of semantic information do not necessarily need to be completely separate from one other, but prospects of co-operation have been investigated. FrameNet and WordNet were created for different purposes, and their data structures are different. During the past few years, however, there have been collaborative attempts to align FrameNet and WordNet and make them interoperable in order to produce a resource with would combine the strengths of them both (Baker, 2012, p. 275). A case study which describes how the synsets and definitions of WordNet and the syntagmatic information of FrameNet can complement each other was carried out by Collin Baker and Christine Fellbaum (2009). In addition, there have been plans to utilize FrameNet for the automatic identification and disambiguation of word meanings in the Semantic Web (Narayanan, Fillmore, Baker, & Petruck, 2002).

Similarly, even though thesauri and alphabetical dictionaries differ from each other, they can be used to complement one another. The electronic age has changed the nature of thesauri and alphabetical dictionaries in a revolutionary way. Firstly, the space constraints of the printed format do not apply anymore and, secondly, the abundance of new possibilities to

carry out searches has made it possible to include various novel arrangements to present lexical information. One good example of such a dictionary is the *Macmillan English Dictionary*, first published in 2002, which is targeted at learners of English. In addition to the printed book, it is also available as both CD-ROM (Rundell, 2002) and Web versions (Macmillan, n.d.). Among many useful features to help users with their language skills, all the senses listed in the alphabetical dictionary entries provide a link to the thesaurus containing synonyms and other related words. In addition, all the words in the dictionary definitions and the thesaurus are interlinked, so the user can, for example, click on one of the synonyms the thesaurus has offered him and look up more information on that word in the alphabetical dictionary.

Another way of enriching alphabetical dictionaries with semantic information is the use of domain labels. An example of such a dictionary is the electronic version of the *Collins English Dictionary* (2000) which groups its entries under various subject field codes. The seven major subject field codes are:

1. Arts
2. Business and Economics
3. Recreation and Sports
4. Religion and Philosophy
5. Science and Technology
6. Social Science and History
7. General

These major subject fields are not in themselves coded, but instead, the dictionary entries are coded according to related subfields. The subject field code "General", however, has not been

subdivided at all and thus has been left completely uncoded. In addition, not all the words in the dictionary are coded exhaustively, though most of the domain specific terms are. (Archer et al., 2004, pp. 819–820)

Subject field codes have also been utilized in the WordNet Domains project in which WordNet synsets have been semi-automatically annotated with one or more domain labels. The total number of the categories is 200, and they are organized hierarchically (Fondazione Bruno Kessler, 2009b). The work was motivated in several respects (Magnini & Cavaglia, 2000, p. 1413):

1) Subject field codes provide cross-categorial information which WordNet, for the most part, lacks.
2) Synsets are the appropriate semantic level for subject field code annotation.
3) Subject field codes play an important role in multilingual Wordnet-like resources, since they are considered basically language-independent.

The first two levels of the hierarchy are as follows (Fondazione Bruno Kessler, 2009a):

- Doctrines
  - Archaeology
  - Art
  - Astrology
  - History
  - Linguistics
  - Literature
  - Philosophy
  - Psychology

    - Religion
- Free Time
    - Play
    - Sport
- Applied Science
    - Agriculture
    - Alimentation
    - Architecture
    - Computer Science
    - Engineering
    - Medicine
    - Veterinary
- Pure Science
    - Astronomy
    - Biology
    - Chemistry
    - Earth
    - Mathematics
    - Physics
- Social Science
    - Administration
    - Anthropology
    - Artisanship
    - Body Care
    - Commerce

    - Economy
    - Fashion
    - Industry
    - Law
    - Military
    - Pedagogy
    - Politics
    - Publishing
    - Sexuality
    - Sociology
    - Telecommunication
    - Tourism
    - Transport
- Factotum

The synsets which do not belong to a specific subject field code but rather can appear in almost all of them, were assigned the last subject field code in the list, "Factotum" (Magnini & Cavaglia, 2000, p. 1414).

2.4 UCREL Semantic Analysis System

In this subsection, I will discuss work and research particularly closely related to the topic of this thesis, namely the UCREL Semantic Analysis System (USAS), to which the FST belongs. I will start by presenting the EST and its semantic lexicons which have been used as models when developing the Finnish counterparts. Subsequently, I will briefly introduce other

extensions to the USAS framework which has now evolved into a multilingual semantic annotation system.

2.4.1 English Semantic Tagger

One example of programs undertaking automatic semantic analysis of text is the EST created at UCREL[18] at Lancaster University. It consists of two components: a) semantic lexical resources and b) software which assigns semantic tags to each word in running text. The software achieves this on the basis of information which is contained in the semantic lexical resources as well as in the various rules and algorithms of the EST.

Since the beginning of the 1990s and within the framework of several different projects, the UCREL team has been developing the EST for the annotation of both spoken and written data with the emphasis on general language. The EST was first applied to the analysis of interview transcripts in market research (Wilson & Rayson, 1993) and to the stylistic analysis of written and spoken English (Wilson & Leech, 1993) in projects named ACASD (Automatic Content Analysis of Spoken Discourse) and ACAMRIT (Automatic Content Analysis of Interview Transcripts) as well as to a pilot study of a large corpus of doctor-patient interactions (Thomas & Wilson, 1996). Subsequently, the software was utilized in the REVERE (Requirements Reverse Engineering to Support Business Process Change) project in the area of software engineering (Rayson, Emmet, Garside, & Sawyer, 2001). In our language technology project Benedict, which was introduced in chapter 1, we used semantic taggers for English and Finnish together to build a context-sensitive search tool for a new type of intelligent electronic dictionary. The EST has also been redesigned to create a historical semantic tagger for English (Alexander et al., 2015). In addition, the EST has been applied to:

[18] For more information, see http://ucrel.lancs.ac.uk/.

- analysis of personal weblogs in Singapore English (Ooi, Tang, & Chiang, 2007),
- analysis and standardisation of SMS spelling variation (Tagg, Baron, & Rayson, 2012),
- analysis of the semantic content and persuasive composition of extremist media (Prentice, Taylor, Rayson, Hoskins, & O'Loughlin, 2011),
- corpus stylistics (e.g. Calvo Maturana, 2012),
- detecting gender and spelling differences in Twitter and SMS (Baron et al., 2011),
- discourse analysis (e.g. O'Halloran, 2011; Davis & Mason, 2013; Al-Hejin, 2015),
- finding contextual translation equivalents for words in the Russian and English languages (Sharoff, Babych, Rayson, Mudraya, & Piao, 2006),
- key domain analysis (e.g. Rayson & Smith, 2006),
- language of suicide notes (e.g. Shapero, 2011),
- metaphors in political discourse (e.g. L´Hote & Lemmens, 2009),
- ontology learning (e.g. Gacitua, Sawyer, & Rayson, 2008),
- phraseology (e.g. Granger, Paquot, & Rayson, 2006),
- political science research (e.g. Klebanov, Diermeier, & Beigman, 2008),
- protection of children from paedophiles in online social networks (e.g. Rashid, Greenwood, Walkerdine, Baron, & Rayson, 2012),
- psychological profiling (e.g. Hancock, Woodworth, & Porter, 2014),
- sentiment analysis (e.g. Simm, Ferrario, Piao, Whittle, & Rayson, 2010),
- training chatbots and comparing human-human and human-machine dialogues (Abu Shawar & Atwell, 2003), and
- deception detection (e.g. Markowitz & Hancock, 2014).

The EST is available via the Wmatrix[19] interface, and a complete list of publications and applications using Wmatrix can be found at Lancaster University (n.d.).

2.4.1.1 Semantic tagset

The categories representing different semantic fields are symbolized by codes referred to as "semantic tags", and together these semantic tags form a "semantic tagset". The semantic tagset which the USAS framework employs was loosely based on the categorization used in *LLOCE* (McArthur, 1981; see section 2.3.1.1.2). The UCREL team considered that this offered the most appropriate thesaurus-type classification for the type of sense analysis for which they wanted to develop their semantic tagger. The tagset has since been expanded and amended in the light of lessons learned from the practical tagging problems which were encountered in the course of the research. (Archer, Wilson, & Rayson, 2002, p. 2)

The present USAS tagset has been arranged into a hierarchy of 21 top level semantic categories which further expand into 232 subcategories. With the tagset, everything that exists in the world or can be imagined can be described, whether they be concrete entities or abstract concepts. Each category contains words which are related to each other. These words can be synonyms, antonyms, hyponyms, as well as meronyms, and they represent all parts of speech. Table 1 below displays the top level semantic categories of the hierarchy.

[19] For further information, see Rayson, 2003.

Table 1

Top Level Semantic Categories of the USAS Semantic Tagset

A	General & Abstract Terms
B	The Body & The Individual
C	Arts & Crafts
E	Emotional Actions, States, & Processes
F	Food & Farming
G	Government & The Public Domain
H	Architecture, Building, Houses, & The Home
I	Money & Commerce
K	Entertainment, Sports, & Games
L	Life & Living Things
M	Movement, Location, Travel, & Transport
N	Numbers & Measurement
O	Substances, Materials, Objects, & Equipment
P	Education
Q	Linguistic Actions, States, & Processes
S	Social Actions, States, & Processes
T	Time
W	The World & Our Environment
X	Psychological Actions, States, & Processes
Y	Science & Technology
Z	Names & Grammatical Words

A list of all top level categories and subcategories is presented in Appendix A in English and in Appendix B in Finnish. The reader is advised to consult these appendices if a semantic tag is not explained or clear from context.

A semantic tag consists of various markers as described in Archer et al. (2002, pp. 1–2). A semantic tag always begins with an upper case letter which indicates the top level semantic category. This letter is followed by a digit which indicates the first subdivision in the field. The simplest possible semantic tag contains one upper case letter and one number. For example, the tag for "sentimental" is E1 ("Emotional Actions, States and Processes: General") and the tag for "daffodil" is L3 ("Plants"). If there are more subdivisions, one or two more numbers can be added (e.g. the tag for the verb "reschedule" is T1.1 ("Time: General") and the tag for "tomorrow" is T1.1.3 ("Time: General: Future")). According to Piao et al. (2005a), the depth of the semantic hierarchical structure is limited to a maximum of three layers, since this has been found to be the most feasible approach. In theory, it would be possible to include as many layers of subdivision of meaning until no further subclassification is possible, but semantic field analysis schemes which are too complex may cause problems for practical analysis. That said, the existing semantic categories can be subdivided for a particular task if need be, since the deep hierarchy structure allows to amend the system easily.

In addition to the numbers and digits, it may sometimes be necessary to add one, two, or even three plus or minus markers to the semantic tags to indicate antonymous pairs or a positive or a negative position on a semantic scale. For example, "old" is tagged as T2+, whereas "young" is tagged as T2-; "accessory" is N5++, whereas "inferiority" is A5.1-- ; "archaic" is T3+++, whereas "avant-garde" is T3---. Similarly, comparative and superlative forms of adjectives and adverbs which are formed with inflections are expressed utilizing plus and minus markers. For example, the adjective "easy" has been assigned the semantic tag A12+, the comparative form "easier" is tagged as A12++, and the superlative form "easiest" as A12+++. Moreover, markers "m" and "f" indicating gender are also used. For example, the semantic tag S4f is used for "aunt" and the semantic tag S4m for "bridegroom".

As noted in section 2.3.1.3, not all words always fall neatly into predefined semantic categories but rather are somewhat "fuzzy" sets, where one word can belong to two or even three categories. This multiple membership of categories is indicated in the context of the USAS framework by a "slash tag" (also known as a "portmanteau tag"). By way of illustration, "classroom" is tagged P1/H2, since it can be considered to belong both to the category "Education in General" (P1) and to the category "Parts of Buildings" (H2). "Neurotic" is tagged B2-/X1, where the semantic tag B2 represents the category "Health and Disease", so B2- stands for ill health or disease, and X1 represents the category "Psychological Actions, States, and Processes in General". "Tattoo" is tagged as C1/B1 ("Arts and Crafts" / "Anatomy and Physiology"). The semantic tag for the verb "improve" is represented by A5.1+ ("Evaluation: Good/Bad", with the plus marker indicating "good") and also A2.1 ("Affect: Modify, Change"). Thus, A5.1+/A2.1 stands for "change into good". These markers will be discussed in more detail in section 3.4 with many Finnish-language examples from the equivalent semantic lexical resources for Finnish. In addition, the USAS tagset uses five other symbols (Archer et al., 2002, p. 2), but these will not be discussed here, since they are relatively rare in the English semantic lexical resources and do not appear at all in the Finnish semantic lexical resources.

Unlike many other present-day semantic taxonomies, the USAS semantic tagset is concept-driven rather than content-driven. This means that it aims at providing a conception of the world that is as general as possible, instead of trying to offer a semantic network for specific domains. (Piao et al., 2005a) If or when it is necessary to have a finer-grained taxonomy for a certain task or purpose, it will be relatively easy to expand the present system simply by adding new levels of subcategories or by using more specific slash tags.

2.4.1.2 Semantic lexical resources

The English semantic lexical resources, the knowledge base for the EST, consist of two different parts: 1) lexicon of single words and 2) lexicon of MWEs which contains verb phrases (e.g. "come across"), noun phrases (e.g. "computer file"), multiword proper names (e.g. "United Kingdom"), idioms (e.g. "kick the bucket") depicting single semantic units or concepts. These resources were created manually by first adding semantic tags to the dictionaries of the CLAWS POS tagger (see sections 2.2.3.1 and 2.4.1.4), and, subsequently, they were expanded by adding words which were collected from large text corpora (Piao et al., 2005a). The EST lexicons contain both basic and inflected forms, since there was no reliable lemmatiser[20] available for the English language when the development of the EST started.

The information about the single word lexicon entries can be found in three different columns. The first column indicates the word and the second column its part of speech generated by CLAWS[21]. For example, as shown in the sample below, "misconceptions" is a plural common noun, "misdirected" is the past participle of a lexical verb, and "misguided" is a general adjective. The third column indicates the semantic category. The simplest scenario occurs when the word has only one sense, in which case only one semantic tag will have been attached to the lexicon entry (e.g. in the case of "misfortune" below). If the word is ambiguous, in that it has more than one sense, the different senses are listed in the third column arranged in frequency order (e.g. in the case of "miserably" below). The following is a sample from the single word lexicon:

[20] A lemmatiser is a program which reduces words in the input text to basic forms.

[21] The full CLAWS tagset can be found at http://ucrel.lancs.ac.uk/claws7tags.html. The architecture of the EST will be discussed in section 2.4.1.4.

misconceptions	NN2	A5.2-/X2.1
misdirect	VVI	M6/A5.3- Q2.2/A5.3-
misdirected	VVN	M6/A5.3- Q2.2/A5.3-
miser	NN1	S1.2.2+/S2mf
miserable	JJ	E4.1- N3.2-
Miserables	NP1	Z3
miserably	RR	E4.1- N3.2- A5.1-
misery	NN1	E4.1-
misfit	NN1	A6.2-/S2mf
misfortune	NN1	A1.4
misguided	JJ	X2.5-
mishear	VVI	X3.2/A5.3-

By comparison, the information in the MWE lexicon entry is presented in two columns. The first column in the lexicon entry includes both the MWE and the relevant grammatical and syntactic information, which will be dicussed in more detail in the following paragraphs, and the second column indicates the semantic category. If the MWE is ambiguous, the semantic tags for the different senses are arranged in frequency order in the same way as in the single word lexicon.

All the MWE lexicon entries are written into templates, whereby they consist of patterns of words and grammatical and syntactic information presented in the first column. Often they also contain symbols known as "wild cards" that can represent any character or group of characters. The following is a sample from the MWE lexicon:

dope_NN1 pusher*_NN*	F3/S2mf
dormer_NN1 bungalow*_NN*	H1
doss*_* {R*} about_RP	K1
doss*_* {R*} around_RP	K1

dot*_* every_AT1 i_ZZ1 and_CC cross*_* every_AT1 t_ZZ1	N5.1+
dotted_* {R*/Np/PP*} about_*	M6
dotted_* {R*/Np/PP*} around_*	M6
dot_NN1 matrix_NN1	Y2
doubl*_* {Np/P*/R*} up_RP	N5+/A2.1 A6.1+ E4.1+ S1.1.2+
double-decker_JJ sandwich*_NN*	F1
double_* breasted_*	B5
double_* check*_*	X2.4/N6+

Wild cards enable the EST to recognize MWEs which have similar structures. For example, the template "*_* shortage*_*" would capture the expressions "labour shortage" and "fuel shortages". Furthermore, the EST recognizes not only continuous MWEs, that is, expressions in which it is not possible to add any embedded elements between the constituents (e.g. "dope pusher") but also discontinuous MWEs, that is, expressions inside which it is possible to add varying embedded elements (e.g. "double up"). By way of illustration, the template "doubl*_* {Np/P*/R*} up_RP" would capture both the expression "double up the reward" and the expression "double the price up". As a result, the MWE lexicon covers many more MWEs than is the number of individual entries. All these symbols will be presented and discussed in more detail in section 5.2.2.2 in which I draft guidelines for writing templates for Finnish MWEs.

The semantic lexical resources for the EST were last significantly updated and expanded in 2006, and in the present form they contain 54,953 single words and 18,921 MWEs (Mudraya, Babych, Piao, Rayson, & Wilson, 2006, p. 6). Additionally, the resources include a small autotagging lexicon. This comprises around 50 fixed patterns which can have many possible instantiations. Such expressions can be tagged effectively through the use of wild cards. (Rayson, Archer, Piao, & McEnery, 2004, p. 9) For example, the autotagging lexicon

entry "*kg" (kilograms) would tag all combinations of numbers and the abbreviation "kg" as N3.5 which represents the semantic category "Measurement: Weight".

2.4.1.3 Disambiguation procedures

Just as with POS tagging, the task of semantic tagging can also be broadly subdivided into two phases (Garside & Rayson, 1997, p. 188):

1) Tag assignment: All potential semantic tags are to be attached to each word.
2) Tag disambiguation: From this set of all potential semantic tags, the contextually appropriate one is selected[22].

If a word in a text is included in the semantic lexical resources, if it has only one sense, and if is not a part of a MWE, tagging it correctly is a straightforward task for a semantic tagger. If this is not the case, the task becomes far more difficult, since successful semantic tagging entails being able to both recognize if a word is a single word or part of a MWE and to identify which of the senses is the appropriate sense in a given context if the word has more senses than one.

There are seven procedures which the EST can utilize for the task of semantic tag disambiguation, in other words, for finding the correct semantic tag for the given sense (Garside and Rayson, 1997, pp. 190–192; Piao, Rayson, Archer, Wilson, & McEnery, 2003):

[22] Note, however, that not all systems, such as LIWC, include this phase.

1) POS tag

The first disambiguation method is the POS tagging already introduced in section 2.2. which takes place prior to semantic tagging and is carried out by the CLAWS POS tagger. By way of illustration, "address" can be either a singular common noun or a basic form of a lexical verb:

address	NN1	H4 Q2.2
address	VV0	Q1.2 Q2.2 A1.1.1

If CLAWS determines that the tag NN1 representing a singular common noun is the relevant grammatical tag, this simplifies the task of the semantic tagger by leaving it with only two candidate semantic tags to choose from: the tag H4 representing the category "Residence" and the tag Q2.2 representing the category "Speech Acts".

2) General likelihood ranking for single word and MWE tags

The senses in the semantic lexicon entries have been arranged in frequency order according to information obtained from frequency-based dictionaries, past tagging experience, and intuition of the compilers. The most frequent and thus the most likely semantic tag is placed first, the second most frequent and thus the next likely semantic tag is placed second, etc. As a consequence, if there is no other disambiguation method which the program can apply, it is wisest to use the first tag, since that represents the most common sense and is thus most likely to be the correct tag. By way of illustration, the lexicon entry for the noun "mouse" contains the following tags:

mouse	NN	L2mfn Y2 S1.2.3-/S2mf

The tag NN1 is a POS tag assigned by the CLAWS component and indicates a singular common noun. The POS tag is followed by the relevant semantic tags. The first semantic tag, L2, represents the category "Living Creatures Generally", so the first and thus the most common sense is that of a rodent. The second semantic tag, Y2, represents the category "Information Technology and Computing", so here it refers to the pointing device for the computer. The third and the least likely sense is that of a quiet or timid person, which is represented by the semantic tag S1.2.3-/S2mf.

3) Overlapping idiom resolution

Normally, MWEs take priority over single word tagging. In other words, the semantic tagger first matches the text against the MWE templates, and if it discovers words which match a template and thus together form a MWE, it tags these words together as a unit having the same sense. If no suitable MWE template is discovered, a word is considered to be a single word and tagged individually. However, in some cases, MWE templates can overlap, in that some MWE templates can produce more than one set of possible taggings for the same set of words. To resolve such situations, a set of rules has been developed, whereby these rules help to determine which of the MWE templates is the most likely one and should therefore be favoured. The rules take account of both the length and the span of the MWEs and of how much of the template is matched in each case.

4) Domain of discourse

If the domain or topic of discourse in a given text is known, this information can be used to "weight" tags, in other words, to alter the order of semantic tags in the single

word lexicon and MWE lexicon for a particular domain. Taking the noun "mouse" again as an example:

mouse NN1 L2mfn Y2 S1.2.3-/S2mf

If the topic of discourse in the text dealt with computing, it would be sensible to weight the category Y2 ("Information Technology and Computing") to automatically raise its likelihood, since this would be the most likely sense in this context.

5) Text-based disambiguation

Gale, Church, and Yarowsky (1992, pp. 233–237) carried out experiments with polysemous words to support their hypothesis that well-written discourses tend to avoid multiple senses of polysemous words. Indeed, they discovered that this tendency was as strong as 98%. One of their test words was "sentence", and the same sense repeatedly appeared both in texts which deal with grammar and in texts which deal with the law. If this hypothesis continued to hold in other cases, it would represent an important addition to the methods for determining word senses. This approach has not, as yet, been implemented in the EST, but it resembles the above-mentioned procedure number 4 with the exception that, while in procedure 4 the weighting is adjusted manually, in this approach the weighting would be determined by the program.

6) Template rules

The same type of template rules that are written for the identification of MWEs can also be used for detecting certain senses of words. For instance, when the noun

"account" occurs in a sequence, such as "someone's account of something", it is very likely to mean "narrative explanation" and not "bank account".

7) Local probabilistic disambiguation

It is generally supposed that the local surrounding context determines the correct semantic tag for a given word. Thus, the surrounding context can be identified in terms of a) the words themselves, b) their grammatical tags, c) their semantic tags, or d) some combination of all three. An application of this method named the "Domain Detection System" was developed in the Benedict project, where the most probable sense of a word was calculated by making use of information about the other words in the same sentence. The Domain Detection System is described in more detail in Löfberg et al., 2004.

2.4.1.4 Program architecture and evaluation

The EST is built on four components. They are: 1) the CLAWS POS tagger, 2) the lemmatiser, 3) the semantic tagging component, and 4) the auxiliary manipulating components, such as the disambiguation template rules described in the previous subsection and the small auto-tagging lexicon described in section 2.4.1.2. The lemmatiser was incorporated into the EST during the Benedict project for the dictionary look-up function. The semantic tagging component consists of the semantic lexical resources (described in section 2.4.1.2) and the software that implements algorithms of semantic disambiguation which then automatically links words in a text to one or more semantic categories. The following figure illustrates the multi-level structure of the EST:

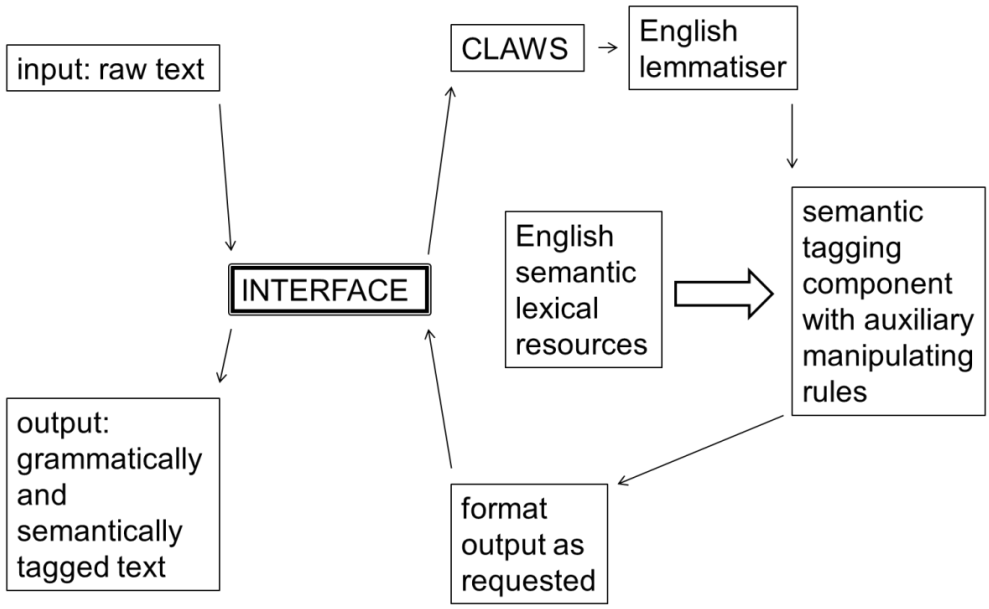

Figure 1. Architecture of the English Semantic Tagger

When text is entered into the EST, CLAWS analyses the text grammatically and assigns each word all possible grammatical tags. In the next phase, in order of likelihood, the lemmatiser finds the basic form of the word. Thereafter, the semantic tagging component matches the patterns of the output against the patterns in the single word and MWE lexicons, utilizing the auxiliary manipulating components, and then assigns each single word and MWE a semantic tag which denotes its meaning. For example, the sentence "It was very warm and summery yesterday, and many people sat on a park bench to enjoy the warm weather." is tagged in the following way:

0000001 002	-----	-----	
0000002 010	PPH1	It	Z8
0000002 020	VBDZ	was	A3+ Z5
0000002 030	RG	very	A13.3
0000002 040	JJ	warm	O4.6+ O4.2+ S1.2.1+

0000002 050	CC	and	Z5
0000002 060	JJ	summery	T1.3
0000002 070	RT	yesterday	T1.1.1
0000002 071	,	,	
0000002 080	CC	and	Z5
0000002 090	DA2	many	N5+
0000002 100	NN	people	S2mfc
0000002 110	VVD	sat	M8 C1 P1 G1.1 G2.1 M6 A9+
0000002 120	II	on	Z5
0000002 130	AT1	a	Z5
0000002 140	NN1	park	M7/L3
0000002 150	NN1	bench	H5 G2.1c
0000002 160	TO	to	Z5
0000002 170	VVI	enjoy	E2+ A9+ E4.1+
0000002 180	AT	the	Z5
0000002 190	JJ	warm	O4.6+ O4.2+ S1.2.1+
0000002 200	NN1	weather	W4
0000002 201	.	.	

The EST has been tested many times with excellent results. The latest evaluation of the semantic lexical resources for the EST was carried out by Piao, Rayson, Archer, and McEnery (2004) in which an indicator referred to as "lexical coverage" was calculated. Lexical coverage shows how many of the single words and MWEs in the test corpus the EST recognizes, in other words, how many of those single words and MWEs are included in the semantic lexical resources. The results from tagging modern English were reported to range between 98.49% for the BNC Sampler Corpus (1,956,171 words) and 95.38% for the METER Corpus of law and court journalism reports (241,311 words), which is an outstanding result and demonstrates that the semantic lexicons are able to deal with most general domains. The

lexicons were also tested on six different historical corpora, and the results ranged between 92.76% and 97.29%. The overall performance of the EST was evaluated by Rayson et al. (2004), where a corpus containing 124,900 words from transcriptions of 36 informal conversations was used for the experiment. The corpus was tagged with the EST, after which it was manually corrected by a team of post-editors. The accuracy[23], that is the number of single words and MWEs the EST is able to recognize and tag correctly, was calculated at 91.05% which is an excellent result. Moreover, Piao et al. (2003) carried out an evaluation of the accuracy of the MWE component of the EST using the newspaper section of the METER corpus, which consists of more than 250,000 words, as test material. In this case, the result was 90.39%[24] which was shown to be comparable to other existing systems.

2.4.2 Extension of the semantic tagger framework for other languages

The EST has functioned as a basis for the development of equivalent semantic taggers for other languages. Such equivalent tools also enable the development of multilingual NLP, text mining, and other information and communications technology (ICT) systems. The first non-English semantic tagger was the FST described in this thesis, while the second was the Russian Semantic Tagger (RST). The latter was developed in the ASSIST (Automatic Semantic Assist for Translators) project to provide contextual examples of translation equivalents for words from the general lexicon between English and Russian languages (Mudraya et al., 2006). The development of the FST and the RST was a relatively similar process, involving the modification of the software framework originally created for English to meet the needs of the analysis of Finnish and Russian respectively. In addition, this work

[23] Also the term "precision" has been used in the same context when evaluating the performance of the USAS semantic lexicons.

[24] All in all, 3,792 MWEs out of 4,195 were tagged correctly.

involved the manual construction of the semantic lexicons, the knowledge base for the programs, which is a very time-consuming task.

During recent years, new semantic taggers have been developed, and new methods have been utilized to carry out the lexicon development much more rapidly. These methods involve bootstrapping new semantic lexical resources via automatically translating the English semantic lexicons into other languages. This proved a very successful approach for languages for which there are appropriate high-quality bilingual lexicons available (Piao, Bianchi, Dayrell, D'Egidio, & Rayson, 2015). At present, there are also equivalent semantic taggers Czech, Chinese, Dutch, French, Italian, Malay, Portuguese, Spanish, Urdu, and Welsh. The lexical coverage potential of 12 languages was recently evaluated in Piao et al. (2016). Furthermore, there are plans now to extend the USAS framework next for Arabic, Norwegian, and Swedish. The semantic taggers are available via the USAS Web interface[25].

2.5 Key Points on the Finnish Language

In this section, I will briefly introduce the Finnish language and provide a brief overview of some of its specific grammatical features which have had an effect on the development of the Finnish semantic lexical resources and the FST software more generally. This will help non-Finnish speakers to understand the discussion about the grammar and structure of Finnish in the subsequent chapters. Following that, I will review previous work and research which has been reported on the creation of related lexical resources for Finnish.

[25] For more information, see http://ucrel.lancs.ac.uk/usas/.

2.5.1 Origins and structure of Finnish

Finnish belongs to the Finno-Ugric language family together with, for example, Estonian, Hungarian, the Sámi languages (spoken in the North of Finland, Norway, and Sweden, and in the far north-west of Russia), as well as Karelian, Vepsian, Ludian, Votian, and Livonian (spoken in Russia, around the south and east of the Gulf of Finland) (Karlsson, 1999, p. 1). Finnish is spoken by the vast majority of the people living in Finland whose population is at present almost 5,500,000 people, with the second official language being Swedish. Finnish is one of the official languages of the European Union.

Finnish differs from English in many respects. For example, Finnish does not contain grammatical gender, with the pronoun *hän* being used to refer to both males and females. Finnish uses the Latin alphabet set similar to the English alphabet with three additions: the characters *å*, *ä*, and *ö*. The writing mainly corresponds to the pronunciation, and the main word stress is on the first syllable.

A very distinctive feature of Finnish is its rich morphology. Finnish is an agglutinative, synthetic language, whereas English is by and large an analytic language. Finnish predominately uses inflections where English uses prepositions, which often results in fairly long words in Finnish. In their study, Heikkinen, Lehtinen, and Lounela (2001, p. 13) used the Finnish Parole corpus to investigate various aspects of Finnish. Among other things, they calculated that the average length of a single Finnish word is 8.5 characters. By comparison, the average word length in English is approximately five characters (Kornai, 2007). At the same time, Finnish sentences usually contain fewer words than the equivalent sentences in English (see, for instance, the example sentences in section 3.5). The discussion about Finnish morphology will continue in section 2.5.1.1. Furthermore, the fact that Finnish does not use articles, which are usually very short words, increases the average word length. Only

approximately 10% of Finnish words in running text are grammatically ambiguous, because the stems of Finnish words are relatively long compared to English, and often the inflections reveal the part of speech in question (Koskenniemi, 2013, p. 18). The rich morphology permits a relatively flexible word order in a sentence, since abundant information about the part of speech and the syntactic function of a word is usually attached to the stem of the word in the form of various endings and enclitic particles. It is often, therefore, possible to change the order of words without changing the core meaning of the sentence or making it incomprehensible. In many other languages, such as English, this would not be possible because of all the separate articles, prepositions, etc., but it would simply result in confusion. The discussion about Finnish word order will continue in section 2.5.1.3.

The most important methods for forming new words in Finnish are compounding and derivation. Approximately 10–15% of dictionary entries are basic words, 20–30% are derivatives, and 60–70% are compounds. (Koskenniemi et al., 2012, p. 47) By way of illustration, the words *käsittää* ("to understand", "to comprise"), *käsitys* ("impression", "opinion"), *käsitellä* ("to handle", "to treat", "to deal"), *käsittely* ("handling", "treatment", "hearing"), *käsitteellinen* ("conceptual"), *käsitettävyys* ("comprehensibility"), and *käsittämättömyys* ("incomprehensibility") have all been derived from the word *käsi* ("hand") (Ruppel, forthcoming). Compounding will be discussed in more detail in section 2.5.1.2.

In the Benedict project, we adapted the semantic tagger which was originally developed for the analysis of the English language to meet the needs of semantic analysis of Finnish. With regard to the software component, we noted in the course of the development process that the basic architecture of the EST was applicable for the semantic analysis of Finnish as well, but some modifications were still necessary. Moreover, the semantic lexical resources for Finnish were created from scratch by the author. In the following three subsections, I will provide a brief overview of those particular features of Finnish which had an effect on the

development of the FST, both in terms of the semantic lexicons and the software. The reader interested in a more comprehensive account of the Finnish grammar is referred to Karlsson (1999, written in English) and Hakulinen et al. (2004, written in Finnish).

2.5.1.1 Rich morphology

Finnish often uses endings where many Indo-European languages make use of independent words, such as prepositions, postpositions, and possessive suffixes. The number of case endings in Finnish nominals is quite high. There are 15 of them in all, whereas, for example, English uses only one: the genitive. In addition, there are possessive suffixes as well as various enclitic particles which can be used to indicate emphasis or to form a direct question. All these types of endings can appear attached to a nominal, but their order is always fixed. It is: 1) number, 2) case ending, 3) possessive suffix, and 4) one or two enclitic particles. (Karlsson, 1999, pp. 4–6, 20, 228–230) For example, it is possible to produce the following combinations from the noun *kutsu* ("invitation") and the endings *-i* (number: plural), *-ssa* (case ending for inessive which indicates "in"), *-ni* (possessive suffix for singular first person: "my") and *-kin* (enclitic particle: "too"):

kutsu/ssa	in the invitation
kutsu/i/ssa	in the invitations
kutsu/ssa/ni	in my invitation
kutsu/i/ssa/ni	in my invitations
kutsu/kin	the invitation too
kutsu/t/kin	the invitations too
kutsu/ni	my invitation

kutsu/ni	my invitations
kutsu/ni/kin	my invitation too
kutsu/ni/kin	my invitations too
kutsu/ssa/kin	in the invitation too
kutsu/i/ssa/kin	in the invitations too
kutsu/ssa/ni/kin	in my invitation too
kutsu/i/ssa/ni/kin	in my invitations too

Finnish verbs are formed in similar manner. Finite verb forms (in other words, forms with a personal ending) inflect for person (6 personal endings), mood (4 moods which express, for example, the speaker's attitude), tense, and the passive. In addition, these endings can be followed by enclitic particles. (Karlsson, 1999, p. 20–23) By way of illustration, the verb *ostaisitkohan* ("I wonder if you would buy") consists of the following components:

osta	stem of the verb *ostaa* ("to buy")
isi	ending indicating the conditional mood
t	ending indicating the singular second person
ko	enclitic particle indicating a direct question
han	enclitic particle used for softening the request

There are also non-finite verb forms (in other words, forms which do not contain personal endings): infinitives, of which there are three important types[26], and two types of participles. Some non-finite forms can be inflected in the passive voice like finite verbs, but unlike finite verb forms, non-finite verb forms often take a possessive suffix and a case ending, since

[26] A fourth infinitive does exist, but it is very rare.

infinitives act in the same manner as nouns and participles act in the same way as adjectives. Moreover, participles can be inflected for number. (Karlsson, 1999, pp. 24–25) As is the case for nominals, the order of the ending types is fixed for verbs as well.

Altogether, Finnish nouns can have approximately 2,000 different forms, adjectives approximately 6,000 different forms (comparatives and superlatives triple the number), and verbs a total of 12,000–18,000 different forms (Koskenniemi, 2013. pp. 10–11)[27]. Needless to say, Finnish words can thus be very long and contain a considerable amount of information. However, since the endings are added on one after another systematically, it is not difficult to analyze them, either for a human being or for a computer, once it is determined which part of the word is the stem and which are the different endings attached to that stem. The solution which we adapted to process these in the Benedict project will be described in section 3.3.1.

2.5.1.2 Productive use of compounding

A very common means of forming new words in Finnish is compounding. In this thesis, the term "compound" is used to refer to those words which are formed by concatenating two or more words without a space between them. Compounds are most often formed from nouns, but other parts of speech can also appear in compounds. By comparison, where Finnish uses compounds, English often uses MWEs, for example, *eläin=laji*[28] ("animal species" or "species of animal") and *laki=kirja* ("statute book").

Hakulinen et al. (2004, p. 388) differentiate between two main types of compounds. The most common type, determinative compounds, consists of constituents which have a semantically non-symmetrical relationship with each other. More precisely, the latter element

[27] The 30th anniversary seminar of Lingsoft, a Finnish language technology company, was held 25 November 2016, and, in fact, Kimmo Koskenniemi mentioned in his presentation that there are in all a quadrillion different word forms in Finnish.

[28] The symbol "=" is used in this subsection to mark boundaries between the compound constituents.

is dominant and more significant for the meaning, whereas the former element modifies the latter part. (Hakulinen et al. (2004, p. 396). By way of illustration, *ruoka=lusikka* ("table spoon") is a kind of spoon and *kana=keitto* ("chicken soup") is a type of soup. In the above examples, the first constituent of the compound is in the nominative, but other cases can appear as well, most often the genitive which is indicated by the ending *-n*. This is the case, for instance, in the compounds *ruoa<u>n</u>=laitto* ("cooking"; literally "food's=making"), *koira<u>n</u>=ilma* ("bad weather"; literally "dog's=weather"), and *taivaa<u>n</u>=sininen* ("sky-blue"; literally "sky's=blue"). It can also be the case that compound constituents differ from the basic form and never appear in the language in isolation or in an inflected form. This phenomenon is known as "casus componens" (Hakulinen et al., 2004, pp. 393–394, 402–404). By way of illustration, this is evident in the compounds *hevos=jalostus* ("horse breeding") and *kolmi=loikka* ("triple jump") in which the first constituent never appears in isolation or inflected. Similarly, in the compounds *kuusi=vuotias* ("six-year-old") and *vihreä=silmäinen* ("green-eyed"), the second constituent does not appear in isolation or inflected. Moreover, in the compounds *kansallis=mielinen* ("nationalistic"; literally "national=minded") and *seitsen=kertainen* ("seven-fold"), neither the first nor the second constituent appears in isolation.

The above examples comprise determinative compounds which have meanings that are more or less the sum of the compound constituents. However, there are also such determinative compounds, where the meanings cannot easily be deduced from the sum of the meanings of the compound constituents. Examples of such compounds are *tieto=kone* ("computer"; literally "knowledge=machine") and *potku=housut* ("playsuit" (for a baby); literally "kick=trousers"). Such items are referred to as "lexicalized compounds" in this thesis.

The second common compound type is that of copulative compounds. These consist of two or more compound constituents which are in a symmetrical relationship with each other.

In other words, they represent the same part of speech and their relationship is semantically additive. A hyphen is often used to differentiate between the constituents. (Hakulinen et al., 2004, p, 416) Examples of such compounds are the noun *tutkija-opettaja* ("researcher and teacher") and the adjective *sini=vihreä* ("blue and green"). Furthermore, numerals are also often written as one single word in Finnish and thus resemble copulative compounds, for instance, *viisi=tuhatta=viisi=sataa=kuusi=kymmentä=kuusi* ("five thousand five hundred sixty six") (Hakulinen et al., 2004, p, 388).

Compounding is indeed a very productive means of word formation. For example, it is possible to form names for various soups by combining the names of the main ingredients with the noun *keitto* ("soup"). Thus, we get *kala=keitto* ("fish soup"), *parsa=keitto* ("asparagus soup"), etc. Similarly, an abundance of names for injuries can be produced by combining the names of different body parts with the noun *vamma* ("injury"): *nilkka=vamma* ("ankle injury"), *kallo=vamma* ("skull injury"), etc. Nor does the evident wealth of possibilities end here. By way of illustration, one can add the noun *resepti* ("recipe") at the end of all different types of soups resulting in a multitude of new nouns, such as *kala=keitto=resepti* ("fish soup recipe"). Or one can add the noun *spesialisti* ("specialist") at the end of the above compounds indicating different types of injuries, again resulting in many new combinations, such as *kallo=vamma=spesialisti* ("skull injury specialist"). This type of productivity can eventually lead to very long words, such as *kala=keitto=resepti=valikoima* ("selection of fish soup recipes") and *kallo=vamma=spesialisti=ryhmä* ("group of skull injury specialists"). Complex compounds can even correspond to complete sentences. An example of such a case by Karlsson (1999, p. 242) is the compound *prahassa=käymättömyys=kompleksi* which translates into English as "a complex about not having been to Prague". It is clearly evident that the number of possible compounds is

nnumerable, and it would thus not be sensible or even possible to try to include all of them in a dictionary as entries, but only the most commonly appearing ones are included.

As pointed out above, usually a compound functions as a single word in a clause. In an elliptic compound construction, however, one or more compound constituent, either at the beginning or at the end of the compound, can be omitted and replaced by a hyphen for abbreviation purposes in a list of compounds (Hakulinen, 2004, p. 420). Examples of such compounds are:

- *viini=pullo ja -lasi* which is abbreviated from *viini=pullo ja viini=lasi* ("wine bottle and wine glass")
- *kana- liha- tai kasvis=keitto* which is abbreviated from *kana=keitto*, *liha=keitto tai kasvis=keitto* ("chicken soup, meat soup, or vegetable soup")

Elliptic compound constructions are very seldom found in dictionaries, but they appear relatively frequently in running text.

Understanding the meaning of a compound which consists of many constituents and is not included in a dictionary is not usually very difficult for a human being, since he can intuitively split such words and look for the meaning of the constituents separately, if need be. However, this task is far more complicated for a computer, since if a word is not included in a dictionary or a lexicon, it remains unidentified. Thus, where there is a need to analyze Finnish text automatically, it is necessary to develop mechanisms which help the program to identify and process all possible instances of Finnish compounds. One such mechanism is the "compound engine" which we developed in the Benedict project. The compound engine will be described in section 3.3.2.

2.5.1.3 Relatively flexible word order

In a Finnish sentence, the subject, verb, and object or predicate usually follow the order: 1) subject, 2) verb, 3) object/predicate. Generally, the clause indicating ownership comes before the verb, whereas the clause indicating what is owned follows the verb. (Hakulinen, 2004, p. 1303) Examples of such sentences are:

Hän (S) *osti* (V) *kengät* (O). ("He bought shoes.")

Auto (S) *on* (V) *likainen* (P). ("The car is dirty.")

Minulla (owner) *on* (verb) *koira* (what is owned). ("I have a dog.")

These most common types of orders are termed neutral or unmarked. However, the elements in the above sentences could also be placed in other orders. These other, less common orders would be equally grammatically correct, but the thematic structure of the clause would change, resulting in new emphases and nuances. Sentences can also contain various adverbials, attributes, and adjuncts which may be positioned quite freely within the sentence. However, if the sentence elements have modifiers (e.g. *uudet* *kengät* ("new shoes"), these remain attached to the headword.

Understanding sentences in which the sentence elements have varying orders is not demanding for a human being, but where the purpose is to develop computer software for the analysis of Finnish, the numerous possible variations in the word order need to be taken into account. This issue will be elaborated on further in section 5.2.2 in which the creation of the MWE lexicon for Finnish will be discussed.

2.5.2 Previous work related to large machine-readable semantic lexical resources for Finnish

So far, relatively little work and study has been reported on the creation of large machine-readable semantic lexical resources for the Finnish language apart from two national research projects. In the following subsections, I will describe briefly the lexical resources of these projects: the Finnish Semantic Web and the Finnish WordNet. These lexicons are quite different from the semantic lexicons dealt with in this thesis. Firstly, the semantic lexical resources developed for the FST are intended for full text analysis and thus contain words and MWEs representing all parts of speech, whereas these other lexicons contain words and MWEs representing a limited set of parts of speech. The second difference is that the semantic lexical resources developed for the FST use semantic fields as an organizing principle, while the others are built applying other organizing principles.

2.5.2.1 Finnish Semantic Web

The National Semantic Web Ontology Project in Finland (FinnONTO), which lasted from 2003 to 2012, was launched to develop a Finnish-language open-source foundation for a national metadata ontology, an ontology service, and a linked data framework in Finland, as well as to demonstrate its usefulness in practical applications. The work was based on the Semantic Web Project which was initialized by the World Wide Web Consortium and briefly introduced in section 2.3.3. The consortium consisted of over forty public organizations, companies, and universities representing a wide area of the functions of society, such as museums, libraries, business, health organizations, government, media, and education (Semantic Computing Research Group, n.d.-a). The Semantic Computing Research Group

(SeCo) at the Aalto University and the University of Helsinki was, until the end of 2013, responsible for the development of the ontologies as well as for the ontology server framework named ONKI. From the beginning of 2014, the National Library of Finland has been in charge of its maintenance and further development as a thesaurus and ontology service named Finto[29]. (Seppälä & Hyvönen, 2014, p. 1)

The core content of the FinnONTO lexicons is formed by a shared top ontology named the Finnish General Upper Ontology (YSO) and various domain ontologies which together form the KOKO ontology cloud (Seppälä & Hyvönen, 2014, p. 1). The YSO ontology is based on the General Finnish Thesaurus (YSA) which is maintained by the National Library of Finland. The relations between its entries can be either "subclass-of" relations (hyponymy), "part-of" relations (meronymy), or "instance-of" relations (Hyvönen, Viljanen, Tuominen, & Seppälä, 2008, p. 98). The horizontal top level of the ontology is divided into the categories displayed in the following figure, and the related vertical domain ontologies extend its class hierarchy into a framework of different application domains:

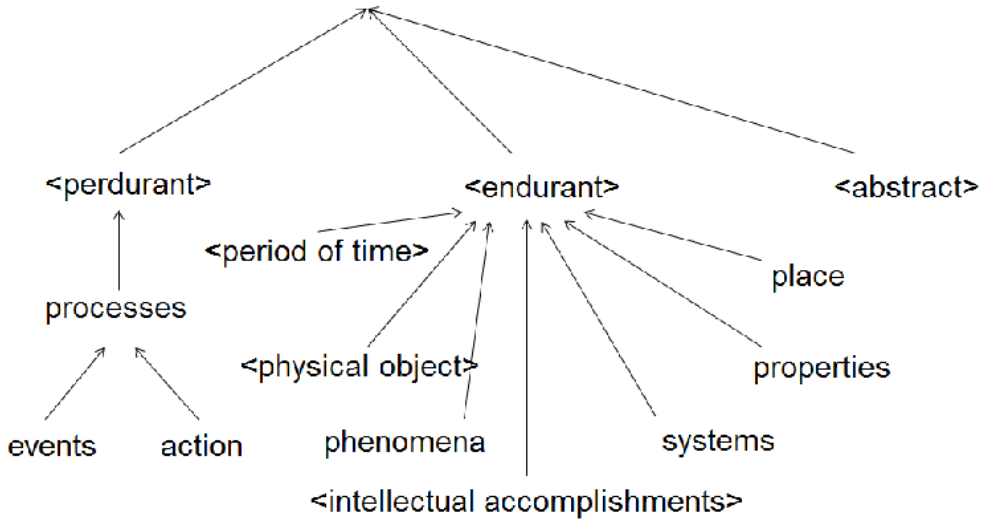

Figure 2. Top level of the Finnish General Upper Ontology (Katri Seppälä, personal communication, August 24, 2011)

[29] For more information, see http://finto.fi/en/about.

At present, the KOKO ontology cloud in its entirety contains the following manually constructed subject matter ontologies (Semantic Computing Research Group, n.d.-b):

- General (Finnish General Upper Ontology YSO; 26,000 concepts)
- Agriculture and forestry (AFO; 6,000 concepts)
- Government (JUHO; 6,400 concepts)
- Fiction literature (KAUNO; 5,100 concepts)
- Literature research (KITO; 850 concepts)
- Linguistics (KTO: 950 concepts)
- Culture research (KULO; 1,500 concepts)
- Business (LIITO; 3,400 concepts)
- Seafaring (MERO; 1,400 concepts)
- Cultural heritage (MAO/TAO; 6,000 concepts)
- Music (MUSO; 1,400 concepts)
- Defence (PUHO; 2,000 concepts)
- Applied arts (TAO: 2,500 concepts)
- Health (TERO; 6,500 concepts)
- Photography (VALO; 2,000 concepts)

In addition, SeCo has also developed, for example, actor, place, time, event, and biological ontologies.

The ontologies consist of nouns only. In general, these nouns consist of a single word; MWEs are not common. Verbs are nominalized, and, for example, the words *hyvä* ("good") and *paha* ("evil"), which can be both adjectives and nouns, are included as nouns. (Katri Seppälä, personal communication, August 24, 2011)

2.5.2.2 Finnish WordNet

Another type of a large lexical resource is WordNet, with the Finnish-language version Finnish WordNet (FiWN[30]) being first released in December 2010 (Lindén, Niemi, & Hyvärinen, 2012, p. 68) for the purposes of language technology research and applications within the FIN-CLARIN consortium[31] at the University of Helsinki. FiWN conforms to the WordNet framework (see section 2.3.3) which was originally created at Princeton University (Lindén & Carlson, 2010, p. 119)

Different approaches to creating a WordNet have been taken when the network has been extended for new languages. The FiWN opted for manual translation of the 117,659 English synsets of the Princeton WordNet 3.0 into Finnish, resulting in aligned WordNets (Lindén et al., 2012, p. 68). The work was carried out in four months by professional translators with SDL Trados program (Lindén & Carlson, 2010, pp. 126, 129). The creators of the FiWN based the direct translation approach on the assumptions that 1) most synsets in the Princeton WordNet represent language-independent real-world concepts and 2) the structure of the Princeton WordNet is reusable, since the semantic relations between synsets are mostly independent of the language (Lindén et al., 2012, p. 68). By comparison, the Polish WordNet was created from scratch: they compiled the synsets by utilizing information drawn from vast corpora. A third possible approach has been to apply the top ontology of the Princeton WordNet by using a selection of its 5,000 basic concepts translated and then expanding the resource further with the aid of a local dictionary. This approach was adopted, for instance, in the creation of the Danish WordNet. (Lindén & Carlson, 2010, pp. 121–123)

---

[30] For more information, see http://www.ling.helsinki.fi/en/lt/research/finnwordnet/.

[31] The FIN-CLARIN consortium belongs to the European CLARIN collaboration which aims to build an infrastructure for language resources and technology. For more information, see https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiFrontpage.

The current version 2.0 of the FiWN was released in October 2012. The resource has been expanded by using Wikipedia and Wiktionary as sources for automatically finding new additions, such as words and senses still found to be missing. This has been carried out by utilizing the interlanguage links between the Finnish and English Wikipedia and the explicitly marked translations between the Finnish and English Wiktionary. (Lindén et al., 2012, pp. 68, 70)

2.5.3 Related notions

In addition, there are also other systems developed for Finnish which rely on semantic ontologies. This subsection contains a brief overview of those systems which are most interesting in regard to the topic of this thesis.

The FrameNet framework, which was presented in section 2.3.3, is now being extended to Finnish as well within the FIN-CLARIN consortium. Similarly to the above-mentioned FiWN, the Finnish FrameNet is developed utilizing translation. Approximately 615,000 translation units from the English FrameNet data have now been translated into Finnish. In the next phase, the developers have been searching for the translation matching examples from corpora to verify that the frames can indeed be moved to a second language without resulting in poor quality "translationese". (Krister Lindén, personal communication, June 10, 2014) The finalized database will be available open source under the Creative Commons licence.

Another recent creation for semantic role labeling of Finnish text is the Finnish Proposition Bank which is an annotated corpus of semantic roles. It is being developed by the Turku BioNLP Group[32]. The Finnish Proposition Bank utilizes much more generic labels than

[32] For more information, see http://bionlp.utu.fi/index.html.

FrameNet, and it is intended for corpus annotation rather than as a lexical resource. (Haverinen et al., 2013)

Some commercial applications utilizing semantic analysis and thesauri for Finnish exist as well. These have been developed, for example, by Connexor[33], Etuma[34], Leiki[35], and Lingsoft[36]. However, they are not discussed here due to the lack of objective, comparable evaluations.

As I already pointed out in section 2.3.1.3, there are no thesauri for Finnish. Two synonym dictionaries exist, the *Synonyymisanakirja* ("Synonym Dictionary") (Jäppinen, 1989) and the *Synonyymisanasto* ("Synonym Lexicon") (Leino & Leino, 1990). These are, however, mere synonym finders which list only total and partial synonyms in their entries. Overall, Finnish lexicography is not as advanced as for some other languages. The history of Finnish lexicography is considered to have begun in 1637 with the publication of Ericus Schroderus's Finnish word list *Lexicon Latino-Scondicum*, which contained 2,400 words, and it was subsequently followed by some bilingual dictionaries. However, the first completely monolingual professionally compiled large-scale dictionary for Finnish was published only between the years 1951 and 1961 in six volumes. The work on this, the *Nykysuomen sanakirja* ("The Dictionary of Modern Finnish"), however, had already been started in 1929. After being ruled by Russia and Sweden, Finland finally gained independence in 1917. In 1927, the Finnish Parliament decided that a dictionary of the modern Finnish language should be compiled, and thus the work on the *Nykysuomen sanakirja* was started two years later. (Ruppel & Sandström, 2014, pp. 143–145, 151) Later, two large-scale monolingual dictionaries have been published with state funding: the *Suomen kielen perussanakirja* ("The Basic Dictionary of the Finnish Language"), which was published in 1990s, and the

[33] For more information, see http://www.connexor.com/nlplib.

[34] For more information, see http://www.etuma.com/.

[35] For more information, see http://www.leiki.com/.

[36] For more information, see http://www.lingsoft.fi/?lang=en.

Kielitoimiston sanakirja ("The New Dictionary of Modern Finnish") which is the most comprehensive and up-to-date monolingual Finnish dictionary available nowadays. These dictionaries were compiled by the Institute for the Languages of Finland, a state-funded organization, which carries out language planning and lexicography for the languages spoken in Finland[37]. As to rivals compiled by private publishers, Ruppel (forthcoming) comments that "These dictionaries are left beyond the scope of this discussion, since they would require a space that would not match their importance." Furthermore, according to Ruppel (personal communication, September 22, 2016), Finnish publishing houses have now forsaken the publication of large bilingual dictionaries altogether. As a result, in the future, Finns can familiarize themselves with different languages only indirectly via other languages, such as English, which is a worrying development.

2.6 Chapter Summary

In this chapter, I have established the background for this thesis. I have first defined the most important related concepts starting with corpus linguistics and then moved on to successively more specialized concepts which are: corpus annotation, linguistic annotation, POS tagging, parsing, and semantic tagging. Semantic tagging is one method of carrying out linguistic annotation, and the necessary pre-processing for semantic tagging is provided by POS tagging and parsing. Thereafter, I have reviewed some examples of semantic ontologies which represent the conceptual analysis method and the content analysis method. The semantic lexical resources, which this thesis focuses on, are arranged according to an ontology which represents the conceptual analysis method. However, the content analysis method is also relevant for this study, since possible domain-specific extensions to the system

[37] For more information, see http://www.kotus.fi/en.

would represent this approach (see section 5.3). In addition, I have also briefly discussed some other, less related systems which rely on semantic ontologies.

Following that, I have presented the USAS framework. Their most important undertaking has been the development of the EST and its applications to various fields and purposes; they represent the state-of-the-art in the field. The EST and its semantic lexical resources have functioned as a model for the development of the FST and the Finnish semantic lexical resources. In addition, I have introduced briefly other extensions to the USAS framework which has now evolved into a multilingual semantic annotation system.

I have concluded this chapter with a brief account of the key points of the Finnish language, concentrating on those specific grammatical features of the language which have had an effect on the development of the FST, both in terms of the semantic lexical resources and the software. These features include its rich morphology, the productive use of compounding, and the relatively free word order. Coping with these various features is not difficult for a human being, but for a computer it is a challenging task which requires a variety of solutions. These solutions, as well as the development of the Finnish semantic lexical resources, will be described in the following chapter.

3 Semantic Lexical Resources for the Finnish Semantic Tagger

3.1 Introduction

This chapter describes the Finnish semantic lexical resources that form the main contribution of this thesis. I will provide a detailed description of the Finnish semantic lexicons, and I will also elucidate how these resources differ from the English semantic lexicons, both in terms of content and of construction. This chapter answers **RQ1 (What do the Finnish semantic lexical resources consist of, what type of principles and practices have been followed in their creation, and how do these resources differ from their English counterparts both in terms of content and construction?)**.

Similarly to the EST described in section 2.4.1, which has functioned as a model for our Finnish counterpart, the FST also consists of two components:

1) semantic lexical resources and
2) software which assigns semantic tags to each word in running text on the basis of information contained in the semantic lexical resources as well as in the various rules and algorithms of the program.

The semantic lexical resources were created by the current author, whereas the software component was developed collaboratively by Lancaster University, Kielikone, and the current author.

I will first look at the initial phases of the development process of the FST, and, subsequently, I will summarize the development and the structure of the software component. Although the FST software is not the main focus of this thesis, it is essential to start from it. The reason for this is that it is not possible to develop semantic lexical resources such as ours in isolation, but the software in which it will be applied needs to be taken into account in many respects throughout the development process. Thereafter, I will provide a detailed description of the principles and practices which I have followed when creating the semantic lexical resources, and I will then proceed to depict their contents. The semantic lexical resources are my most important contribution to the FST and the main focus of this thesis. Finally, I will illustrate the output of the FST.

3.2 Initial Phases

In the Benedict project (see section 1.1), we began the practical development of the FST by building parallel semantically tagged test and training corpora for Finnish and English in order to test the feasibility of the USAS software and of the USAS tagset for the analysis of Finnish. For these pilot parallel corpora, we decided to choose texts that deal with coffee, since the theme fell within the semantic areas of food and drink that we first started experimenting with in this project. Although we realized that this was a small specific domain and not representative of the whole taxonomy, it allowed us to investigate the plans on real data. The Finnish corpus was compiled from texts collected from the Internet[38], and the English corpus was produced by translating the Finnish corpus into English. Thereafter, the texts constituting the Finnish corpus were further edited to some extent in order to make them

[38] The texts were collected from http://www.kahvilasi.net in the year 2002. The website is no longer available, but some of the texts can now be found at http://www.helsinki.fi/kemia/opettaja/aineistot/kahvi/kartta.html.

lexically match the English corpus as perfectly as possible for our testing purposes. The resulting Finnish "coffee corpus" consisted of 2,063 words, and the parallel English "coffee corpus" consisted of 3,473 words. The difference between these numbers is due to the fact that Finnish as an agglutinative language predominately uses inflections instead of prepositions as is the case for English, and Finnish does not use articles (see section 2.5.1); as a result, Finnish sentences usually contain fewer words than their translations into English, and words in Finnish sentences are usually longer than in English sentences. Finally, both corpora were tagged grammatically and semantically. The Finnish corpus was tagged manually, whereas the English corpus was tagged using the EST, after which it was manually post-edited.

In the following phase, we compared the two parallel tagged corpora, and we were able to draw two significant conclusions from them. Firstly, since the languages are very different from each other, as became evident in section 2.5, it was obvious that we would have to implement some changes in the software to enable it to process the specific features of Finnish successfully. These changes will be discussed in sections 3.3.1 and 3.3.2. However, the second conclusion was that while the software clearly needed some modification, the semantic categories developed originally for the EST did not: they were found entirely suitable for the semantic categorization of objects and phenomena in Finnish as well. For example, I tagged the very "Finnish" concept *kiuas* (the stove which is used for heating the Finnish sauna) as H5/O4.6+ which is a combination of the semantic tags representing the categories "Furniture and Household Fittings" and "Temperature". The plus marker indicates a high temperature. The traditional Easter pudding *mämmi* as well as other typically Finnish dishes fall conveniently into the category F1 ("Food"). The shared semantic categories thus function as a type of a "meta-dictionary" or "lingua franca" between the languages. The experiences were similar a few years later when the equivalent semantic tagger for Russian

was developed (Mudraya et al., 2006, pp. 293–294). This may be partly due to the fact that there are a reasonable amount of similarities in the cultures of these three countries. Another probable reason is the fact that the USAS semantic categories are so general that they can be easily applicable across different cultures. Nevertheless, the semantic categories may well need some adjustment when they are to be applied to the analysis of languages in cultures which are very different from ours. Interesting findings were reported by Qian and Piao (2009) in relation to the development of a semantic annotation scheme for Chinese kinship terms. The work was based on modifying the USAS tagset. They noticed that the Chinese kinship system is quite different and much finer-grained than the English kinship system, and even if the USAS scheme was made finer-grained by subdividing the existing categories further, the scheme would not cover the type of distinctions which are made in Chinese. (Qian & Piao, 2009, pp. 189–191)

Archer et al. (2004, p. 823–824) point out in relation to the USAS category system that its purpose has been to provide a conception of the world that is as general as possible. As a consequence, some of the fine-grained distinctions made by other category systems can be lost. They illustrate this with the example of birds. The USAS category system does not have a specific category for birds, but birds as well as other animals are all grouped together in the category L2 ("Living Creatures Generally") which belongs in the top level category L ("Life and Living Things"). If a particular task requires, the category system can be expanded further by adding more subcategories, such as "Creatures of the Land", "Creatures of the Sea", and "Creatures of the Air", and the subcategory "Creatures of the Air" could be further expanded into subcategories, such as "Wild Birds" and "Domestic Birds". However, the classification of words into finer-grained categories might be problematic, since, for instance, birds which are considered to be wild by one culture may be considered pets by another culture. As was evident from section 2.3, the existing semantic ontologies vary a great deal as to the depth of

the hierarchy and the number of categories they include. Generally, however, the coarser-grained the category system is, the more applicable it is across different cultures.

The building of the parallel test and training corpora also marked the beginning of the semantic lexicon development. The words contained in the Finnish coffee corpus constituted the first entries in the Finnish single word lexicon which will be described in section 3.4.1.

3.3 Development of the Software Component

The specific grammatical features of Finnish that engendered the need to modify the software were rich morphology and productive use of compounding which were discussed in sections 2.5.1.1 and 2.5.1.2. Our solutions for addressing these issues will be described in the following two subsections. Furthermore, we decided to change the encoding system of the whole USAS framework. Issues connected to this will be discussed in section 3.3.3. The third specific grammatical feature of Finnish which I presented in section 2.5.1.3, relatively flexible word order, did not cause a need to modify the software. Instead, it affects the development of the Finnish MWE templates. I will discuss this in more detail in section 5.2.2.2, where I draft guidelines for writing templates which can reliably recognize different types of Finnish MWEs.

3.3.1 Modifications caused by rich morphology

The English and Finnish languages require different algorithms and tools for processing the same type of linguistic information. The processing of text in a semantic tagger starts from the retrieval of POS information that provides the basis for determining the semantic category of a word. In the EST, the CLAWS POS tagger (Garside & Smith, 1997, pp. 102–121) is used

for this purpose. In order to develop an equivalent semantic tagger for Finnish, we needed a Finnish counterpart POS tagger. For this purpose, we used a Finnish morpho-syntactic analyser and parser named TextMorfo[39].

TextMorfo includes several different tools that analyse Finnish text in various aspects. The most important of these tools are Morfo and DC Parser which were introduced in sections 2.2.3.1 and 2.2.3.2. Morfo analyses the morphological structure of Finnish words. It extracts morpho-syntactic information from words and returns the candidate basic forms with all the potential interpretations of the part of speech, inflections, and enclitic particles. This step is especially essential for a language with rich morphology, since it would be totally impossible to try to include all potential inflected forms of Finnish words combined with all potential enclitic particles in the semantic lexicons[40]. DC Parser, in turn, is a full dependency parser of Finnish which returns a "dependency tree", in other words, a structure that indicates the dependency relationships between words in the input sentence, such as predicates and objects. In addition, DC Parser recognizes and lumps together some frequently co-occurring multiword collocations which it processes as one unit; these will be examined in section 3.4.1. Thus, based on the candidate interpretations of the input word which the Morfo component has generated, the DC Parser component selects the correct interpretation in the given context. Finally, TextMorfo converts the output into a user-friendly list of disambiguated words. (Jukka-Pekka Juntunen, personal communication, April 10, 2008[41]) By way of illustration, TextMorfo generated the following output for the sentence *Ajoimmeko liian lujaa risteyksessä?* ("Did we drive too fast in the crossing?")[42]:

---

[39] Similarly, the Russian Semantic Tagger uses a Russian morpho-syntactic analyser named Mystem as the equivalent of the CLAWS POS tagger of English and the TextMorfo POS tagger and parser of Finnish (Mudraya et al., 2006, p. 5).

[40] This issue will be discussed in more detail in connection with the semantic lexicon development in section 3.4.

[41] Further details are unavailable, since TextMorfo is a commercial product.

[42] TextMorfo output was provided by J-P Juntunen from Kielikone.

liian (liian), category: Adverb, case: ; liian , Place: 2, CCat:
_QUESTION (?), category: Delimiter, case: ; _QUESTION , Place:
5, CCat:
risteys (risteyksessä), category: Noun, case: In; risteys , Place: 4,
SG CCat:
lujaa (lujaa), category: Adverb, case: ; lujaa , Place: 3, CCat:
Ajaa (Ajoimmeko), category: Verb, case: ; Ajaa , Place: 1, Imp Act Ind
P 1P ko CCat:
(null) ((null)), category: EndOfSentence, case: (null); (null) (null),
Place: (null), (null) (null) (null) (null) (null) (null) (null) (null)
CCat:(null)[43]

From the above output, we can conclude that the sentence in question consists of the constituents displayed in Table 2 below. Thus, *Ajoimmeko* is a verb in the past tense, active voice, indicative mood, and in the first person plural, and it ends with the enclitic particle *-ko* which indicates a direct question. The words *liian* and *lujaa* are adverbs. The word *risteyksessä* is a noun in the inessive singular. Finally, a question mark concludes the sentence.

[43] The order of the constituents in the TextMorfo output is determined by the dependency tree. The abbreviation CCat stands for compound category; there were no compounds in this example sentence.

Table 2

Breakdown of TextMorfo Output

Place	**Constituent**	**Translation**	**Grammatical information**
1	Ajoimmeko	Did we drive	Imp (verb in the past tense)
			Act (active voice)
			Ind (indicative mood)
			P (plural form)
			1P (first person)
			ko (enclitic particle indicating a question)
2	liian	too	adverb
3	lujaa	fast	adverb
4	risteyksessä	in the crossing	In (noun in inessive case)
			SG (singular form)
5	?	?	QUESTION (question mark)

In the course of the development process, we realized that the POS tagset of TextMorfo was not entirely sufficient for our purposes. TextMorfo uses the tags which are listed in Table 3 below.

Table 3

TextMorfo Tags

Abbreviation	e.g. *CD* ("CD"), *eKr.* ("BC")
Adjective	e.g. *epäitsekäs* ("unselfish"), *puolueeton* ("impartial")
Adverb	e.g. *kaukana* ("far"), *filosofisesti* ("philosophically")
Code	e.g. *b, e, Y*
Conjunction	e.g. *jos* ("if"), *kunnes* ("until")
Interjection	e.g. *aamen* ("amen"), *pahus* ("damn")
Noun	e.g. *keskeytys* ("interruption"), *jääkiekkoilija* ("ice hockey player")
Numeral	e.g. *ensimmäinen* ("the first"), *kolmetoista* ("thirteen")
Preposition	e.g. *ilman* ("without"), *yli* ("over")
Pronoun	e.g. *he* ("they"), *kumpikin* ("both")
Proper	e.g. Elina (female name), Aamuposti (name of a Finnish newspaper)
Verb	e.g. *ryöpätä* ("to parboil"), *kieltää* ("to deny")

The supplementary POS tag that we found necessary for the analysis of Finnish was "CompPart". The tag "CompPart" is used in the FST to mark the specific group of Finnish word forms presented in section 2.5.1.2 which appear solely as the first constituent in compounds and are never used independently. Examples of such words are: *aamiais* ("breakfast") as in *aamiaispöytä* ("breakfast table") and *kuolin* ("death") as in *kuolinaika* ("time of death"). Such words marked as "CompPart" differ from the basic form of the word

they have been derived from (for instance, the basic forms for the above example words are: *aamiainen* and *kuolema*) and do not represent any part of speech[44].

3.3.2 Modifications caused by productive use of compounding

The second need for modification of the software component was caused by compounding. As mentioned in section 2.5.1.2, compounding is a very productive means of word formation in Finnish. The number of possible compounds is infinite, so it would be totally impossible to collect all possible candidates. Attempting to include as many as possible would not be sensible either, since this would inevitably result in an uncontrollable lexicon size. Therefore, we decided to include only the most frequent compounds as well as lexicalized compounds in the single word lexicon. All other possible, less frequently used compounds of a more temporary nature are handled by a new component in the FST software named the "compound engine".

When text is fed into the FST and the program discovers a word that does not exist in the semantic lexical resources, it next checks if the word is possibly a compound consisting of two words. If this is discovered to be the case, the FST assigns the relevant semantic tag/tags for both constituents of the compound separately. At the final stage, the semantic tags of the compound constituents are combined automatically and separated by a slash. The resulting semantic tags resemble the slash tags which were discussed in section 2.4.1.1. For instance, the compound engine generated the following output for the compound *pernatulehdus* ("splenitis"; literally "spleen inflammation") which is not included in the Finnish single word lexicon:

[44] By comparison, such words which are never used independently and which appear as the final constituent in a compound, for example *mielinen* ("minded") as in *uudistusmielinen* ("reformist"; literally "reform-minded"), do represent a part of speech and thus have been assigned the relevant POS tag (in this case adjective).

<w pos="Noun/Noun" mwe="com" sem="B2-/B1" lem="tulehdus/perna">pernatulehdus</w>

As seen above, the second constituent of the compound (here *tulehdus* ("inflammation")) is placed first. The reason for this is that the second constituent is usually more significant in terms of the meaning of the compound than the first constituent. Consequently, the first constituent of the compound (here *perna* ("spleen")) that modifies the second constituent is placed second. Thus, the word *pernatulehdus* is tagged as B2-/B1 (the category "Health and Disease", with the minus marker indicating ill health / the category "Anatomy and Physiology"[45]). The abbreviation "mwe="com"" in the output indicates that the tag has been produced by the compound engine[46].

If the compound constituents are ambiguous and have been assigned more than one semantic tag, the compound engine generates all possible combinations of the semantic tags of the constituents. For example, the compound *talvikenkä* ("winter shoe") would receive the following tags:

<w pos="Noun/Noun" mwe="com" sem="B5/T1.3 O2/L2/T1.3" lem="kenkä/talvi">talvikenkä</w>

The noun *talvi* ("winter") has been assigned one semantic tag, T1.3 ("Time: Period"). In comparison, the noun *kenkä* ("shoe") has been assigned two semantic tags: B5, which represents the category "Clothes and Personal Belongings", and O2/L2 which indicates a horseshoe (this slash tag denotes that the word in question belongs both to the category "Objects Generally" and to the category "Living Creatures Generally"). Thus, the compound

[45] The USAS semantic tagset was presented in section 2.4.1.1, and the discussion continues in section 3.4. Where the semantic tags are not explained or clear from context, the necessary definitions and examples can be found in Appendix C. Additionally, a list of all semantic categories can be found in Appendix A (in English) and in Appendix B (in Finnish).

[46] The output of the FST will be presented in more detail in section 3.5.

engine generates two different combinations of these tags: B5/T1.3 and O2/L2/T1.3. In this case, the first combination is the correct one.

Note that above I wrote about the compound engine that "it next checks if the word is possibly a compound consisting of two words". Most often Finnish compounds consist of two words, but there is also a large number of compounds which consist of three or more words, as became evident in section 2.5.1.2. However, the compound engine splits a compound only into two constituents. Thus, it regards as one constituent the final word in the compound which it recognizes and as the other constituent all that comes before it. Examples of this are the compounds *kallo=vamma=spesialisti*[47] ("skull injury specialist") and *kallo=vamma=spesialisti=ryhmä* ("group of skull injury specialists") which the compound engine would regard as compounds consisting of the constituents *kallovamma* and *spesialisti* and *kallovammaspesialisti* and *ryhmä*. Such a result is not wholly satisfying, but, nevertheless, I believe that this approach is the wisest, since if all possible combinations of the semantic tags of the compound constituents were generated, the end result might become more confusing than helpful. In addition, as I noted earlier, the final constituent of a compound is usually the most relevant constituent for the meaning of the compound.

3.3.3 Other modifications to the software

During the early stages of the development process, we also came to the conclusion that the encoding system of the software needed to be changed. Although most of the letters of the Finnish alphabet are the same as in the English alphabet, there are three additional characters in Finnish whose values fall outside the basic ASCII code set that the EST used to employ. These characters are: *å*, *ä*, and *ö*. To address this issue, we adopted the Unicode (UTF-8)

[47] The symbol "=" is used in this subsection to mark boundaries between the compound constituents.

encoding scheme for the whole USAS framework. This freed us from a complex conversion problem in encoding. Moreover, this type of preparation of the core components also made it easier to extend the framework to other languages. Indeed, the new semantic taggers in the USAS framework (see section 2.4.2) were also encoded using Unicode.

3.3.4 Program Architecture

The following figure illustrates the architecture of the FST. The first phase is grammatical analysis which is carried out by the TextMorfo component. Grammatical analysis provides the basis for semantic analysis occurring in the second phase.

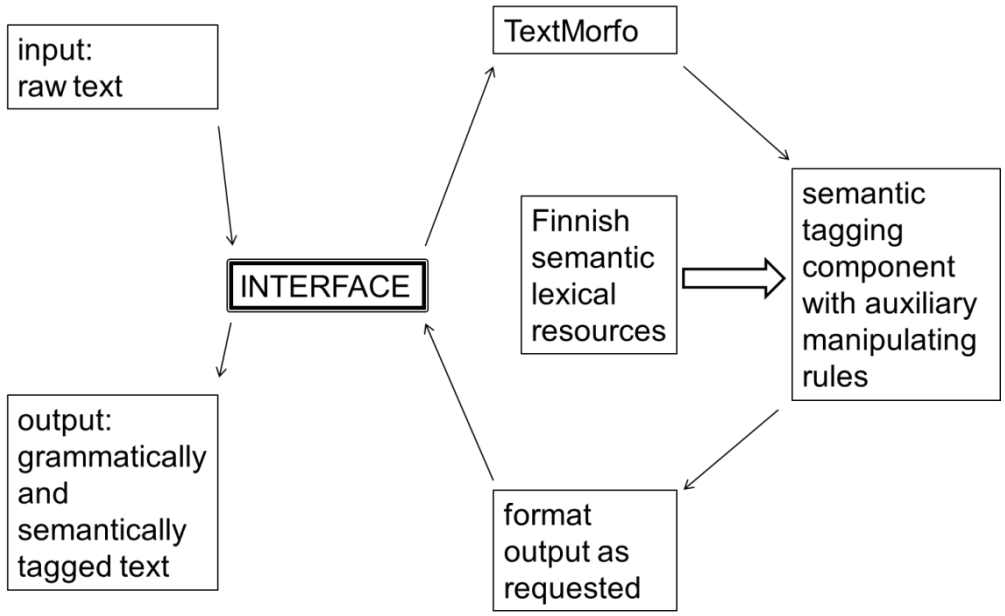

Figure 3. Architecture of the Finnish Semantic Tagger

The FST is parallel to the EST in terms of structure (see Figure 1 in section 2.4.1.4), with the exception that the EST employs a POS tagger (CLAWS) and a lemmatiser separately, whereas TextMorfo contains both these tools. A parallel architecture of the two semantic

taggers naturally requires compatible semantic lexical resources. The development of the semantic lexicons for Finnish will be discussed in the following subsection.

3.4 Development of the Semantic Lexical Resources

The creation of the semantic lexical resources has been by far the most laborious and time-consuming task in the FST development. At the beginning of the Benedict project, we envisaged that we might be able to make use of some automated methods, such as producing a machine translation of the entries in the English lexicon to Finnish and then editing the resulting list. However, quite soon we decided that these types of "conversion table approaches" were not feasible, and the development was undertaken from scratch[48]. Despite the large amount of work involved, I decided to carry out the lexicon development by myself, even though I was offered a possibility to enlist a student or two for help. The reason for this was that by doing so I hoped that the end result would be as coherent as possible. The English semantic lexicons have been developed during two decades by various people. Even though the tagset used was the same, people tend to perceive things somewhat differently, which has caused slight incoherence when assigning semantic tags to words. For example, the singular form of the noun "trance" is tagged as X1 ("Psychological States, Actions, and Processes: General"), whereas the plural form of the same noun, "trances", is tagged as X2 ("Mental Actions, and Processes"). In addition, the noun "hunger" has been tagged as F1-/B1 ("Food", with the minus marker indicating the lack of it / "Anatomy and Physiology"), whereas the

[48] Interestingly, during the past few years, automated methods have been applied successfully. Equivalent semantic lexicons have now been developed for Chinese, Italian, and Brazilian Portuguese by bootstrapping new semantic lexical resources via automatically translating the existing English semantic lexicons into these languages (Piao et al., 2015). However, this method requires appropriate, high-quality bilingual dictionaries or lexicons. To the best of my knowledge, such resources are not yet freely available for Finnish. The semantic lexical resources for the latest semantic taggers in the USAS framework have been created utilizing automatic translation and crowdsourcing, after which they have been manually cleaned and improved (Piao et al., 2016).

noun "thirst" has been tagged as B1/F2 ("Anatomy and Physiology" / "Drinks"). The UCREL team carried out an experiment to measure the inter-rater reliability for semantic annotation in a subsection of the EST lexicon. For this purpose, they utilized a crowdsourcing methodology. They engaged multiple people to perform the tagging for a part of the semantic lexical resources for the EST to measure how general native users of English are able to replicate the categorisation. A similar experiment for the Finnish semantic lexical resources will be reported in section 4.5, and the results for English will be reported in connection with it.

As is the case with the English model, the Finnish semantic lexical resources also contain two separate lexicons: one consisting of single words and one consisting of MWEs. An adaptation of Introduction to the USAS Category System (Archer et al., 2002) is included in this thesis as Appendix C. This appendix displays the top level semantic categories as well as all their subcategories with many prototypical Finnish language examples of both single words and MWEs.

I have used Microsoft Excel in the lexicon construction, and I have found it a very useful tool for this purpose. There are also various other tools such as XML editors and databases, for example, Protégé[49], which allow maintenance of lexical resources.

When creating the Finnish semantic lexical resources, I have followed the principles and practices used in the development of the English semantic lexical resources as closely as possible. Similarly to the English semantic lexicons, the aim in the development of the Finnish counterparts as well has been to build them primarily into a resource representing general language. General language in this context could be defined as the type of language which a native speaker can understand without any special mastery. Such language can be found, for example, in newspaper text. However, there is one significant difference between

[49] For more information, see http://protege.stanford.edu/.

the English and Finnish semantic lexicons in terms of structure: the English semantic lexicons contain both basic forms as well as their inflectional variants, whereas the Finnish counterparts consist of basic forms only. This is due to the fact that at the initial phase of the EST construction, the developers had no reliable automatic English lemmatiser available, and therefore they had to include also the inflected forms in the semantic lexicons. This has not created any problems, however, since the number of inflected forms in English is limited[50]. For Finnish this approach would have been totally impossible due to its highly inflectional and agglutinative nature. If all inflectional variants were included in the Finnish semantic lexicons, combined with all possible enclitic particles, this would result in an unmanageable lexicon size. Thus, the FST uses the Finnish morpho-syntactic analyser and parser TextMorfo described in section 3.3.1 to reduce Finnish words to basic forms first, and only after that are these basic forms compared to the semantic lexicon entries which are also in basic form.

Similarly to the English semantic lexical resources, the Finnish semantic lexical resources also employ the USAS semantic tagset which was introduced in section 2.4.1.1. Hence, the semantic tags in the Finnish lexicons as well are composed of an upper-case letter indicating the top level semantic category (e.g. T ("Time")), a digit indicating a first subdivision of the field (e.g. T1 ("Time")), and optionally, a decimal point followed by a further digit (e.g. T1.1 ("Time: General")) or two decimal points and two digits (e.g. T1.1.1 ("Time: General: Past")) which indicate a finer subdivision in the field. The depth of the semantic hierarchical structure is limited to a maximum of three layers, since this has been found to be the most feasible approach (Piao et al., 2005a). In addition to the upper-case letters and digits, the Finnish semantic lexical resources also contain two optional markers that can be attached at the end of a semantic tag. These are "f" indicating females and "m" indicating males[51]. By way of illustration, the noun *naishenkilö* ("female person") is tagged as S2.1f and the noun *hieho*

[50] A lemmatiser was included in the EST only during the Benedict project.

[51] These markers were originally created for the purpose of experiments with anaphor resolution.

("heifer") as L2f, whereas the noun *poikamies* ("bachelor") is tagged as S2.2m and the noun *ori* ("stallion") as L2m. If a word can be used for both sexes, for example *kirjailija* ("author") or *afrikkalainen* (the noun "African"), the Finnish semantic lexicons do not use both these markers, as is the procedure in the English semantic lexicons, since they were not found necessary for our purposes. Thus, these nouns have received the semantic tags Q4.1/S2 and Z2/S2 respectively. Furthermore, the markers "%" and "@" (rarity markers), "c" (potential antecedents of conceptual anaphors[52]), "n" (neuter), and "i" (semantic idiom) used in the EST (Archer et al., 2002, p. 2) were found unnecessary in the Finnish semantic lexicons for the time being. However, if need be, these can be added later.

Moreover, one, two, or three pluses or minuses can be attached to semantic tags to indicate antonymous pairs or a positive or a negative position on a semantic scale. By way of illustration, *kohtelias* ("polite") has received the tag S1.2.4+, whereas *epäkohtelias* ("impolite") has received the tag S1.2.4-, and *hyödyllinen* ("useful") has been tagged as A1.5.2+, whereas *hyödytön* ("useless") has been tagged as A1.5.2-. Two pluses indicate an increased amount of something. For example, *lisä* ("addition") has been assigned the semantic tag N5++, and *jatkuvasti* ("continuously") the semantic tag T2++. Two minuses, in turn, indicate the opposite, as is, for instance, in the case for the noun *huonommuus* ("inferiority") A5.1--. Three pluses or minuses indicate the upper and lower extremes, for instance, in the case of the words *identtinen* ("identical") A6.1+++, *ikuisuus* ("eternity") T2+++, *jättikokoinen* ("gigantic") N3.2+++, *ainutlaatuinen* ("unique") N5---, *rutiköyhä* ("poor as a church mouse") I1.1---, and *äskettäin* ("recently") T3---. Moreover, comparative and superlative forms of adjectives and adverbs are expressed by pluses and minuses. In the English semantic lexical resources, the comparatives and superlatives are included as individual entries. For example, in the English single word lexicon, the adjective "fast" has received the semantic tag N3.8+,

[52] This was used to mark candidate pronouns which could possibly be linked to their related referents.

the comparative form is tagged as N3.8++, and the superlative form as N3.8+++. However, since the TextMorfo component automatically reduces comparatives and superlatives into basic forms before passing the output on to the semantic tagging component, we decided to include only the basic forms of adjectives and adverbs in the Finnish semantic lexical resources. If the FST is further developed, a new component will need to be built which then adds the relevant pluses and minuses into the semantic tags of those adjectives and adverbs which appear in comparative or superlative form[53].

Sometimes words do not fall neatly into predefined semantic categories, but they can belong in two or even three categories. In such cases, the semantic tags representing these categories are combined with a slash into one single semantic tag; these are referred to as slash tags. Slash tags were introduced in section 2.4.1.1 in connection with the USAS semantic tagset. By way of illustration, the verb *varastaa* ("to steal") as well as the corresponding noun *varastaminen* have been tagged as G2.1-/A9+, in which the semantic tag G2.1- signifies something illegal and the semantic tag A9+ signifies getting and possession. Hence, the semantic tag G2.1-/A9+ means that something is taken possession of illegally. The following verbs and their derivations, among others, have also received this same tag in the single word lexicon: *kaapata* ("to hijack"), *kidnapata* ("to kidnap"), as well as *anastaa*, *kähveltää*, *näpistellä*, and *varastella*, all of which denote stealing. The semantic tag which indicates the actor for these verbs can, in turn, be formed by adding a third semantic tag, S2, indicating a person: thus, for instance, the nouns *anastaja*, *kaappari*, *kidnappaaja***,** and *varas* have all been assigned the semantic tag G2.1-/A9+/S2. Some other examples of slash tags in the Finnish single word lexicon include:

[53] Since this function has not yet been available, some comparative and superlative forms of adjectives and adverbs which were necessary for our testing purposes were added into the semantic lexical resources.

algerialainen	Noun	Z2/S2 ("Algerian")
anniskella	Verb	I2.2/F2 ("to sell alcohol")
arvojärjestys	Noun	N4/S7.1 ("ranking order")
helppokäyttöinen	Adjective	A1.5.1/A12+ ("easy-to-use")
herätyskokous	Noun	S9/S1.1.3+ ("revivalist meeting")
kaikkialla	Adverb	M6/N5.1+ ("everywhere")
likaantua	Noun	O4.2-/A2.1 ("to get dirty")
lomauttaa	Verb	I3.1-/T1.3 ("to lay off")
onnenhetki	Noun	T1.2/E4.1+ ("moment of happiness")
pedanttisesti	Adverb	A4.2/N5.2+ ("pedantically")
Volkswagen	Proper	Z3/M3

The general practice in the English semantic lexicons has been to place first the semantic tag which is the most relevant for the meaning. In my work, I have followed this practice in order to have uniform lexicons. There are, however, some cases in which a different ordering has been applied in both the English and Finnish semantic lexical resources. An example of this is the semantic tag S2 which indicates people. This semantic tag is always placed last, for instance, as can be seen in the case of *algerialainen* among the above examples. Another example is the semantic tag Z3 which indicates proper names other than personal or

geographical names. Sometimes, the semantic tag Z3 has been used alone, but sometimes it has been complemented by another semantic tag which indicates the field in this the proper name belongs. An example of this is *Volkswagen* in the above list. In such a case, the semantic tag Z3 is always placed first, and the complementary semantic tag is placed second. As many as 2,996 tag types, in other words, different combinations of letters, digits, pluses, minuses, and, slashes, appear currently in the Finnish semantic lexical resources.

Choosing the correct semantic tag for the lexicon entries has often been a very complex task that has involved consulting various types of reference material. When I have had a word to tag before me, I have generally first looked into a monolingual dictionary of Finnish to identify all possible senses. The dictionary which I have used over the past few years is the electronic version of the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish") which is the most comprehensive modern monolingual dictionary of the Finnish language[54]. Before its publication, I used the electronic version of the *Gummeruksen uusi suomen kielen sanakirja* ("The Gummerus New Dictionary of the Finnish Language"; Nurmi, 1998). By comparison, the linguists in charge of the development of the English semantic lexical resources used a number of different knowledge sources and tools in the task of selecting relevant semantic tags for the new lexicon entries. These included, for example, large electronic dictionaries, such as the *Collins English Dictionary*, and concordance lines from representative corpora, such as the BNC (Piao et al., 2005a). Unfortunately, there have been no such large, freely accessible reference corpora available for Finnish.

Furthermore, I have also often cross-checked the English semantic lexicons to find out which semantic tag the UCREL team has chosen for the sense in question, in order to make the Finnish semantic lexicons as compatible with them as possible. There have been some cases, however, in which I have decided on a slightly different interpretation. For example,

[54] This dictionary is updated constantly. The last update was carried out February 29, 2016.

the word *geisha* exists in both languages. In the English single word lexicon, it has been tagged as S3.2/S2.1f which is a combination of the category "Relationship: Intimate/Sexual" and of the category "People: Female". For the Finnish single word lexicon, however, I chose the semantic tag K1/S2.1f ("Entertainment Generally" / "People: Female", because that matched better the definition in the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish"), and I myself considered this semantic tag a more relevant choice. When necessary, I have also consulted the online versions of *MOT Englanti*, MOT *Collins English Dictionary*, *Merriam Webster Dictionary*, and *Collins English Dictionary*, as well as Google.

It is worth noting that the number of semantic tags in the semantic lexicon entries is not necessarily the same as the number of senses in the reference sources which I have used. The decisions have not been based on large-scale analysis but on my intuition and on what I have considered practical solutions for this type of semantic lexical resources. Firstly, I have left out senses which I considered infrequent, archaic, dialectal, or representing jargon and thus not relevant additions to such lexicons. Secondly, as is the case with the English semantic lexicons, the sense distinction in the Finnish semantic lexicons is more coarse-grained than in large dictionaries, such as my reference sources, and thus I have grouped together some close senses which can be considered to belong in the same semantic field. By way of illustration, the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish") lists the following senses for the noun *koulu***.** To be concise, I have included here only the first one of the usage examples for each sense.

1. (lower) educational institution

   Maamme koulut. ("The schools in our country.")
2. in some noun compounds of tuition in course format

   Pyhäkoulu. ("Sunday school.")

3. school building, school house, school premises

   Koulu on torin laidassa. ("The school is located next to the market place.")

4. teaching and studying at school, schoolwork; school attendance

   Koulu alkaa, päättyy. ("School starts, ends.")

5. school system

   Koulun uudistaminen. ("School reform.")

6. especially in music

   a. systematic course; school book / series of school books containing such a course

      Kitarakoulu. ("Guitar school.")

   b. schooling; proficiency produced by schooling

      Viulistin mainio koulu kävi ilmi jo ensi tahdeista. ("The excellent schooling of the violinist was apparent from the first notes onwards.")

7. school of thought

   Rafaelin koulu. ("The School of Raphael").

By comparison, I have assigned *koulu* the following two semantic tags in the Finnish single word lexicon:

1. P1 ("Education in General")
2. S5+ ("Groups and Affiliation", with the plus marker indicating belonging in a group)

The semantic tag P1 covers the senses number 1, 2, 3, 4, 5, and 6a, whereas the semantic tag S5+ covers the sense number 7. I have disregarded the sense number 6b altogether when compiling the lexicon entry, since that sense is very infrequent and thus not a relevant addition to such a general language resource. Nevertheless, if a particular task requires, it is

possible to add further levels of subdivision and thus enable a much more detailed analysis and description including even more distinctive features. Alternatively, the semantic tags can be made more specific by using slash tags. I will return to this topic of granularity in section 5.3 in which I draft guidelines for tailoring the Finnish semantic lexical resources for domain-specific applications.

The simplest type of a semantic lexicon entry is that for an unambiguous word or MWE, that is, the word or the MWE has been assigned only one semantic tag. In case a word or a MWE has been assigned more senses than one in the semantic lexicons, the semantic tags representing the senses have been organized in perceived frequency order. This order is based on information received from the above-mentioned Finnish reference sources, my native-language intuition, and my work experience as a lexicographer. To the best of my knowledge, there is no dictionary of the Finnish language in which information about the frequency of the different senses of ambiguous words would be systematically available. According to Eija-Riitta Grönros (personal communication, August 3, 2012), the editor of the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish"), they arranged the senses in the entries for ambiguous words according to various principles. The primary aim was to place the most frequent sense first. However, at times, the most logical solution was to place the concrete sense before the figurative sense, even though the former is less frequently used than the latter. Additionally, if a dictionary entry contains many meaning groups, the compilers considered the most practical option to place the closely connected senses one after another, even though one of them might be less frequently used than the ones following it. And finally, in some cases they had noticed later on that the perceived frequency order was after all not correct. According to Grönros, this was due to the fact that the work was based on an earlier large monolingual dictionary, the *Suomen kielen perussanakirja* ("Basic Dictionary of the Finnish Language"), which was compiled before the 1990s, and some of the entries had not

been edited since, so the frequency order of the senses had changed over time, because language changes constantly. All things considered, I believe that the order of senses in the Finnish semantic lexicons should relatively well reflect the general situation in ordinary Finnish language usage. An evalution of this will be presented in section 4.4.3.2.1.

Nevertheless, it must be noted that the order of the senses in a semantic lexicon entry is not a crucial issue. Instead, it is more important to develop effective disambiguation mechanisms which can enable the FST to recognize the correct sense in a given context and select the relevant semantic tag for it from the semantic lexicon entry. Various types of solutions to address these issues will be presented in section 4.4.3 in connection with the analysis of the errors which occurred in the application-based evaluation. Furthermore, the order of senses in a semantic lexicon is not essential in many applications, such as in an information retrieval setting where only certain features in the text need to be recognized. In such a case, it is not necessary to disambiguate between the senses of ambiguous words in a given context, but it is sufficient that the semantic tag for the relevant sense is included among the semantic tags in the lexicon entry.

The FST is case-sensitive and can differentiate between general and proper nouns. By way of illustration, the single word lexicon contains the following entries:

terttu	Noun	L3
Terttu	Proper	Z1f

The former word signifies a bunch (such as a bunch of grapes), whereas the latter is a female name. Similarly, the noun *kuusi* denotes a spruce tree, and *Kuusi* is a Finnish family name. In addition, *kuusi* can also be a numeral meaning the number six.

kuusi	Noun	L3

kuusi	Numeral	N1
Kuusi	Proper	Z1

The words *Aurinko*, *Kuu* and *Maa* ("the Sun", "the Moon" and "the Earth") are capitalized when they denote the proper names of heavenly bodies. When used as common nouns, however, they are started with a lower-case letter. Thus, the single word lexicon contains, for instance, the following entries:

kuu	Noun	W1	T1.3[55]
Kuu	Proper	Z3/W1	

Sometimes both a lower-case variant (common noun) and upper-case variant (proper name) to refer to the same concept. In such a case, both variants have been included in the single word lexicon:

internet	Noun	Y2
Internet	Proper	Z3/Y2

In the following section, I take a closer look at the Finnish single word lexicon.

3.4.1 Single word lexicon

I have carried out the development process of the single word lexicon in various phases. The grammatically and semantically tagged word list that I created from the manually tagged test and training corpus, which was described in section 3.2, marked the beginning of the

[55] The noun *kuu* has also the sense "month".

work. Next, I assigned both POS and semantic tags to a list generated by Kielikone consisting of 5,000 most frequently used words in a corpus which they had compiled of newspaper texts published in *Helsingin Sanomat*, the biggest daily newspaper in Finland. This was done to focus on increasing coverage for corpora as efficiently as possible (the evaluation of the lexical coverage of the single word lexicon will be presented in section 4.3). Thereafter, I collected word lists of different fields, such as plants, animals, languages, foods, drinks, currencies, as well as personal, geographical, and other proper names from various freely available Internet sources. In addition, I intuitively listed words belonging to many other fields such as weekdays, months, colours, and body parts. I read through all the words in each of the lists, on one hand deleting the least frequently occurring words that I found unnecessary for such a general language lexicon and, on the other hand, adding new words belonging to the same meaning groups that I considered worthwhile additions. Following this procedure, I used TextMorfo to assign POS tags for these words, after which I manually added the relevant semantic tag for each word in the list, for example, L2 for every word in the list of animals and T1.3 for all weekdays and months. I then read through the lists again carefully to find words that have any other senses. An example of such a case is the noun *hiiri* that has two senses: 1) a small furry animal (L2) and 2) a pointing device for the computer (Y2). Similarly, the noun *sammakko* most often refers to the animal frog (L2), but it can also refer to the way children swim the breaststroke (M4), or to a mistake (A5.3-). The resulting single word lexicons entries thus are:

hiiri	Noun	L2	Y2	
sammakko	Noun	L2	M4	A5.3-

When in this way I had managed to construct a "seed lexicon" of approximately 15,000 entries, I saved it into the software component of the FST, and thus the "working prototype" of the FST was ready.

Thereafter, I have been collecting candidates for new lexicon entries in the following way. I have fed different types of newspaper texts, articles on different fields, as well as online fiction and non-fiction into the FST. I have collected these from various sources to be able to provide a coverage as wide as possible. I have found current newspaper texts particularly beneficial, because they offer plenty of topical vocabulary and proper names as well as words which have recently entered the Finnish language. When text is entered, the FST assigns both POS and semantic tags for each word. In most cases, the words have been included in the TextMorfo lexicons, and as a result, the POS tagging component based on TextMorfo has recognized the part of speech of these words and has been able to assign the correct POS tag to them. However, if a word is not yet included in the semantic lexical resources, the semantic tagging component does not recognize such a word, but it assigns the word the semantic tag Z99 which represents the category "Unmatched". From the tagged output, I have then sorted out all these instances of words tagged as Z99, from the resulting list I have deleted the infrequent and misspelt words, and, thereafter, I have assigned semantic tags to the remaining words which I have considered valuable additions to the single word lexicon, consulting the reference sources which I mentioned in the previous section. Periodically, I have saved the latest version in the software component. In this way, I have incrementally built the single word lexicon into a database containing 45,781 entries, all both grammatically and semantically tagged, which, based on the method described above, will include the core lexicon of the Finnish language.

The creation of the English single word lexicon was a relatively similar process. The initial version was created by utilizing information which was contained in the lexical

resources of the CLAWS POS tagger. Subsequently, new entry candidates were collected from spoken and written corpora with similar methods as I have used in the development of the Finnish counterpart, in other words, by collecting unmatched words tagged as Z99 and including as new lexicon entries the words which were considered valuable additions (Piao et al., 2005a; Paul Rayson, personal communication, November 23, 2011).

The Finnish single word lexicon differs slightly from its English counterpart, not only because of the absence of inflectional variants, which was discussed in the previous section, but also because of the fact that some frequently co-occurring MWEs have been included in the single word lexicon and not in the MWE lexicon. These are fixed expressions in which the constituent words cannot be inflected, no enclitic particles are used[56], and where no embedded elements[57] are allowed between the constituents. In principle, all entries that consist of two or more words with intervening spaces between them do belong in the MWE lexicon, just as in the English equivalent, but since TextMorfo in the POS tagging phase processes some fixed expressions as single units and then assigns a POS tag to the entire expression[58], it was

[56] In principle, a creative mind would find it possible to add at least some enclitic particles to the end of nearly every word. However, here I concentrate on at least relatively frequently appearing formations and ignore cases which are in principle possible but appear very marginally.

[57] An embedded element is an item which can intervene in a discontinuous MWE. A typical example of such an element is a pronoun, noun, adverb, or proper name which is embedded in a verb phrase, for example:

annoin hänelle lopputilin ("I sacked him".; literally "I gave him the pay-off")
annoin Matille lopputilin ("I sacked Matti"; literally "I gave Matti the pay-off")

A MWE can contain more embedded elements than one, for example:

annoin tänään sille laiskalle Matille lopputilin ("I sacked that lazy Matti today"; literally "I gave today that lazy Matti the pay-off").

Embeddings were mentioned briefly in connection with the EST in section 2.4.1.2, and they will be examined in more detail in section 5.2.2.2 in which I draft guidelines for writing templates for Finnish MWEs.

[58] These resemble the ditto tags which are used in CLAWS (University Centre for Computer Corpus Research on Language (n.d.-d)).

practical to treat these as single units in the semantic tagging component as well. For this reason, they are included in the single word lexicon, and the spaces between the constituents are replaced by underscores. The single word lexicon currently contains 764 such entries, for example:

aamusta_iltaan	Adverb	T1.3+ ("from morning to evening")
herranen_aika	Adverb	Z4 ("good heavens")
heti_kun	Conjunction	Z5 ("as soon as")
hyvää_huomenta	Interjection	Z4 ("good morning")
joka_ikinen	Pronoun	Z8 ("every single one")
kuin_kaksi_marjaa	Adjective	A6.1+ ("like two peas in a pod")
olipa_kerran	Verb	Z4 ("once upon a time")

The following is a sample from the Finnish single word lexicon:

hauraasti	Adverb	O4.1	S1.2.5-
hauras	Adjective	O4.1	S1.2.5-
haurastua	Verb	O4.1/A2.1	S1.2.5-/A2.1
haurastuminen	Noun	O4.1/A2.1	S1.2.5-/A2.1
haurastuttaa	Verb	O4.1/A2.2	S1.2.5-/A2.2
haurastuttaminen	Noun	O4.1/A2.2	S1.2.5-/A2.2

hauraus	Noun	O4.1	S1.2.5-	
Hausjärvi	Proper	Z2		
hauska	Adjective	E4.1+	S1.2.1+	
hauskasti	Adverb	E4.1+	S1.2.1+	
hauskuus	Noun	E4.1+	S1.2.1+	
hauta	Noun	M7/L1-	W3	
hautaaminen	Noun	L1-/A1.1.1	A10-	
hautajais	CompPart	L1-/S1.1.1		
hautajaiset	Noun	L1-/S1.1.1		
Hautala	Proper	Z1		
hautamuistomerkki	Noun	L1-/C1		
hautaus	Noun	L1-/A1.1.1		
hautausmaa	Noun	M7/L1-		
hautautua	Verb	A10-	X5.2+	
hautautuminen	Noun	A10-	X5.2+	
hautoa	Verb	O4.6+	L2	B3
	X2.1			
hautominen	Noun	O4.6+	L2	B3
	X2.1			
hautua	Verb	F1	O4.6+	X2.1
hautuminen	Noun	F1	O4.6+	X2.1
hauva	Noun	L2		
Havaiji	Proper	Z2		
havaijilainen	Adjective	Z2		
havaijilainen	Noun	Z2/S2		
havaijilaispaita	Noun	B5		
havaijipaita	Noun	B5		
havainnoida	Verb	X3		
havainnoija	Noun	X3/S2		
havainnoiminen	Noun	X3		

havainnointi	Noun	X3
havainnollinen	Adjective	A12+

From this sample we learn, for instance, the following facts. The adjective *hauras* and the corresponding adverb *hauraasti* have been assigned two senses: physically fragile and figuratively fragile. Thus, both are tagged as O4.1 S1.2.5-. The category O4.1 is "General Appearance and Physical Properties", whereas the category S1.2.5 is "Toughness: Strong/Weak" which is complemented by a minus marker to indicate a negative position on the semantic scale. The verb *haurastua* means "to become fragile", so the semantic tags O4.1 and S1.2.5- are complemented by another semantic tag which represents the category "Affect: Modify, Change" (A2.1). Another example of a slash tag is the entry for the verb *haurastuttaa* ("to cause something to become fragile"). The category A2.2 ("Affect: Cause/Connected") indicates a causal relationship, so the tags necessary are O4.1/A2.2 and S1.2.5-/A2.2. The words *haurastuminen* and *haurastuttaminen* are nouns which are derived from these verbs and indicate "the act of..." *Hausjärvi* (a municipality in Finland) and *Havaiji* ("Hawaii") are geographical names (Z2). *Hauska* means "enjoyable" (E4.1+) or "personable" (S1.2.1+), so both these senses represent the positive side of the semantic scale in their respective categories. The adverb *hauskasti* and the noun *hauskuus*, in turn, are its derivations. *Hauta* ("grave") and *hautausmaa* ("graveyard") are tagged as combinations of the categories M7 ("Places") and L1 ("Life and Living Things"), and since it is deceased people at issue here, a minus marker is attached to the semantic tag L1. *Hauta* has also a second sense "trench", for which the relevant semantic tag is W3 ("Geographical Terms"). *Hautajaiset* ("funeral") is tagged as L1-/S1.1.1, in which the semantic tag S1.1.1 stands for the category "Social Actions, States, & Processes: General". *Hautala* is a Finnish personal name (Z1). In the case of the noun *hautamuistomerkki* ("sepulchral monument"), the semantic tag L1- is complemented by the semantic tag C1 ("Arts and Crafts"). The verb *hautautua* and the

derived noun *hautautuminen* mean "being covered" (A10-) or "immersing oneself in something" (X5.2+). The verb *hautoa* as well as the derived noun *hautominen* have been assigned a total of four different senses. These are: 1) warming something (O4.6+), 2) incubating eggs (L2), 3) bathing related to medical treatment (B3), and 4) pondering (X2.1).

The distribution of different POS categories in the single word lexicon is shown in Table 4 below. Nouns are by far the largest group constituting 57.70% of the entries. The second largest group is that of proper nouns (17.28%), followed by three other substantial groups: adjectives (7.35%), verbs (7.21%), and adverbs (6.98%). The remaining groups are notably smaller.

Table 4

Distribution of Part-of-Speech Categories

in the Single Word Lexicon of the Finnish Semantic Tagger

POS categories	**Entries (types)**	**%**
Abbreviation	381	0.83
Adjective	3,366	7.35
Adverb	3,194	6.98
Code	17	0.04
CompPart	606	1.32
Conjunction	116	0.25
Interjection	85	0.19
Noun	26,417	57.70
Numeral	91	0.20
Preposition	223	0.49
Pronoun	73	0.16
Proper	7,913	17.28
Verb	3,299	7.21
Total	45,781	100.00

It would be interesting to be able compare these figures to the distribution of POS categories in a general dictionary of Finnish. Unfortunately, such a dictionary in which POS information would be systematically included does not exist. However, Eija Riitta Grönros, the editor-in-chief of the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish"), was able to offer some help (personal communication, August 26, 2008). The edition published in 2008 contains in all almost 100,000 entries, and its electronic version also holds

the separate *The Dictionary of Finnish Place Names* which contains approximately 21,000 place names. The POS categories differ slightly from those used by TextMorfo and thus by the FST, but the POS categories which are shared are: "Adjective", "Noun", "Numeral", "Pronoun", and "Verb". According to their rough estimate, 72% of the entries were classified as nouns, 10% as adjectives, 10% as verbs, and less than one per cent as numerals and pronouns. As is evident from Table 5 below, if this estimated distribution of POS categories is compared to the distribution of POS categories found in the Finnish single word lexicon, the figures are actually quite similar, once the categories "Proper" and "CompPart", which can be expected not to exist in a general dictionary, have been excluded.

Table 5

Distribution of the Shared Part-of-Speech Categories in the Kielitoimiston sanakirja (KS) and in the Single Word Lexicon of the Finnish Semantic Tagger (FST)

Shared POS categories	**KS%**	**FST%**
Adjective	10	9.03
Noun	72	70.90
Numeral	<1	0.24
Pronoun	<1	0.20
Verb	10	8.85

In addition, I compared the distribution of POS categories in the Finnish single word lexicon to the distribution of POS categories in a list of 9,996 words found to be the most common in Finnish newspaper texts (Kielipankki, n.d.). This frequency list was created by CSC (IT Center for Science[59]) in 2004, and the source material consisted of 43,999,826 words of newspaper text. Even though a comparison to a frequency list is not as relevant as a

[59] For more information, see https://www.csc.fi/home.

comparison to a dictionary, the fact that TextMorfo had been used for the analysis of this newspaper corpus (Sami Salonen, personal communication, January 2, 2013) made the case quite interesting, since, consequently, the words in the frequency list and the words in the Finnish single word lexicon have been classified utilizing the same POS categories. As Table 6 below shows, here as well the overall distribution of POS categories was fairly similar, except that the number of nouns was somewhat higher and the number of verbs was somewhat lower in the Finnish single word lexicon.

Table 6

Distribution of Part-of-Speech Categories in the CSC Frequency List (CSC) and in the Single Word Lexicon of the Finnish Semantic Tagger (FST)

POS categories	**CSC%**	**FST%**
Abbreviation	1.40	0.83
Adjective	9.16	7.35
Adverb	7.76	6.98
Code	0.00	0.04
CompPart	0.02	1.32
Conjunction	0.49	0.25
Interjection	0.03	0.19
Noun	44.69	57.70
Numeral	0.57	0.20
Preposition	1.64	0.49
Pronoun	0.48	0.16
Proper	19.53	17.28
Verb	14.23	7.21

Thus, it is evident from the above two tables that the POS distribution in the Finnish single word lexicon is largely similar to that found in a comprehensive monolingual dictionary and in a very large corpus.

The second and third column of Table 7 below show the distribution of the Finnish single word lexicon entries in the 21 top level semantic categories. If a lexicon entry has been assigned more than one semantic tag, here the lexicon entry has been counted in the top level category of the first semantic tag representing the sense which has been considered to be the most frequent and thus the most representative sense for the lexicon entry in question. The category Z is by far the largest constituting 21.31% of the entries. The categories A (9.93%), B (8.16%), S (7.43%), L (6.11%), and O (6.11%) are also substantial. The smallest categories are P (0.79%), Y (0.84 %), and C (0.85%). The corresponding figures for the English single word lexicon (Paul Rayson, personal communication, October 6, 2010) are shown in the fourth and fifth columns. In the English single word lexicon as well, the category with most entries is Z (18.64%). It is followed by the categories A (13.17%), S (8.89%), and O (7.35%), while the smallest categories are C (0.57%), Y (0.93%), P (1.00%), and W (1.02%).

Table 7

Distribution of Entries in the Top Level Semantic Categories in the Single Word Lexicons of the Finnish Semantic Tagger (FST) and the English Semantic Tagger (EST)

Semantic Categories	**FST Entries**	**FST %**	**EST Entries**	**EST %**
A General & Abstract Terms	4,544	9.93	7,330	13.17
B The Body & the Individual	3,734	8.16	3,074	5.52
C Arts & Crafts	389	0.85	317	0.57
E Emotional Actions, States, & Processes	1,509	3.30	2,042	3.67
F Food & Farming	1,167	2.55	1,515	2.72
G Government & the Public Domain	1,235	2.70	2,057	3.69
H Architecture, Buildings, Houses, & the Home	676	1.48	875	1.57
I Money & Commerce	1,004	2.19	2,018	3.62
K Entertainment, Sports, & Games	1,420	3.10	1,188	2.13
L Life & Living Things	2,798	6.11	1,277	2.29
M Movement, Location, Travel, & Transport	2,185	4.77	3,012	5.41
N Numbers & Measurement	1,916	4.19	2,185	3.92
O Substances, Materials, Objects, & Equipment	2,796	6.11	4,091	7.35
P Education	360	0.79	554	1.00
Q Linguistic Actions, States, & Processes	2,035	4.45	2,927	5.26
S Social Actions, States, & Processes	3,401	7.43	4,949	8.89
T Time	1,418	3.10	1,444	2.59
W The World & Our Environment	718	1.57	568	1.02
X Psychological Actions, States, & Processes	2,336	5.10	3,354	6.02
Y Science & Technology	385	0.84	517	0.93
Z Names & Grammatical Words	9,755	21.31	10,376	18.64
Total	45,781	100.00	55,670	100.00

The chart below shows a comparison in the 21 top level semantic categories between the Finnish and English single word lexicons. The distribution is quite similar, which shows the maturity of the Finnish single word lexicon.

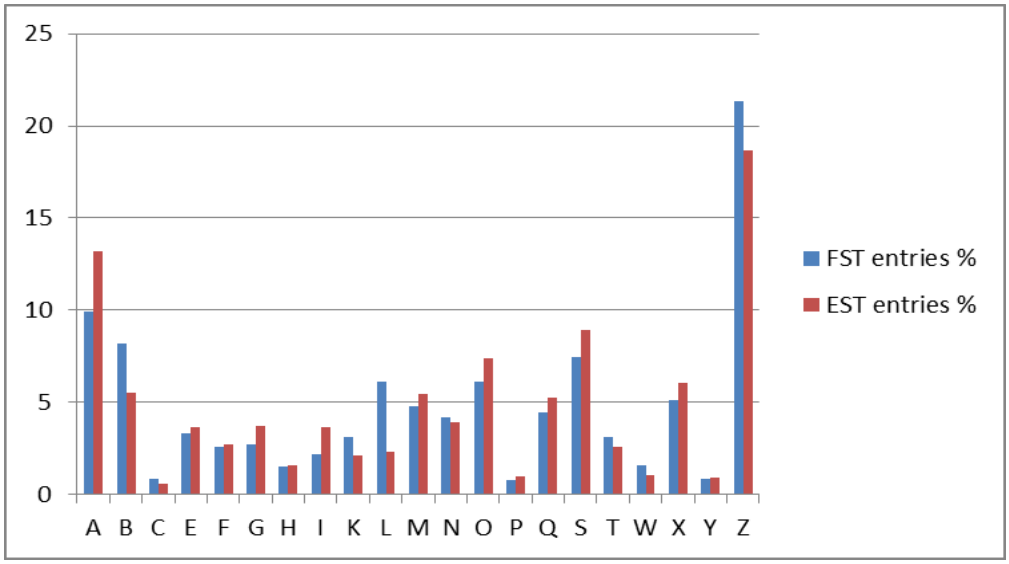

Figure 4. Distribution chart of single word lexicon entries of the Finnish Semantic Tagger (FST) and of the English Semantic Tagger (EST) in the top level semantic categories

By far most of the Finnish single word lexicon entries are unambiguous; 83.50% of them have been assigned only one semantic tag. The highest number of semantic tags, 11 in all, has been assigned to four high-frequency verbs (*mennä*, *pitää*, *tulla*, *vetää*). Table 8 below shows the distribution of the number of semantic tags per entry in the single word lexicon.

Table 8

Distribution of the Number of Semantic Tags per Entry in the Single Word Lexicon of the Finnish Semantic Tagger

Number of tags/entry	**Entries (types)**	**%**
1	38,225	83.50
2	5,285	11.54
3	1,399	3.06
4	500	1.09
5	186	0.41
6	108	0.24
7	44	0.10
8	18	0.04
9	8	0.02
10	4	0.01
11	4	0.01
Total	45,781	100.00

It is not possible to directly compare the number of semantic tags in the Finnish semantic lexicons to the number of senses in the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish") which I have most often used as my reference source since the year 2008. This is due to the fact that even though the entries in the *Kielitoimiston sanakirja* contain numbered senses, these often contain both concrete and figurative senses as well as special field senses grouped together under the same number, so the number of actual different senses is often much higher than the number of numbered senses. Nevertheless, as examples I list the following words. The highest number of numbered senses in the *Kielitoimiston sanakirja*, 26

in all, has been assigned to the verb *olla* ("to be"), whereas in the single word lexicon the number of semantic tags is four. The noun *henki* ("breath", "life", "spirit", "ghost", "person", etc.) has been assigned 11 numbered senses in the *Kielitoimiston sanakirja*, whereas it has been assigned six semantic tags in the single word lexicon. The adjective *vahva* ("strong", "powerful", "robust", "potent", etc.) has been assigned 14 numbered senses in the *Kielitoimiston sanakirja*, whereas it has been assigned six semantic tags in the single word lexicon. That said, as I pointed out in section 3.4, similarly to the English semantic lexicons, the sense distinction in the Finnish semantic lexicons is more coarse-grained than in large dictionaries, and thus some close senses which can be considered to belong in the same semantic field have been grouped together.

3.4.2 Multiword expression lexicon

In contrast to the Finnish lexicon of single words described above, the Finnish lexicon of MWEs contains entries which are units of thought consisting of two or more separate orthographic words which depict one semantic concept. Finnish MWEs include most of all multiword proper names, noun and verb phrases, idioms, and proverbs.

The MWE lexicon, which at present contains 6,113 entries, has this far been generated more or less as a by-product of the single word lexicon. In other words, when I have collected word lists of different fields for the semantic lexical resources, I have divided the items into two groups: 1) single words and 2) expressions consisting of two or more separate words with intervening spaces between them. In addition, I have included some lists of noun and verb phrases, idioms, and proverbs into it from various freely available Internet sources. On the whole, I have prioritized the development of the single word lexicon, and the MWE lexicon, as it exists at present, could perhaps better be described as a "beginning of a lexicon" or even

as a "seed lexicon". It is nevertheless a beginning. By comparison, the initial version of the English MWE lexicon was produced by exploiting the information contained in the lexical resources of the CLAWS POS tagger. Subsequently, candidate MWEs have been extracted from corpora using statistical tools, after which the selected MWEs have been manually classified into semantic categories. (Piao et al., 2005a) This corpus-driven approach to the detection of MWEs is described in more detail in Piao, Rayson, Archer, & McEnery (2005b). Similar methods might be useful for extending the Finnish MWE lexicon as well.

The following list is a sample from the Finnish MWE lexicon. The lexicon entries consist of a MWE and the relevant semantic tags. The lexicon entries have been assigned automatically generated templates, but I have deleted them from these examples. The reason for this is that this "quick MWE template solution", which was adopted in the Benedict project due to lack of time, was found to be neither very useful nor intelligent. If the FST is developed further, these automatically generated MWE templates will need to be replaced by accurate manually written templates, as is the case in the English MWE lexicon. I will discuss the "quick MWE template solution" in more detail in connection with the evaluation in section 4.2. In section 5.2.2.2, I will draft guidelines for writing accurate templates which would enable the FST to reliably recognize Finnish MWEs. Expanding the MWE lexicon and writing templates for all its entries manually is mammoth task which is beyond the scope of this thesis. MWEs are less easy to discover automatically than unknown single word entries, and writing accurate templates is very time-consuming.

persona non grata	X7-/S2
perustavaa laatua oleva	A11.1+
perä perää	N4
perään haikaileminen	X7+
pestä tiskit	B4

peukun pitäminen	A1.4
pidellä pihdeissään	S7.1+[60]
pidellä vallassaan	S7.1+
pidemmittä puheitta	Z4
Pieksämäen Lehti	Z3/Q4.2
Pieksämäen maalaiskunta	Z2
pieleen meneminen	X9.2-
pienellä äänellä	X3.2-
pienen ikänsä	T1.3+
pienen pieni	N3.2---

From this sample we learn, for example, the following. *Persona non grata* has received a slash tag: the semantic tag X7- indicates something unwanted or unchosen, and it is complemented by the semantic tag S2 which represents the category "People", so *persona non grata* thus means an unwanted person. *Perustavaa laatua oleva* ("fundamental") is something very important (A11.1+), whereas *pienen pieni* ("tiny") is something very small (N3.2---). *Perä perää* ("one after another") has been assigned the semantic tag N4 which stands for the category "Linear Order", and *pestä tiskit* ("to do the dishes") has been assigned the semantic tag B4 which represents the category "Cleaning and Personal Care". *Perään haikaileminen* ("yearning for") means the act of wanting something (X7+), *peukun pitäminen* the act of keeping one's fingers crossed for luck (A1.4), and *pieleen meneminen* the act of failing (X9.2-). *Pieksamäen Lehti* and *Pieksamäen maalaiskunta* are proper names. *Pieksamäen Lehti* (name of the local newspaper in the region of Pieksamäki, a town in Central Finland) belongs both in the category "Other Proper Names" (Z3) and in the category "The Media: Newspapers etc." (Q4.2), whereas *Pieksamäen maalaiskunta* is a geographical name (Z2). The idioms

[60] This MWE can in principle have a literal meaning as well, "to hold in one's pliers", but it is highly unlikely to appear in text.

pidellä pihdeissään and *pidellä vallassaan* ("to keep someone under one's thumb") have both been assigned the semantic tag S7.1+ which indicates having power. Finally, the idiom *pidemmittä puheitta* ("without further ado") belongs into the category "Discourse Bin" (Z4).

The second and third column of table 9 below show the distribution of Finnish MWE lexicon entries in the 21 top level semantic categories. Similarly to Table 7, if a lexicon entry has been assigned more than one semantic tag, the lexicon entry has been counted in the top level category of the first semantic tag representing the sense which has been considered to be the most frequent and thus the most representative sense for the lexicon entry in question. The corresponding figures for the English MWE lexicon (Paul Rayson, personal communication, October 6, 2010) are shown in the fourth and fifth columns.

Table 9

Distribution of Entries in the Top Level Semantic Categories in the MWE Lexicons of the Finnish Semantic Tagger (FST) and the English Semantic Tagger (EST)

Semantic Categories	**FST Entries**	**FST %**	**EST Entries**	**EST %**
A General & Abstract Terms	798	13.05	2,160	11.53
B The Body & the Individual	283	4.63	1,141	6.09
C Arts & Crafts	6	0.10	110	0.59
E Emotional Actions, States, & Processes	550	9.00	582	3.11
F Food & Farming	186	3.04	652	3.48
G Government & the Public Domain	218	3.57	781	4.17
H Architecture, Buildings, Houses, & the Home	12	0.20	430	2.30
I Money & Commerce	118	1.93	891	4.76
K Entertainment, Sports, & Games	91	1.49	815	4.35
L Life & Living Things	138	2.26	222	1.19
M Movement, Location, Travel, & Transport	170	2.78	1,552	8.29
N Numbers & Measurement	226	3.70	714	3.81
O Substances, Materials, Objects, & Equipment	49	0.80	600	3.20
P Education	219	3.58	316	1.69
Q Linguistic Actions, States, & Processes	287	4.69	784	4.19
S Social Actions, States, & Processes	581	9.50	1,559	8.32
T Time	215	3.52	818	4.37
W The World & Our Environment	27	0.44	97	0.52
X Psychological Actions, States, & Processes	623	10.19	1,036	5.53
Y Science & Technology	11	0.18	255	1.36
Z Names & Grammatical Words	1,305	21.35	3,137	16.75

Df[61]	0	0	78	0.42
Total	6,113	100.00	18,730	100.00

As with the Finnish single word lexicon, in the Finnish MWE lexicon as well the category Z is by far the largest group constituting 21.35% of the entries. Other substantial categories are A (13.05%), X (10.19%), S (9.50%), and E (9.00%). The smallest categories are C (0.10%), Y (0.18%), H (0.20%) and W (0.44%). In regard to the English MWE lexicon, there, too, the largest category is Z (16.75%), followed by A (11.53%), S (8.32%), and M (8.29%), while the smallest number of entries were contained in the categories W (0.52%), C (0.59%) and Y (1.36%). As to the total number of entries, it is useful to bear in mind that MWEs are more common in English than in Finnish, since in addition to multiword proper names (e.g. "United Kingdom") and idioms (e.g. "out of this world"), the English language also uses a large number of phrasal verbs (e.g. "die out"), which are not common in Finnish, as well as noun phrases (e.g. "general studies") the equivalent of which would be a compound in Finnish and written as one orthographic word (*yleisopinnot* "general studies")**.** Thus, even though the

[61] The marker "Df" is an abbreviation from the word "default", and it indicates templates for both single words and MWEs in which the category is defined by the word which replaces "Df", in other words, the first constituent word. For example:

*_JJ style_NN1 Df

This template is able to catch MWEs with a general adjective and the singular common noun "style", such as "Caribbean style", "musical style", or "alpine-style", and each expression is assigned the semantic tag which the respective adjective would receive. Thus, "Caribbean style" would be assigned the semantic tag Z2, "musical style" the semantic tag K2, and "alpine-style" the semantic tag W3. The following template, in turn, would catch expressions like "brutal-looking" and "untidy-looking", and they would respectively be tagged as E3-/A8 and O4.2-/A8 ("seems brutal" and "seems untidy").

_JJ looking_* Df/A8

The Df-marker has not been found necessary in the Finnish MWE lexicon this far.

Finnish MWE lexicon needs expanding, a smaller MWE lexicon would very likely be sufficient for Finnish than for English.

The figure below shows a comparison in the 21 top level semantic categories between the Finnish and English MWE lexicons. It is quite obvious that the Finnish MWE lexicon is less mature than the English single word lexicon, but this outcome could be expected, since its development has not been prioritized.

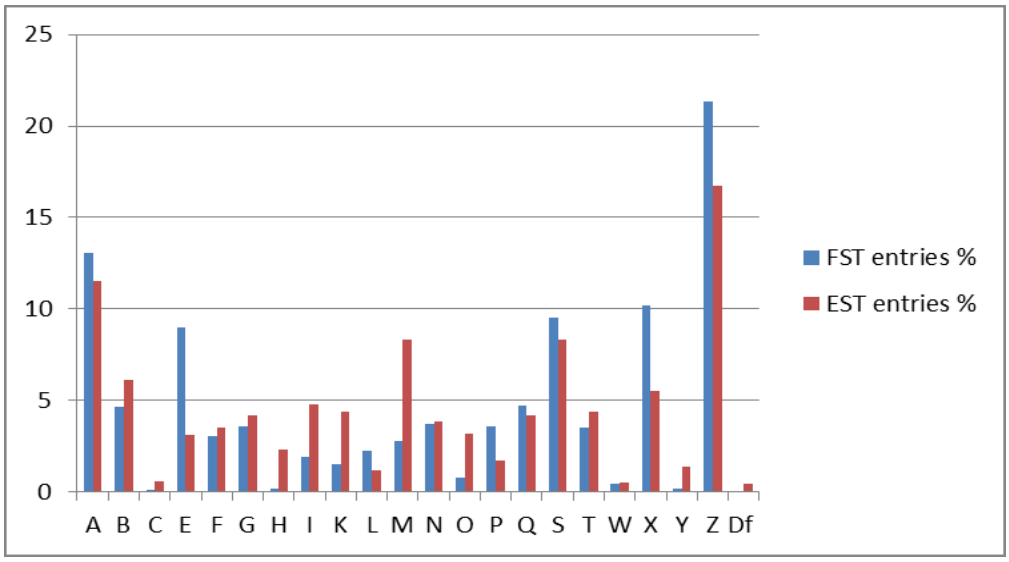

Figure 5. Distribution chart of MWE lexicon entries of the Finnish Semantic Tagger (FST) and of the English Semantic Tagger (EST) in the 21 top level semantic categories

A total of 96.14% of the Finnish MWE lexicon entries have been considered unambiguous and thus have been assigned only one semantic tag. Two semantic tags have been assigned to 234 entries, and three semantic tags have been assigned to two entries as Table 10 below shows.

Table 10

Distribution of the Number of Semantic Tags per Entry in the MWE Lexicon of the Finnish Semantic Tagger

Number of tags/entry	**Entries (types)**	**%**
1	5,877	96.14
2	234	3.83
3	2	0.03
Total	6,113	100.00

MWEs take priority over single word tagging. In other words, similarly to the English counterpart, the FST as well should first match the input text against the templates in the MWE lexicon. If it discovered a word sequence which matches one of the templates and thus together form a MWE, it should tag these words together as a unit having the same sense. If no suitable MWE template is discovered, a word is considered to be a single word and tagged individually using the single word lexicon. While the present Finnish MWE lexicon is a good "preliminary version", it is not very useful as such (see section 4.2), since the accurate templates for MWEs have not yet been written. The FST would, nevertheless, be able, with this present information, to recognize some MWEs such as the expression *perä peräät* ("one after another"), since this expression does not allow embedded elements. In comparison, the FST would not often recognize, for instance, the verb phrases *pidellä pihdeissään* and *pidellä vallassaan* ("to keep someone under one's thumb"; see page 144) because of embedded elements which such expressions often contain. For this reason, it is of utmost importance to devote sufficient time for template development so as to add intelligence to the FST.

3.5 Sample Output

This subsection illustrates the output produced by the FST. The two example sentences are:

Seuraava vuosi tuo isoja muutoksia kotimaan matkustajaliikenteeseen. Esimerkiksi junareittejä lopetetaan, uusia bussireittejä perustetaan ja halpoja matkoja on tarjolla niille, jotka ostavat lippunsa hyvissä ajoin.

Table 11 below shows how these sentences translate into English. As I pointed out in section 3.2, Finnish uses fewer orthographical words than English to convey the same information. This is clearly evident from the translated sentences in the table and due to the fact that Finnish as an agglutinative language predominately uses inflections instead of prepositions, and it does not use articles either. Note also that Finnish does not have the future tense, but the present tense is used for future time as well.

Table 11

Sample Output of the Finnish Semantic Tagger: Example Sentences

Finnish	**English**
Seuraava vuosi	The coming year
tuo	will bring
isoja muutoksia	major changes
kotimaan matkustajaliikenteeseen	in domestic passenger traffic.
.	.
Esimerkiksi	For example,
junareittejä	train routes
lopetetaan	will be eliminated
,	,
uusia bussireittejä	new bus routes
perustetaan	will be created,
ja	and
halpoja matkoja	inexpensive trips
on	will be
tarjolla	on offer
niille	for those
,	
jotka	who
ostavat	buy
lippunsa	their tickets
hyvissä ajoin	in good time
.	.

The FST tags the two example sentences as follows:

<w pos="Verb" mwe="0" sem="T2+" lem="alkaa">Alkanut</w>

<w pos="Noun" mwe="0" sem="T1.3" lem="vuosi">vuosi</w>

<w pos="Verb" mwe="0" sem="M2 I2.2/M2 A10+ M6 A2.2" lem="tuoda">tuo</w>

<w pos="Adjective" mwe="0" sem="N3.2+ T3+ N5+ A11.1+" lem="iso">isoja</w>

<w pos="Noun" mwe="0" sem="A2.1+ B2-" lem="muutos">muutoksia</w>

<w pos="Noun" mwe="0" sem="H4/M7" lem="kotimaa">kotimaan</w>

<w pos="Noun/Noun" mwe="com" sem="M3/M1/S2 M5/M1/S2 M4/M1/S2 M1/M1/S2"

lem="liikenne/matkustaja">matkustajaliikenteeseen</w>

<w pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.</w>

<w pos="_EndOfSentence" mwe="0" sem="Z99" lem="NULL">NULL</w>

<w pos="Adverb" mwe="0" sem="Z5" lem="esimerkiksi">Esimerkiksi</w>

<w pos="Noun/Noun" mwe="com" sem="M6/M3" lem="reitti/juna">junareittejä</w>

<w pos="Verb" mwe="0" sem="T2- L1-" lem="lopettaa">lopetetaan</w>

<w pos="_Delimiter" mwe="0" sem="PUNC" lem=",">,</w>

<w pos="Adjective" mwe="0" sem="T3- N5++" lem="uusi">uusia</w>

<w pos="Noun/Noun" mwe="com" sem="M6/M3" lem="reitti/bussi">bussireittejä</w>

<w pos="Verb" mwe="0" sem="T2+ X2.1 A1.1.1" lem="perustaa">perustetaan</w>

<w pos="Conjunction" mwe="0" sem="Z5" lem="ja">ja</w>

<w pos="Adjective" mwe="0" sem="I1.3- A11.1- G2.2-" lem="halpa">halpoja</w>

<w pos="Noun" mwe="0" sem="M1 F3 L1 N3.3 K5.1" lem="matka">matkoja</w>

<w pos="Verb" mwe="0" sem="A3+ A1.1.1 M6 A8" lem="olla">on</w>

<w pos="Adverb" mwe="0" sem="A9+ A3+" lem="tarjolla">tarjolla</w>

<w pos="Pronoun" mwe="0" sem="Z8" lem="ne">niille</w>

<w pos="_Delimiter" mwe="0" sem="PUNC" lem=",">,</w>

<w pos="Pronoun" mwe="0" sem="N5.1+ Z8" lem="joka">jotka</w>

<w pos="Verb" mwe="0" sem="I2.2 G2.2-/A9-" lem="ostaa">ostavat</w>

<w pos="Noun" mwe="0" sem="Q1.1" lem="lippu">lippunsa</w>

<w pos="Adverb" mwe="0" sem="T4+" lem="hyvissä_ajoin">hyvissä</w>

<w pos="Adverb" mwe="0" sem="T4+" lem="hyvissä_ajoin">ajoin</w>

<w pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.</w>

<w pos="_EndOfSentence" mwe="0" sem="Z99" lem="NULL">NULL</w>

This output displays the following types of information:

- POS tag (pos"=XXX"). For example, "pos"=Noun"" indicates a noun, such as in the case of *lippu* ("ticket").
- Membership (mwe="XXX"). Single words are marked as "mwe="0"", such as in the case of *vuosi* ("year"). The membership value "mwe="com"" indicates that the output has been produced by the compound engine (see section 3.3.2), such as in the case of *reitti/juna* ("route/train"). If a word was a constituent word in a MWE, this would be marked as "mwe="mwe"".
- Semantic tags (sem="XXX"). The semantic tags in the lexicon entry are listed in perceived frequency order, for example, "sem="I2.2 G2.2-/A9-" for the verb *ostaa* ("to buy").
- Basic form of the input word (lem="XXX">XXX). For example, "lem="ostaa">ostavat; *ostavat*" ("they buy"): *ostavat* is the present tense, active voice, indicative mood, and the third person plural form of the verb *ostaa* ("to buy").

Furthermore, the output contains the following information:

- Markers <w [XXX] </w> are XML word tags.
- Delimiters, such as commas, are indicated in the following manner: <w pos="_Delimiter" mwe="0" sem="PUNC" lem="XXX">XXX</w>.
- The end of the sentence is indicated in the following manner:

<w pos="_EndOfSentence" mwe="0" sem="Z99" lem="NULL">NULL</w>.

The sample output also displays how compounds are treated in the FST. As I noted in section 3.3.2, the number of possible Finnish compounds is infinite, so it would be totally impossible to collect all potential candidates. For this reason, the Benedict team decided to include only the most frequent compounds as well as lexicalized compounds in the single word lexicon, and all other possible, less frequently used compounds of a more temporary nature are handled by the compound engine. An example of a frequent compound is *kotimaa* ("home country") which appeared in the above example sentence in the genitive singular (*kotimaan*). This compound is included in the single word lexicon and was tagged by the FST in the following way:

<w pos="Noun" mwe="0" sem="H4/M7" lem="kotimaa">kotimaan</w>

By comparison, the compound *matkustajaliikenne* ("passanger traffic"), which appeared in the above example sentence in the illative singular (*matkustajaliikenteeseen*), is a less frequently used compound which is not included in the single word lexicon. The compound engine has generated the following interpretation for it:

<w pos="Noun/Noun" mwe="com" sem="M3/M1/S2 M5/M1/S2 M4/M1/S2 M1/M1/S2"
lem="liikenne/matkustaja">matkustajaliikenteeseen</w>

Another useful feature in the FST can be seen in the following excerpt from the sample output:

<w pos="Adverb" mwe="0" sem="T4+" lem="hyvissä_ajoin">hyvissä</w>

<w pos="Adverb" mwe="0" sem="T4+" lem="hyvissä_ajoin">ajoin</w>

As I wrote in section 3.4, some frequently co-occurring MWEs have been included in the single word lexicon and not in the MWE lexicon. These are fixed expressions in which the constituent words cannot be inflected, no enclitic particles are used, and where no embedded elements are allowed between the constituents. Since TextMorfo in the POS tagging phase processes some such fixed expressions as single units and then assigns a POS tag to the entire expression, it was practical to treat these as single units in the semantic tagging component as well. Therefore, they have been included in the single word lexicon, and the spaces between the constituents have been replaced by underscores. This expression, *hyvissä ajoin* ("in good time"), is a good example of such a case.

3.6 Chapter Summary

This chapter has described the development and the structure of the Finnish semantic lexical resources and has thus answered **RQ1 (What do the Finnish semantic lexical resources consist of, what type of principles and practices have been followed in their creation, and how do these resources differ from their English counterparts both in terms of content and construction?)**. The overall aim in the development process of the FST has been to adapt the semantic analysis system originally developed for English to meet the needs of semantic analysis of Finnish. The semantic lexical resources which I have created in this thesis provide the knowledge base for the FST and are my most important contribution to it. Consequently, we now have at our disposal equivalent semantic taggers based on equivalent semantic lexicons which are suitable for processing both English and Finnish.

Subsequently, equivalent semantic taggers have been developed also for Czech, Chinese, Dutch, French, Italian, Malay, Portuguese, Russian, Spanish, Urdu, and Welsh.

At first, I have looked at the initial phases of the development process. Thereafter, I have provided a brief summary of the development and the structure of the software component. It was essential to start from this, since semantic lexical resources such as ours cannot be developed in isolation, but the software in which they will be applied needs to be taken into account in many respects all through the development process. The software component, which was developed during the Benedict project, needed some modification to enable it to process the specific features of the Finnish language. The English POS tagger CLAWS was replaced by TextMorfo, a Finnish counterpart, and a new component named the compound engine was included in the software to process Finnish compounds which are not included in the semantic lexical resources. In addition, the encoding system of the whole USAS framework, to which all these semantic taggers belong, was changed to Unicode to allow the software to cope with the Scandinavian characters *å*, *ä*, and *ö*. However, the semantic categories originally developed for the analysis of the English language did not need any modification at all. Indeed, they were found to be suitable in every respect to the semantic categorization of objects and phenomena in the Finnish language and culture as well.

I have then proceeded to presenting the semantic lexical resources for Finnish which are my major contribution to the FST and the main focus of this thesis. They were built from scratch and made compatible with the English lexical resources. I have detailed the steps which have been taken during the development process. Consequently, I have described the content of both the single word lexicon and the MWE lexicon respectively as well as the principles and practices which I have followed in their creation. In addition, I have explained how these resources differ from the English counterpart both in terms of content and construction.

The present single word lexicon contains in all 45,781 entries, and the MWE lexicon contains 6,113 entries. The single word lexicon is already mature, and the MWE lexicon, in turn, is a good "preliminary version" which will become much more useful when it is expanded and when accurate templates are written manually for all its entries. I will discuss the further development of the Finnish semantic lexical resources in more detail in chapter five.

I have concluded this chapter by presenting a sample of the output of the FST. The output shows the POS and semantic categories which the input words have been assigned. It also displays how compounds and MWEs are treated in the FST. In the following chapter, I will evaluate how extensive the semantic lexical resources are and how well they perform when they are applied in the FST software.

4 Evaluation

4.1 Introduction

This chapter reports the results of the evaluations of the Finnish semantic lexical resources which were described in the previous chapter. I will begin by briefly summarizing the results of the formative evaluation carried out at the end of the Benedict project during which the prototype of the FST was developed. Subsequently, I will describe the results of two new evaluations which have been carried out after extending and improving the single word lexicon. These evaluations are:

1) the final evaluation which measures the lexical coverage of the Finnish single word lexicon, in other words, the proportion of words which are covered by this lexicon, and
2) the application-based evaluation which measures the accuracy of the FST, in other words, how well the Finnish single word lexicon performs when it is applied in the FST software.

Both of these evaluations are quantitative by nature, and they answer **RQ2 (How extensive is the Finnish single word lexicon in terms of lexical coverage?)** and **RQ3 (How suitable is the Finnish single word lexicon for use in the semantic analysis of Finnish in the FST software?)**. There is no standard for evaluating such semantic lexical resources, but various methods have been used. For example, the developers of LIWC (see section 2.3.2.2) used a panel of judges to review the classified words (Pennebaker et al., 2015, pp. 5–6). In

this thesis, I have carried out evaluations which are similar to the evaluations reported for the EST, since I believe that the performance of the EST and its semantic lexical resources are the best comparison point for the Finnish equivalents. Although the results are limited by the size and scope of the test corpora, I believe that they are sufficiently diverse to provide an insight into the overall performance of the Finnish single word lexicon. After presenting the results of the evaluations, I will analyze the different types of errors which I discovered in the application-based evaluation and suggest solutions for addressing them. Finally, I will conclude the chapter by presenting the results of the third new evaluation of the Finnish single word lexicon. This evaluation is referred to as the "semantic labeling experiment", and it measures how general native speakers of Finnish are able to replicate the categorisation of the Finnish single word lexicon. This tests the inter-rater reliability to replicate the experiment as well as the understandability of the USAS taxonomy in the Finnish context.

4.2 Formative Evaluation at the End of Benedict Project

During the Benedict project, we evaluated the FST and its semantic lexical resources on various test data and on several occasions. In the evaluation at the end of the project in January 2005, we examined the lexical coverage and the accuracy. At that point of development, the semantic lexical resources consisted of 33,627 single words and 8,912 MWEs[62]. In this section, I will discuss the results only briefly; for more information, see Löfberg et al. 2005.

[62] The number of entries in the MWE lexicon has been reduced since the Benedict project. This is due to the fact that some of the entries were not found to be very useful. For example, all the MWEs which TextMorfo processes as one unit (such as *aamusta iltaan* ("from morning to evening"; see section 3.4.1) were moved from the MWE lexicon to the single word lexicon, and the spaces between the constituents were replaced by underscores.

In the first experiment, we examined the lexical coverage, in other words, the extent of the Finnish semantic lexical resources. The calculation was performed in the following way. The test corpora were entered into the FST, after which the FST grammatically and semantically tagged all the words they contained. When the FST encountered a word which it did not recognize as a single word or as a constituent of a MWE, in other words, the word was discovered to be missing from the semantic lexical resources, the FST assigned such word the semantic tag Z99 which represents the category "Unmatched". In addition, a word was considered to belong in the Z99 category if the compound engine had produced a slash tag for it in which the first semantic tag was Z99, for example:

pos="Noun/Noun" mwe="com" sem="Z99/I2.2/S2" lem="omistajuus/asiakas">asiakasomistajuus</w> and pos="Adjective/Noun"

To compare, a word was considered not to belong in the Z99 category if only the last semantic tag was Z99, for example:

pos="Noun/Noun" mwe="com" sem="H1/F1/Z99" lem="ravintola/gourmet">gourmet-ravintoloista</w>

As I noted in section 3.3.2, the compound engine places the second constituent of the compound first, because the second constituent is usually more significant in terms of the meaning of the compound than the first constituent. Consequently, the first constituent of the compound, which usually modifies the second constituent, is placed second. The former of the above two examples is the compound *asiakasomistajuus* ("co-operative membership"; literally "client ownership") which consists of the constituents *asiakas* ("client") and *omistajuus* ("ownership"). The first constituent *asiakas* is tagged correctly, but since this constituent does not convey the meaning of the compound but only modifies the second

constituent *omistajuus*, which is missing from the single word lexicon and has thus received the semantic tag Z99, this has been considered an error. By comparison, the latter of the above two examples is the compound *gourmet-ravintola* ("gourmet restaurant", here in the plural elative) which consists of the constituents *gourmet* ("gourmet") and *ravintola* ("restaurant"). Here the latter constituent *ravintola* has been tagged correctly, and thus the whole compound has been considered to be correctly tagged, since the semantic tag H1/F1 for the latter constituent conveys well the meaning of the compound even though the word *gourmet* is missing from the single word lexicon; a gourmet restaurant is indeed a type of a restaurant. It is not necessary to provide a finer-grained definition for this compound, but this level of granularity is appropriate for such a general language lexicon. Finally, from the tagged output, all the instances of words tagged as Z99 were extracted, and their percentage of the total number of words was calculated[63].

The disregarding of the comparison of adjectives and adverbs was not considered an error in this evaluation, because this shortcoming in the FST is very insignificant. In the present state, the FST tags the comparative and superlative forms of adjectives and adverbs in the same way as it tags the basic forms. By comparison, the EST takes into account the comparison of adjectives and adverbs by using double pluses or double minuses for comparatives and triple pluses or triple minuses for superlatives[64]. This is a straightforward task, however, since the English semantic lexical resources contain entries not only for basic forms but also for comparative and superlative forms which are formed with inflections as well as for the irregular forms, for example[65]:

quick	JJ	N3.8+

[63] ((total words – unmatched words) / total words) x 100% = lexical coverage%

[64] For more information, see section 2.3.1.

[65] The abbreviations in the second column come from the CLAWS tagset. JJ indicates a general adjective, JJR a general comparative adjective, and JJT a general superlative adjective.

quicker	JJR	N3.8++
quickest	JJT	N3.8+++

When text is entered into the EST, the program compares it to the semantic lexicon entries and is thus able to find the correct interpretation for all instances of comparative and superlative forms. In contrast, in the FST, the TextMorfo component, which carries out the grammatical tagging prior to the semantic tagging, processes all adjectives and adverbs in the input text first into basic forms and only after that passes these basic forms on to the semantic tagging component (see Figure 3 in section 3.3.4). Thus, all adjectives and adverbs irrespective of the comparison receive the same semantic tag. For this reason, it would be pointless to add the comparative and superlative forms of adjectives and adverbs into the Finnish semantic lexical resources. However, should there be a need to differentiate between comparisons, a component could be developed which would enable the FST to recognize the comparative and superlative forms and automatically add the relevant pluses and minuses at the end of the semantic tag for the basic form. This would be technically feasible, since even though the information about the comparison does not show in the POS tag produced by TextMorfo, it is included in the output and could thus be utilized for this purpose.

The Benedict team compiled two different test corpora for examining the lexical coverage in the formative evaluation. The first corpus was a random collection of articles from *Helsingin Sanomat*, the biggest Finnish daily newspaper, which consisted of 24,452 words and represented general modern standard Finnish language in the evaluation. The second corpus was a random collection of Internet texts which dealt with the past and present of Finnish cooking. It consisted of 4,264 words and represented domain-specific data in the evaluation. The lexical coverage achieved was 90.68% on the Helsingin Sanomat Corpus and 94.58% on the Finnish Cooking Corpus, as shown in Table 12 below:

Table 12

Formative Evaluation of Lexical Coverage at the End of the Benedict Project

Test corpus	**Total words**	**Unmatched words**	**Lexical coverage(%)**
Helsingin Sanomat Corpus	24,452	2,278	90.68
Finnish Cooking Corpus	4,264	231	94.58

We considered these results promising. They suggested that already at that point of development the Finnish semantic lexicons provided a practically useful resource in terms of lexical coverage.

Secondly, we examined the accuracy of the FST, in other words, how well the Finnish semantic lexical resources perform when they are applied in the FST software. For testing purposes, we compiled a small corpus which consisted of 3,044 words from a subset of the Finnish Cooking Corpus. This test corpus was tagged automatically with the FST, after which the output was checked manually. In all, we found 515 errors resulting in an accuracy of 83.08% which we considered an encouraging outcome for a prototype tool.

The errors occurred were analyzed and categorized. In all, 45.24% of the errors were found to be lexicon-related. They were caused by the fact that the given single word or MWE was not included in the semantic lexicons or by the fact that the semantic tag for the given sense was not included in the relevant semantic lexicon entries. Furthermore, 54.77% of the errors were caused by mis-performance of the program and lack of disambiguation methods. After analyzing the errors, we concluded that improving the accuracy of the FST would require:

1) expanding and editing the semantic lexical resources,
2) enhancing the performance of the software by improving the accuracy of TextMorfo and the compound engine, and
3) developing effective disambiguation procedures, such as the disambiguation procedures used in the EST (see section 2.4.1.3), which would include, for example, context rules, context-based disambiguation algorithms, and components for recognizing auxiliaries.

Since the termination of the Benedict project, I have expanded and improved the single word lexicon. My aim has been to cover the majority of the core vocabulary of general standard modern Finnish, and I have also edited many of the entries which I had written during the Benedict project to include missing senses and to correct mistakes. In regard to the MWE lexicon, it had become very clear from the formative evaluation that the "quick MWE template solution", which we had adopted in the Benedict project, was neither very useful nor intelligent. Thus, the MWE lexicon would not only need expanding in terms of adding new entries into it, but, most of all, it would also be necessary to create an entirely novel system for writing templates in order to be able to reliably recognize different types of Finnish MWEs. However, this result was not in any way unexpected, considering the limited amount of time and effort the Benedict team had had to devote to the development of the MWE component.

This "quick MWE template solution" did not include manually written MWE templates as is the case with the English MWE lexicon. Instead, all the words in the MWE lexicon entries were processed into basic forms and tagged grammatically with TextMorfo, and the resulting list was then saved into the FST as a MWE lexicon. As a result, this lexicon consists of both the actual MWEs and their lemmatized TextMorfo outputs, the "templates", for example:

Euroopan investointipankki	Z3/I1
Eurooppa_Proper investointipankki_Noun	Z3/I1

The constituents of the MWE *Euroopan investointipankki* ("European Investment Bank"; literally "Europe's Investment Bank") have thus been reduced to their TextMorfo outputs to form a "template" in the following way:

Eurooppa_Proper investointipankki_Noun
(literally "Europe_Proper Investment Bank_Noun")

When text is entered into the FST, the FST tags it and simultaneously compares the tagged output to these "templates", and when the FST discovers a matching pattern, it tags the pattern as a MWE. Sometimes this solution manages to produce a correct interpretation, for example, probably in most of the cases with the above example "template". The reason for this is that when the words *Eurooppa* and *investointipankki* appear in text consecutively, they usually refer to the name of this particular lending institution, *Euroopan investointipankki*.

However, for many MWEs the case is more complicated. This can be exemplified with the idiom *antaa kenkää* ("to give the sack"; literally "to give shoe") for which TextMorfo has produced the following "template" in the MWE lexicon:

antaa kenkää	I3.1-/A2.2
antaa_Verb kenkä_Noun	I3.1-/A2.2

This "template" *antaa_Verb kenkä_Noun* ("to give_Verb shoe_Noun") not only captures the idiom in question but also the literal meaning of the constituent words, for example, in the sentence *Annoin kengän Marialle*. ("I gave the shoe to Maria."). This is because it is not

specified in the "template" in any way that if the idiom is in question, the noun always appears in the partitive singular (*kenkää*), whereas in other cases the literal meaning is the correct meaning. Thus, at present, the literal meaning would also be tagged as I3.1-/A2.2 which would, naturally, be an incorrect semantic tag.

Moreover, with these current "templates", the FST can only recognize such MWEs in which the constituent words appear consecutively in text. In case there is an embedded element, such as *hänet* ("him") in the sentence *Puhuin hänet ympäri* ("I persuaded him"; literally "I spoke him around"), the FST misses the idiom *puhua ympäri* completely and, instead, tags the constituent words separately again resulting in a wrong interpretation.

In the following subsections, I will evaluate the new expanded and improved Finnish semantic lexical resources in terms of lexical coverage and accuracy. These two evaluations differ from the above described formative evaluation in one significant respect: I decided to omit the MWE lexicon completely in the new evaluations and use the single word lexicon only. This decision arose from the fact that, as described above, the "quick MWE template solution" was found to be impractical during the Benedict project. The MWE lexicon needs to be written manually into accurate templates, as is the case in the English MWE lexicon, but carrying out this task at this stage was not possible, since it would have been far beyond the scope of this thesis. However, I have drafted guidelines for creating such templates which would reliably recognize different types of Finnish MWEs. These guidelines will be presented in section 5.2.2.2.

The experiments measuring the lexical coverage in the final evaluation have been carried out in exactly the same way as in the formative evaluation at the end of the Benedict project; thus, the results are comparable. In contrast, the results of the application-based evaluation of accuracy are not directly comparable to the results of the formative evaluation of accuracy. The reason for this is that I have categorized the errors occurred in a slightly different way. In

the formative evaluation, the errors were grouped into ten categories, whereas in the application-based evaluation, I have used in all fourteen categories. The four additional categories which I considered necessary in this new evaluation are: 1) Existing MWE Templates Not in Use, 2) Errors Caused by the Lack of Context Rules, 3) Errors Caused by Archaic Use of Language, and 4) Errors Caused by Colloquial Use of Language.

Lastly, it should be noted that the word count in all of these evaluations has been carried out according to the FST output which is based on TextMorfo. The FST splits and lumps words slightly differently than, for example, Microsoft Word. By way of illustration, the FST usually considers hyphenated words, such as *15-vuotias* ("15-year-old"), *1950-luvulla* ("in the 1950s), and *uv-säde* ("UV ray"), as two separate words, whereas it counts numbers like 4 100 (in Finnish, thousands are often separated by a space) as one word. Nevertheless, the end result of the word count according to the FST is practically the same as when using Microsoft Word. In regard to the fixed expressions which are included in the TextMorfo lexicon and which TextMorfo processes in the POS tagging phase as single units (e.g. *muun muassa* ("among other things"), *puolin ja toisin* ("reciprocally"), and *sen sijaan* ("instead")[66]), the FST tags all the constituents within the fixed expression separately, and thus this does not cause confusion to the word count, as the following example illustrates:

<w pos="Adverb" mwe="0" sem="S1.1.2+" lem="puolin_ja_toisin">puolin</w>

<w pos="Adverb" mwe="0" sem="S1.1.2+" lem="puolin_ja_toisin">ja</w>

<w pos="Adverb" mwe="0" sem="S1.1.2+" lem="puolin_ja_toisin">toisin</w>

[66] The single word lexicon contains 764 such entries; these were discussed in section 3.4.1.

4.3 Final Evaluation of Lexical Coverage

In the first experiments of the final evaluation, I examined the lexical coverage of the new expanded and improved Finnish single word lexicon on a variety of corpora to determine how extensive this lexicon is at present. My aim in the lexicon development has been to cover the majority of the vocabulary of general modern standard Finnish to make the resource useful for various purposes. For this reason, I have selected the test corpora from different sources to ensure that the result of the evaluation would reflect the lexical coverage of the resource in different types of practical annotation tasks. In the following subsection, I will present these corpora and, subsequently, will report the results obtained from the experiments. Note that this evaluation is concerned with words which are missing from the single word lexicon. Senses which are missing from the existing single word lexicon entries will be discussed in sections 4.4.2 and 4.4.3.1.2 in connection with the application-based evaluation.

4.3.1 Test corpora

I compiled the test corpora for the final evaluation myself for two reasons. Firstly, unlike the case for English, there are no large general reference corpora for Finnish. Secondly, the selection of freely downloadable corpora containing clean data for Finnish is very limited because of copyright issues, especially with respect to modern Finnish. I chose various types of texts in order to be able to investigate the lexical coverage potential of the Finnish single word lexicon from such different aspects as genre, domain, and historical period. I based my choice of text types on my knowledge of reference corpora of English, and the choice of text types largely reflects the text types included in the Lancaster-Oslo/Bergen Corpus[67]. I selected

[67] For more information, see http://www.hit.uib.no/icame/lob/lob-dir.htm#lob4.

some of the test material employed in the following experiments from freely downloadable corpora, but the majority consists of texts which I collected manually from various freely accessible Internet sources which are not password protected. The Internet is a rich source of information, and Seale, Charteris-Black, MacFarlane, and McPherson (2010, p. 598) contend that, even though the opinion in the Internet research community is rather divided, they consider that neither informed consent nor ethical review is required to use in research such messages which are in the public domain. Moreover, I do not intend to release my test data in its entirety, but only separate single words and some sentences are displayed in the examples of this thesis.

Firstly, I compiled five corpora to represent general modern standard Finnish in this evaluation. These corpora are:

1) The Helsingin Sanomat Corpus: a random collection of body text in news headlines, articles on various topics, columns, editorials, etc. from the online edition of *Helsingin Sanomat* (Helsingin Sanomat, n.d.), the biggest Finnish daily newspaper. The texts were dated 5 June 2009–22 November 2011. This corpus consists of 24,688 words. Thus, it is from the same source and approximately of the same size as the Helsingin Sanomat Corpus which was used in the formative evaluation at the end of the Benedict project.
2) The President Halonen's New Year's Speeches Corpus: former president Tarja Halonen's speeches on New Year's Eves from 2001 to 2007. The speeches are available in the corpus collection of modern Finnish texts provided by the Institute for the Languages of Finland (Kotimaisten kielten keskus, 2007). This corpus consists of 6,045 words.

3) The Ellit Corpus: a random collection of body text in articles on different topics from *Ellit* (Ellit, n.d.), an interactive Internet magazine for women. The texts were dated 20 September 2006–24 November 2012. This corpus consists of 18,385 words.
4) The La Habanera Corpus: a reality-based novel *La Habanera*, written by Päivi and Santeri Kannisto (2005), in which the authors tell about their lives and the reasons which lead to their decision to drop out of the rat race. This corpus consists of 7,760 words.
5) The Kauniita Valheita Corpus: the Finnish-language translation *Kauniita valheita* (2007) by Tiina Sjelvgren from the English original of the fictional novel *Beautiful Lies* written by Lisa Unger. This corpus consists of 88,657 words.

As mentioned earlier, the Finnish single word lexicon has been developed for the analysis of general Finnish language. However, I wanted to evaluate its lexical coverage on a specific domain as well. Therefore, for the second experiment, I compiled a Finnish Culinary Culture Corpus of headings and body text from the report *Suomalaisen ruokakulttuurin ulottuvuuksia* ("Dimensions of Finnish Culinary Culture") (Ruokatieto, n.d.) which deals with various topics related to culinary culture, such as Finnish food and taste, values, opinions, seasons, festivities, meals, food production, and table manners. This corpus contains 7,518 words and by and large reflects the same themes as the texts which constituted the Finnish Cooking Corpus used in the formative evaluation at the end of the Benedict project.

Moreover, even though the Finnish single word lexicon has been developed for the analysis of modern Finnish language, I wanted to investigate how it manages to process older text. If the results were promising, the single word lexicon could provide a helpful resource for the analysis of historical text as well. For this purpose, I selected as test material classic

novels and newspaper articles written by five different authors—Juhani Aho, Minna Canth, Arvid Järnefelt, Teuvo Pakkala, and Kyösti Wilkuna—between the years 1884 and 1930. These works are available in the corpus collection of classic Finnish literature which is provided by the Institute for the Languages of Finland (Kotimaisten kielten keskus, 2006). I compiled the following five corpora of them:

1) The Juhani Aho Corpus: *Katajainen kansani ja muita uusia ja vanhoja lastuja* (1891–1899), *Minkä mitäkin Italiasta* (1906), *Minkä mitäkin Tyrolista* (1908), and *Lohilastuja ja kalakaskuja* (1921). This corpus consists of 120,717 words.
2) The Minna Canth Corpus: *Naiskysymyksestä—Lehtikirjoituksia* (1884–1896), *Hanna* (1886), *Köyhää kansaa* (1886), *Kauppa-Lopo* (1889), *Lain mukaan* (1889), *Lehtori Hellmanin vaimo* (1890), and *Agnes* (1892). This corpus consists of 124,910 words.
3) The Arvid Järnefelt Corpus: *Isänmaa* (1893), *Elämän meri* (1904), *Maaemon lapsia* (1905), *Veneh'ojalaiset* (1909), *Greeta ja hänen herransa* (1925), and *Vanhempieni romaani I–III* (1928–1930). This corpus consists of 336,844 words.
4) The Teuvo Pakkala Corpus: *Elsa* (1894), *Lapsia* (1895), and *Pikku ihmisiä* (1913). This corpus consists of 85,539 words.
5) The Kyösti Wilkuna Corpus: *Tapani Löfvingin seikkailut isonvihan aikana* (1911). This corpus consists of 42,071 words.

Lastly, even though the Finnish single word lexicon has been developed for the analysis of standard Finnish, in the fourth experiment, I wanted to investigate the effect of the more informal type of language often found in Internet discussions on the lexical coverage. If the results were promising, the single word lexicon could also be useful for various applications analyzing Internet content. For this experiment, I collected three corpora containing online

discussions from three different forums which are publicly accessible, in other words, reading the messages is not password protected.

1) The Death and Mourning Corpus: headings and body text from various threads under the topic Kuolema ja suru ("Death and Mourning") at the online discussion forum Suomi24[68] (Suomi24, n.d.). The texts were dated 22 November 2006–10 February 2012. This corpus consists of 30,295 words.
2) The Niinistö or Haavisto Corpus: body text of the 413 posted messages from the thread Toinen kierros: Niinistö vai Haavisto ("The second electoral round: Niinistö or Haavisto") at A-Tuubi (Yle, n.d.), the online discussion forum of Yleisradio Oy, Finland's national public service broadcasting company. The texts were dated 23 January 2012–9 February 2012. This corpus deals with the presidential election in Finland at the beginning of the year 2012 and consists of 51,157 words.
3) The Separate Pool Times for Muslims Corpus: body text of the 1,252 posted messages from the thread Muslimeille omat uintivuorot ("Separate pool times for Muslims") at the online discussion forum of *Iltalehti* (Iltalehti, n.d.), a Finnish yellow press newspaper. The texts were dated 29 December 2011–5 January 2012. This corpus consists of 114,003 words.

4.3.2 Results

I conducted the first experiments in the final evaluation to investigate the lexical coverage of the Finnish single word lexicon on general modern standard Finnish. I performed the

[68] All the texts from the discussion forums of Suomi24 from 2001 to June 2015 were published as a corpus by the Language Bank of Finland in 2015. This corpus is freely available for academic use and can be downloaded from https://www.kielipankki.fi/aineistot/.

calculation in the same way as in the formative evaluation at the end of the Benedict project (see section 4.2). The five test corpora, which were described in the previous subsection, produced the results displayed in Table 13 below:

Table 13

Final Evaluation: General Modern Standard Finnish

Test corpus	**Total words**	**Unmatched words**	**Lexical coverage(%)**
Helsingin Sanomat	24,688	1,240	94.98
President Halonen's New Year's Speeches	6,405	134	97.91
Ellit	18,340	913	95.02
La Habanera	7,760	373	95.19
Kauniita Valheita	88,657	4,806	94.58

The highest lexical coverage, 97.91%, was achieved on the corpus which consisted of President Halonen's New Year's speeches. The results were very good also on the other corpora, ranging from 95.19% to 94.58%. This shows that the Finnish single word lexicon in its present state indeed covers the majority of the core vocabulary of Finnish and is thus capable of dealing with most general domains which appear in general modern standard Finnish text. In comparison, the English semantic lexical resources obtained the lexical coverage of 97.59% when they were applied to general modern English represented by the written section of the BNC Sampler corpus[69] which contains a total of 970,532 words (Piao et al., 2004). The result is quite similar to the results obtained in the final evaluation of the Finnish single word lexicon. However, it must be noted that the results are not strictly comparable, since the test data for English was almost seven times larger than the test data for

[69] For more information, see http://www.natcorp.ox.ac.uk/corpus/sampler/sampler.pdf.

Finnish. Recently, the lexical coverage potential of all the twelve semantic lexicons belonging to the extension of the USAS framework for English was measured (Piao et al., 2016). The results were very encouraging, with the top coverage of 95.93%[70] for Finnish.

Lindén and Niemi (2014) report a set of evaluations to measure another existing large semantic lexical resource for Finnish, Finnish WordNet, which was discussed in section 2.5.2.2. For testing the lexical coverage of the FiWN, they used a large text corpus of Finnish newspaper text. They discovered that the entries in FiWN 2.0 covered 57.3% of all the words in their corpus. After excluding proper names, which do not belong in the FiWN lexicon, and including only nouns, verbs, adjectives, and adverbs, the coverage was 82.4% of running text. To my knowledge, an evaluation of lexical coverage has not been carried out for the lexicons of the Finnish Semantic Web which was another large semantic lexical resource for Finnish discussed in section 2.5.2.2. However, it must be noted that the semantic lexical resources for the FiWN and for the Finnish Semantic Web differ from the semantic lexical resources dealt with in this thesis in that their purpose is not to cover words representing all parts of speech. For this reason, the evaluation of the lexical coverage of the FiWN is not comparable to the final evaluation discussed in this section.

In the second experiment, I examined the lexical coverage of the Finnish single word lexicon on a specific domain. For this purpose, I used the corpus which consisted of texts on various aspects of Finnish culinary culture. The lexical coverage obtained was 95.36%, as Table 14 below reveals:

[70] This is an average of the results 95.89% for a corpus containing newspaper text and 95.98% for a corpus containing blog text.

Table 14

Final Evaluation: Specific Domain of Finnish Culinary Culture

Test corpus	**Total words**	**Unmatched words**	**Lexical coverage(%)**
Finnish Culinary Culture	7,518	349	95.36

This was also a good result, even though the text contained some jargon, technical terms, and other domain-specific vocabulary. Nevertheless, the lexical coverage would undoubtedly drop if the single word lexicon was tested, for example, on a more specialized domain, such as food science or molecular gastronomy. However, should there be a need to use the single word lexicon for the analysis of such specialized domains, it would be possible to tailor it for this particular purpose. Issues connected with this will be discussed in section 5.3.

The English semantic lexical resources have also been evaluated on a specific domain. For this purpose, the developers used journalistic reports on law and court stories collected from the UK Press Association newswire service and from nine UK mainstream newspapers which were included in the METER corpus[71]. The test corpus consisted of 241,311 words and produced a lexical coverage of 95.38%, even though the texts did contain a considerable amount of technical terms and jargon. Thus, the lexical coverage was not equally good as could be expected on general language corpora, but the drop was surprisingly small. (Piao et al., 2004)

Next, I wanted to investigate how the Finnish single word lexicon manages to process older texts by examining the lexical coverage on four corpora which consist of classic novels and newspaper articles written between the years 1884 and 1930. The results obtained in this third experiment were almost congruent, as is evident from Table 15 below:

[71] For more information, see http://nlp.shef.ac.uk/meter/.

Table 15

Final Evaluation:

Classic Novels and Newspaper Articles Written between 1884 and 1930

Test corpus	**Total words**	**Unmatched words**	**Lexical coverage(%)**
Juhani Aho	120,717	8,919	92.61
Minna Canth	124,910	8,948	92.84
Arvid Järnefelt	336,844	26,569	92.11
Teuvo Pakkala	85,539	5,944	93.05
Kyösti Wilkuna	42,071	3,163	92.48

The standard Finnish language as it exists today began to establish itself around the 1880s (Häkkinen, 1994, p. 15), but only around the 1920s, after Finland had gained independence, were modern spelling norms and grammar standardized (Pulkkinen, 1972, pp. 57–65). This is evident in these test corpora: the language used is still slightly different from what it is in the present day. In addition to the differences in spelling and grammar, the texts also contained a fair amount of archaic vocabulary which does not exist in the modern Finnish language anymore. Despite all this, the single word lexicon developed for the analysis of modern general Finnish functioned surprisingly well at this task, which suggests that the single word lexicon could already be of some help in various tasks requiring automatic semantic analysis of older Finnish text.

The lexical coverage of the English semantic lexical resources with respect to older English text has also been evaluated. The UCREL team drew their first historical test material from the Lancaster Newsbooks Corpus[72], which contains newsbooks of the mid-seventeenth century, and tagged it using the same semantic lexicons that were used for modern English,

[72] For more information, see http://www.lancs.ac.uk/fass/projects/newsbooks/.

obtaining the lexical coverage of 94.40%. A similar drop in the lexical coverage as on the domain-specific METER Corpus was detected, which could be expected due to the historical variants in the test text. (Piao et al., 2004) I believe that this test corpus by and large represents approximately the same phase of development in the English language as the test corpora used in the final evaluation of this thesis. Baron, Rayson, and Archer (2009, p. 51) report that there was not a significant amount of spelling variation anymore in the English language by the mid-seventeenth century, and thus it can be assumed that the spelling norms of the English language were becoming standardized around that time.

The research and development work revolving around the semantic analysis system has been a vast undertaking since the year 1990 in the UCREL research centre at Lancaster University. A very interesting spin-off has been the development of a historical semantic tagger for English. In January 2003, the UCREL team began to explore the feasibility of redirecting the EST so that it could be applicable to the study of historical texts dating from 1600 onwards (Archer et al., 2003). In the first phase, "historical" lexicons containing items peculiar to earlier periods of English were added to the existing EST which had originally been created for the analysis of modern English (Piao et al., 2004). This already improved the results. The lexical coverage obtained in the experiments ranged between 92.36% and 97.01%, and when applied on a 5-million-word corpus of 19th century fiction, the lexical coverage was as high as 97.29%. Since then the system has been further redirected to process Early Modern English texts in other respects but the vocabulary used as well by developing a new tool called Variant Detector (VARD) which acts as a pre-processor for text containing spelling variation (Baron & Rayson, 2009). In section 4.4.3.4.2, I will consider how these innovations could perhaps facilitate the processing of older Finnish text as well in which, similarly to English, both the spelling variants and the vocabulary have evolved over time. In addition, in the SAMUELS project, a semantic tagger has been developed which uses the

Historical Thesaurus of English (Historical Thesaurus of English, n.d.-a) presented in section 2.3.1.1.3 as its core dataset. This historical semantic tagger assigns each word in the input text the reference code which the *Historical Thesaurus of English* provides for the concept in question. (Alexander et al, 2015) UCREL has participated in the work, and the historical semantic tagger compliments the semantic tags used in the EST (see section 2.4.1.1) by offering finer-grained meaning distinctions for use in WSD.

In the fourth and last experiment, I examined how suitable the Finnish single word lexicon is for the analysis of texts collected from Internet discussions. The results obtained on the three test corpora are presented in Table 16 below:

Table 16

Final Evaluation: Texts from Internet Discussions

Test corpus	**Total words**	**Unmatched words**	**Lexical coverage(%)**
Death and Mourning	30,295	1,775	94.14
Niinistö or Haavisto	51,157	4,109	91.97
Separate Pool Times for Muslims	114,003	7,279	93.62

Again, this was a surprisingly good result considering the often more informal nature of the language used. Part of the texts in Internet discussions is written in Finnish which is correct in terms of vocabulary, grammar, spelling, and punctuation, but very commonly the texts also contain peculiar features which may cause problems for this type of software and semantic lexical resources that have been developed for the analysis of standard Finnish. Firstly, Finnish "Internet language" often resembles spoken colloquial language. By way of illustration, typical features of colloquial spoken Finnish are omission and assimilation of sounds as well as differences in form (Karlsson, 1999, pp. 245–248). Secondly, the

vocabulary used is often quite different, and the use of emoticons is very common in Internet discussions. Thirdly, such texts often contain plenty of misspellings and typographical errors, since the purpose has been to share something quickly instead of trying to produce "polished language". Errors appear both in texts written by native and by non-native users of Finnish. Despite these differences, the results prove that the Finnish single word lexicon functioned relatively successfully in this task and thus it could already be applied for the analysis of more informal writing contained on the Internet. In addition, the results can be further improved by training the FST and the semantic lexical resources to process the features mentioned above which differentiate "Internet language" from standard Finnish. This can be achieved by incorporating into it mechanisms similar to the mechanisms used in VARD. I will discuss these possibilities in more detail in sections 4.4.3.4.3, 4.4.3.4.4, and 5.3.3.

4.4 Application-Based Evaluation of Accuracy

I conducted the second set of experiments in order to measure the accuracy, that is to determine how well the single word lexicon performs when it is applied in the FST software. I tagged the test material automatically with the FST, after which I checked the output manually. The accuracy was calculated as the percentage of the semantic tags found to be correct from the total number of semantic tags in the automatically tagged text. If the output produced by the compound engine was incorrectly tagged, it was counted as one error, since compounds are single orthographical units. By comparison, if a MWE was tagged incorrectly, each of its constituent words, which are separated by spaces, was counted as one error. Moreover, as was the case with the formative evaluation described in section 4.2, I have not regarded the disregarding of the comparison of adjectives and adverbs as an error.

At present, the FST employs two disambiguation procedures:

1) POS tagging which takes place prior to semantic tagging and is carried out by TextMorfo and
2) general likelihood ranking which means that the senses in the lexicon entries have been arranged in perceived frequency order according to information obtained from dictionaries and intuition.

Thus, if a word has only one sense and if it is included in the single word lexicon, the semantic tag assigned by the FST should be the correct tag. In the case of an ambiguous word, the word is considered to be tagged correctly if the first semantic tag in the list, in other words, the tag indicating the most frequent sense of the word, is the correct tag in the given context.[73]

In the following section, I will present the test material which I selected for the application-based evaluation and, subsequently, I will report the results obtained from the experiments.

4.4.1 Test subsets

For the application-based evaluation of accuracy, I chose as test material subsets of the corpora which I had used for the final evaluation of lexical coverage (see section 4.3.1). In this evaluation as well, the aim was to cover different types of texts in terms of genre, domain, and historical period. The five subsets are all of approximately similar size but not precisely,

[73] In the Benedict project, a third disambiguation method was in use, namely the MWE component, in which case MWEs take precedence over single word expressions in the tagging process. However, as I noted in section 4.2, the MWE component does not function reliably in its present state and was thus omitted from this evaluation.

because I have used full sentences with these corpora rather than cutting the text at exactly 2,000 words. The test subsets include:

1) The Helsingin Sanomat Subset: four final sentences[74] from 51 news headlines and articles covering a wide variety of topics, such as equality, crimes, accidents, holidaymaking, furniture, clothing, human relations, sex, butchery, economy, politics, birds' nesting, fear of flying, sports, tattoos, breast feeding, injuries, and illnesses. In all, this subset consists of 2,009 words and represents general modern standard Finnish non-fiction in the evaluation.
2) The Kauniita Valheita Subset: the first 2,003 words from the Finnish-language translation *Kauniita valheita* (2007) by Tiina Sjelvgren from the English original *Beautiful Lies* written by Lisa Unger. This subset represents general modern Finnish fiction in the evaluation.
3) The Finnish Culinary Culture Subset: a random selection of sections from the report on Finnish culinary culture covering the topics of meals, table manners, food, aesthetics, agriculture, food processing, trade, privately owned restaurants, public food services, and recording knowledge about Finnish culinary culture. This subset consists of 2,014 words and represents domain-specific Finnish text in the evaluation.
4) The Hanna Subset: the first 2,004 words from the fictional novel *Hanna* written by Minna Canth in 1886. This subset represents older Finnish text in the evaluation.
5) The Separate Swimming Pool Times for Muslims Subset: the first 2,014 words from the beginning of the corpus of the 1,252 posted messages from the thread Muslimeille omat uintivuorot ("Separate Pool Times for Muslims") which I collected from the

[74] The final set of sentences contains exceptionally six sentences in order to have the total number of words in the test subset as close to 2,000 as possible.

online discussion forum of *Iltalehti*. This subset represents Finnish "Internet language" in the evaluation.

4.4.2 Results

The results of the experiments are presented in Table 17 below:

Table 17

Application-based Evaluation of Accuracy

Test subset	**Total words**	**Incorrectly tagged words**	**Accuracy(%)**	**Credible intervals(%)**[75]
Helsingin Sanomat (non-fiction)	2,012	348	82.70	81.0–84.3
Kauniita Valheita (fiction)	2,004	331	83.48	81.8–85.1
Finnish Culinary Culture (domain-specific)	2,009	400	80.09	78.3–81.8
Hanna (older Finnish)	2,004	343	82.93	81.2–84.4
Separate Pool Times for Muslims (Internet)	2,011	413	79.46	77.7–81.2

[75] Since these are based on small sub-samples from the corpora, this column indicates the 95% credible intervals for the results (assuming a uniform prior). Credible intervals are the Bayesian equivalent of confidence intervals and refer to the minimum and maximum values within which the true value of parameter must lie, with 95% degree of belief. For more information, see, for example, http://www.statsdirect.com/help/basics/confidence_interval.htm.

The accuracy ranged between 79.46% and 83.48%. By comparison, the accuracy in the formative evaluation at the end of the Benedict project on the subset of Finnish cooking texts was 83.08%. Even though the results of the formative evaluation and the results of the application-based evaluation are not directly comparable, as I pointed out in section 4.2, it is still possible to draw the conclusion that the accuracy had not improved. Thus, it is evident that even though the Finnish single word lexicon is significantly larger and better at present and is very useful in terms of lexical coverage, this alone is not sufficient to improve the accuracy of the FST.

By comparison, the accuracy of the EST has been calculated at 91.05%. It was tested on a corpus which contained approximately 124,900 words of transcriptions of 36 informal conversations, usually between two people in each case. (Rayson et al., 2004, pp. 10–11)

In order to acquire an insight into the errors which occurred, I identified 14 different types among them and classified them accordingly. I grouped the error types further into four major categories. These will be detailed in the following subsections.

The major category "Errors Related to the Semantic Lexical Resources" includes the four error types presented in Table 18 below. These are all errors which can be solved by developing the semantic lexical resources.

Table 18

Major Category: Errors Related to the Semantic Lexical Resources

Error type	**Percentage of all errors**
1) Errors Caused by Missing Words	18.91
2) Errors Caused by Missing Senses	3.00
3) Errors Caused by Existing MWE Templates Not in Use	3.60
4) Errors Caused by Missing MWE Templates	5.56
Total	31.07

By contrast, the major category "Errors Related to the FST Software" requires developing the software components of the FST. This major category comprises the four error types presented in Table 19 below:

Table 19

Major Category: Errors Related to the FST Software

Error type	**Percentage of all errors**
5) Wrong Order of Senses	33.02
6) Errors Caused by the Compound Engine	4.14
7) Errors Caused by Ellipsis in Compound Constructions	0.60
8) Wrong Semantic Tags for Ordinal Numbers	0.44
Total	38.20

The major category "Errors Related Both to the Semantic Lexical Resources and the FST Software" requires developing both of these components. It includes the two error types presented in Table 20 below:

Table 20

Major Category: Errors Related Both to the Semantic Lexical Resources and to the FST Software

Error type	**Percentage of all errors%**
9) Errors Caused by the Auxiliary Verb *olla* in Perfect and Pluperfect Constructions	9.37
10) Errors Caused by the Lack of Context Rules	5.29
Total	14.66

The final major category, "Other Error Types", contains the four error types presented in Table 21 below:

Table 21

Major Category: Other Error Types

Error type	**Percentage of all errors%**
11) Errors Caused by TextMorfo	8.17
12) Errors Caused by Archaic Use of Language	3.54
13) Errors Caused by Colloquial Use of Language	2.62
14) Spelling Errors in the Test Subset	1.74
Total	16.07

It is not possible to resolve these errors by developing the semantic lexical resources and the FST software, but they require other types of solutions.

By far the most frequently occurring single error type, constituting 33.02% of all errors, was "Wrong Order of Senses" which means that in case of an ambiguous word, the first semantic tag in the lexicon entry was not the correct tag in the given context. In addition, these and other errors which were related to the FST software put together constituted 38.20% of all errors, while the second largest major category, 31.07% of all errors, was "Errors Related to the Semantic Lexical Resources". I will present all the error types in more detail with various examples in the following subsections, along with suggestions for resolving them. I will primarily concentrate on the errors related to the semantic lexical resources which are related to the main focus of this thesis.

4.4.3 Analysis of the errors in the application-based evaluation

In general, it was straightforward to carry out the classification into the 14 error types which I identified in the results. In a few cases, however, one error could be classified into two different types. An example of such a case is the noun *ruoanlaitto* ("food preparation") in the sentence *Suomalainen ruoanlaitto- ja ruokaosaaminen ulottuu kodeista teollisuuslaitoksiin ja tutkimuksesta ruokajärjestelmiin*. ("Finnish food preparation and food knowledge extends from the home to industrial establishments and from research to food systems."). Firstly, the noun *ruoanlaitto* belongs in the compound construction *ruoanlaitto-osaaminen* which the FST is not yet able to process correctly because of ellipsis[76]. This represents error type 7, "Errors Caused by Ellipsis in Compound Constructions" (see section 4.4.3.2.3). Secondly, the noun *ruoanlaitto* was processed incorrectly by the compound engine.

[76] For more information on ellipsis in compound constructions, see section 2.5.1.2.

Thus, this noun could also be categorized under the error type 6, "Errors Caused by the Compound Engine" (see section 4.4.3.2.2). I decided to classify this error under the error type 6, since I considered errors caused by the compound engine more serious than errors caused by ellipsis in compound constructions.

4.4.3.1 Errors related to the semantic lexical resources

In all, 31.07% of the errors encountered in the application-based evaluation were related to the semantic lexical resources. In the following subsections, I will analyze and discuss the four error types which belong in this major category.

4.4.3.1.1 Errors caused by missing single words

The first error type, "Errors Caused by Missing Single Words", constituted 18.91% of all errors encountered in the application-based evaluation. Table 22 below displays the number and the percentage of the errors occurred in each test subset. The second last column in the right displays the total number of errors in all the test subsets. The last column in the right displays their percentage of the total number of all errors which occurred in the application-based evaluation.

Table 22

Application-Based Evaluation: Errors Caused by Missing Single Words

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	83	54	72	63	75	347	18.91
Percentage of errors in this error type	23.92	15.56	20.75	18.96	21.61		

The single words which I classified into this error type included both general language and domain-specific vocabulary, and there was also a considerable number of personal, geographical, and other proper names. In addition, several words in other languages appeared; these were mostly English. However, not all single words which I discovered to be missing in the evaluation were included in this error type, but I made two exceptions in my classification. Firstly, such missing single words which I considered to represent historical features of language, I classified under the error type 12, "Errors Caused by Archaic Use of Language" (see section 4.4.3.4.2 for examples). Secondly, such missing single words which I considered to represent colloquial use of language, I classified under the error type 13, "Errors Caused by Colloquial Use of Language" (see section 4.4.3.4.3 for examples). The reason for this decision was that such words do not represent modern standard Finnish language for

which the semantic lexical resources are developed. Rather, their treatment requires other types of solutions which will be addressed in their respective subsections.

The number of missing single words was relatively similar across all test subsets, even though the text types which the test subsets represented were quite different from each other. For example, in the Helsingin Sanomat Subset, the proportion of missing personal, geographical, and other proper names was significantly larger than in the other test subsets. This outcome could be expected, because the Helsingin Sanomat Subset contained a collection of short sections from various news headlines and articles, and personal, geographical, and other proper names occur frequently in this type of text. The Finnish Culinary Culture Subset, in turn, contained a fair amount of vocabulary particular to that specific domain and unfamiliar to general language. Since the semantic lexicons discussed in this thesis are built primarily into a general language resource, it could also be expected that some of this domain-specific vocabulary remains unrecognized.

Even though the results of the formative evaluation at the end of the Benedict project and the results of this application-based evaluation are not directly comparable, as I noted in section 4.2, it is still possible to draw the conclusion that the new, expanded, and improved single word lexicon indeed covers a wider vocabulary than the old version and produces better results when it is applied in the FST. In the formative evaluation, the percentage of errors caused by missing single words was 38.83%, whereas in this application-based evaluation, the percentage of errors caused by missing single words was only 18.91%.

The single word lexicon can be further improved by adding new entries into it. Nevertheless, since the semantic lexical resources are intended to be a general language resource, the purpose is not to try to include all the vocabulary in the language exhaustively; this would only result in an unmanageable lexicon size. Instead, the primary aim is to include only the core vocabulary of the language as well as at least relatively frequently appearing

personal, geographical, and other proper names. Should there be a need to process domain-specific text, the semantic lexicons can be tailored to that task by expanding them with vocabulary relevant to that particular domain. In such a case, it might also be helpful to expand the semantic tagset by setting up new categories or dividing the existing categories into further subcategories. This can be done easily because of the flexibility of the USAS semantic category system. In case there is a need to recognize a larger number of proper nouns, for example, in named entity recognition tasks, it would be reasonably easy to expand the semantic lexicons by utilizing gazetteers, place name dictionaries, or other similar lists, as well as Wikipedia. The further development of the Finnish semantic lexical resources will be discussed in more detail chapter five.

On the basis of the results of the application-based evaluation, I have added, for example, the following new entries into the single word lexicon:

adrenaliini	Noun	B1/O1
		("adrenaline")
juopporetku	Noun	F2+++/S2
		("boozer")
Niinimaa	Proper	Z1
		(a Finnish family name)
pk-yritys	Noun	I2.1/S5+
		("small or medium-sized enterprise")
reumaatikko	Noun	B2-/S2
		("rheumatic" (noun))
surkimus	Noun	X9.2-/S2
		("lame duck")

It would be a practical idea to supplement the single word lexicon by also including derivations of those words which are added, if there are any. For example, the verb *anella* ("to plead") was found to be missing in this evaluation. Thus, in addition to the verb *anella*, I also added the nouns *aneleminen* ("pleading"), *anelija* ("pleader"), and *anelu* ("pleading") which are derived from this verb. The new single word lexicon entries thus are:

aneleminen	Noun	Q2.2 ("pleading")
anelija	Noun	Q2.2/S2 ("pleader")
anella	Verb	Q2.2 ("to plead")
anelu	Noun	Q2.2 ("pleading")

Likewise, the adjective *synkeä* ("gloomy")**,** which was also discovered to be missing in this evaluation, is another valuable addition to the single word lexicon. In addition to the adjective *synkeä*, I also added the adverb *synkeästi* ("gloomily") and the noun *synkeys* ("gloominess") which are its derivations. The new single word lexicon entries thus are:

synkeys	Noun	E4.1-	W2- ("gloominess")
synkeä	Adjective	E4.1-	W2- ("gloomy")
synkeästi	Adverb	E4.1-	W2- ("gloomily")

There was a considerable number of single words classified in the error type "Errors Caused by Missing Single Words" that I do not consider useful additions to this type of a general language resource. I base my decisions on dictionaries and on my work experience as a lexicographer. Such words are, for example, rarely occurring words, such as *isolationismi* ("isolationism"). It would neither be a practical idea to include rarely occurring proper names, such as *Aapraham* (old Finnish male name; a form of Abraham) and *Terra* (the name of a dog in a newspaper article).

Another frequent source of error in this error type were compounds for which the slash tag generated by the compound engine (see section 3.3.2) was not correct. Examples of such compounds are *premiumtuote* ("premium product") and *rusketuspakko* ("tanning compulsion"). In my opinion, such infrequent compounds of a more temporary nature would not be relevant additions to the single word lexicon. Nevertheless, according to my observations, relatively often the compound engine had functioned successfully and had managed to produce the correct interpretation for such compounds.

Finally, the single word lexicon and the TextMorfo lexicon should be expanded concurrently. This is necessary, because if a word is added to the single word lexicon but it does not exist in the TextMorfo lexicon, TextMorfo does not recognize this word, and this automatically results in the semantic tag Z99 for this word even if it existed in the single word lexicon. For example, the words *blogi* ("blog") and *deli* ("deli") were such words in this evaluation which were missing both from the single word lexicon and from the TextMorfo lexicon. It would also be very sensible to make a file comparison using, for example, the Unix diff command between the present versions of the single word lexicon and the TextMorfo lexicon to find missing words and to make the lexicons correspond.

4.4.3.1.2 Errors caused by missing senses

The second type of errors related to the semantic lexical resources was "Errors Caused by Missing Senses". This type accounted for 3.00% of all errors encountered in the application-based evaluation, as is evident from Table 23 below:

Table 23

Application-Based Evaluation: Errors Caused by Missing Senses

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	12	7	13	14	9	55	3.00
Percentage of errors in this error type	21.81	12.73	23.64	25.45	16.36		

The highest number of missing senses, 14 in all, was discovered in the Hanna Subset. Five of them were due to the fact that the single word lexicon entry for the noun *lamppu* ("lamp") included only the semantic tag O3/W2 which indicates an electric lamp. In this classic novel from the year 1886, however, lamps functioned with oil. Thus, in this case, the correct semantic tag would have been O2/W2.

As is the case with the missing words, it would not be a practical solution to try to exhaustively include all possible senses of the words in their lexicon entries. However, it

would be sensible to include the at least relatively frequently occurring senses and develop effective disambiguation mechanisms which could help in determining the relevant sense in the given context. That said, I regarded all cases of missing senses which I discovered in this evaluation as useful additions to the existing single word lexicon entries. By way of illustration, I added the following underlined senses into their respective lexicon entries:

kpl	Abbrev	N5	Q4.1 ("paragraph")
kuve	Noun	M6 ("side")	B1
kaveri	Noun	S3.1/S2	S2.2m ("chap")
lamppu	Noun	O3/W2	O2/W2 ("lamp other than electric")

4.4.3.1.3 Errors caused by existing multiword expression templates not in use

While the previous two error types dealt with single words, this and the following error type deal with MWEs[77].

The errors in the type "Errors Caused by Existing Multiword Expression Templates Not in Use" were caused by the fact that the MWE lexicon was not used at all in this evaluation. This was due to the fact that the "quick MWE template solution", which was generated for the MWE lexicon in the Benedict project, was found to be neither useful nor intelligent (see section 4.2). Therefore, it was omitted completely from this evaluation, and only the single word lexicon was used. In point of fact, this error type is more related to the FST software component than to the semantic lexical resources. However, it was necessary to place it here

[77] It was estimated that 16% of words in English running text are semantic MWEs (Rayson, 2005, p. 4). Unfortunately, corresponding information is not available for Finnish. MWEs are common in Finnish but not as common as in English, since English also contains an abundance of noun phrases the equivalent of which in Finnish would be written as single orthographic compound words.

to help put the following subsection into context. This type caused 3.60% of all errors encountered in the application-based evaluation, as Table 24 below shows:

Table 24

Application-Based Evaluation:

Errors Caused by Existing MWE Templates Not in Use

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	14	14	14	18	6	66	3.60
Percentage of errors in this error type	21.21	21.21	21.21	27.27	9.09		

In case the first semantic tags assigned to the words which constituted a given MWE were all incorrect, I counted all the constituent words as errors. An example of such a case is the idiom *maksaa vaivan* ("to be worth it"; literally "to pay for the hardship") from the Helsingin Sanomat Subset. The FST assigned it the following semantic tags:

pos="Verb" mwe="0" sem="I1.2 I1.3" lem="maksaa">maksaa

pos="Noun" mwe="0" sem="A12- B2- E4.1-" lem="vaiva">vaivan

Nevertheless, the correct semantic tag for the whole expression would be A1.5.2+, as can be found in the MWE lexicon. Thus, this MWE counts as two errors. By comparison, I counted, for example, the MWE *nostaa syyte* ("to press charges"; literally "to lift a charge") as one error. The correct semantic tag for the whole expression would be G2.1, as can be found in the MWE lexicon. Since the other of the words which constitute this MWE has received this correct semantic tag, I have considered the following output as one error:

pos="Verb" mwe="0" sem="M2 N5+/A2.2 A5.1+/A2.2 A2.2 I1/A9+" lem="nostaa">nostaa

pos="Noun" mwe="0" sem="G2.1" lem="syyte">syyte

Above I wrote that this error type contains MWEs which are included as "quick MWE template solutions" in the MWE lexicon but which remained unrecognized, since the MWE lexicon was not in use. In all, 66 errors were classified as belonging in this error type. However, even if the MWE lexicon in its present state had been in use, all of these MWEs would still not have been recognized in the test subsets. Such MWEs in which the constituent words appear consecutively should have been recognized. An example of this was the expression *poissa tolaltasi* ("[you are] upset"; literally "[you are] away from your trail"):

pos="Adverb" mwe="0" sem="M6/A6.1-" lem="poissa">poissa

pos="Noun" mwe="0" sem="Z99" lem="tola">tolaltasi[78]

However, relatively often the constituents of a MWE do not appear consecutively in text, but MWEs may be discontinuous, in other words, there may be embedded elements between the constituents of a MWE. This phenomenon occurs for the most part in verb phrases. The

[78] The word *tola* is archaic and therefore it is not included in the single word lexicon. However, it is commonly used in this expression which is included in the MWE lexicon.

templates which are created with the "quick MWE template solution" do not recognize discontinuous MWEs. An example of such a case is the sentence *Olen tällaisilla perusteilla annettuja vuoroja vastaan*. ("I object to turns distributed on such grounds"; literally "I am on such grounds distributed turns against.") from the Separate Pool Times for Muslims Subset. It includes the phrasal verb *olla vastaan* ("to object"; literally "to be against") and in all four embedded elements between the constituents as can be seen in the following FST output:

```
pos="Verb" sem="A3+ A1.1.1 M6 A8" lem="olla">olen
pos="Adjective" sem="Z5" lem="tällainen">tällaisilla
pos="Noun" sem="A2.2" lem="peruste">perusteilla
pos="Verb" sem="A9- A2.2 A1.1.1 A10+ S7.4+ S3.2" lem="antaa">annettuja
pos="Noun" sem="N4 T1.3/I3.1 M3 M4 M5" lem="vuoro">vuoroja
pos="Preposition" sem="Z5" lem="vastaan">vastaan
```

Hence, even if the present MWE lexicon had been in use, 19 errors of the total of 66 errors representing this error type would still have remained, since those MWEs would not have been recognized because of embedded elements. Furthermore, as I noted in section 4.2, the "quick MWE template solution" caused a considerable amount of errors and confusion. Thus, even if it had managed to produce correct outputs a few times, these would presumably have been outnumbered by the errors. Indeed, a new, more useful, and intelligent solution for creating Finnish MWE templates needs to be developed to be able to successfully identify and tag Finnish MWEs. I will draft guidelines for carrying out this task in section 5.2.2.2.

4.4.3.1.4 Errors caused by missing multiword expression templates

In all, 5.56% of all errors in the application-based evaluation were caused by missing MWE templates, as Table 25 below illustrates:

Table 25

Application-Based Evaluation: Missing MWE Templates

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	24	26	17	24	11	102	5.56
Percentage of errors in this error type	23.53	25.49	16.67	23.53	10.78		

As was the case with the error type "Errors Caused by Existing Multiword Expression Templates Not in Use" discussed above, I counted here as well missing MWE templates as multiple errors if none of the first semantic tags assigned for the constituent words was the correct semantic tag. For instance, I considered the MWE *nähdä sielunsa silmin* ("to see in the mind's eye"; literally "to see with one's soul's eyes"), for which the correct semantic tag would be X2.1, as three errors:

pos="Verb" mwe="0" sem="X3.4 X2.5+ S3.1 X2.6+" lem="nähdä">näki

pos="Noun" mwe="0" sem="S9 S2" lem="sielu">sielunsa

pos="Noun" mwe="0" sem="B1 X3.4 X2.5+ O2 W3 C1 O4.3" lem="silmä">silmin

By comparison, I considered as one error, for instance, the MWE *tarttua kiinni* ("to grab", "to stick to"; literally "to grab hold of", "to stick to") in which the other of the first semantic tags assigned (A1.7+) is the correct semantic tag for the whole expression:

pos="Verb" mwe="0" sem="A1.7+ A1.1.1 S1.1.3+ B2- N5+/A2.1 Y2" lem="tarttua">tarttui

pos="Adverb" mwe="0" sem="A10- A1.7+ N3.3-" lem="kiinni">kiinni

Many MWEs found to be missing in the evaluation would be valuable additions to the Finnish MWE lexicon. In addition to *nähdä sielunsa silmin* and *tarttua kiinni* in the above examples, I would include, for instance, the following MWEs: *ehdoton vankeus* ("unconditional imprisonment"), *kestävä kehitys* ("sustainable development"), *käydä keskustelua* ("to debate"), *neljän seinän sisällä* ("indoors"), *olla mielessä* ("to have [something] in mind"), and *vanha kunnon* ("good old"). In contrast, I would not consider it useful to add very rarely appearing MWEs in the lexicon, for example, Little Angels, the name of a hospital in the USA, which appeared in the Kauniita Valheita Subset.

Once worthwhile additions to the MWE lexicon have been collected, they need to be written into MWE templates according to the guidelines presented in section 5.2.2.2 and assigned the relevant semantic tags.

4.4.3.2 Errors related to the FST software

In all, 38.20% of the errors in the application-based evaluation were discovered to be related to the FST software. This is the largest major category of errors, which indicates that

in addition to improving the semantic lexical resources it is even more necessary to invest in the development of the software component. The four error types which belong in this major category will be discussed in the following subsections.

4.4.3.2.1 Errors caused by wrong order of senses

By far the largest single error type in the application-based evaluation was "Errors Caused by Wrong Order of Senses" which constituted 33.02% of all errors encountered, as Table 26 below illustrates. This error type contains ambiguous words 1) for which the first semantic tag listed in the lexicon entry is not the correct semantic tag but one of the other semantic tags listed[79] and 2) the correct sense of which is not identifiable with the aid of context rules. By contrast, the type of errors which can be resolved by developing context rules are classified into the error type "Errors Caused by the Lack of Context Rules". These will be discussed in section 4.4.3.3.2.

[79] The semantic tags are organized in perceived frequency order; see section 3.4.

Table 26

Application-Based Evaluation: Wrong Order of Senses

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	127	102	162	75	140	606	33.02
Percentage of errors in this error type	20.96	16.83	26.73	12.38	23.10		

Various approaches could be adopted to resolve these errors. By way of illustration, procedures number 4, 5, and 7, which the EST utilizes for the task of semantic tag disambiguation and which were described in section 2.4.1.3, would be applicable to the disambiguation of Finnish as well. Some suggestions based on these procedures will be outlined in the following paragraphs.

Applying procedure number 4, one possible approach to improve disambiguation mechanisms would be to take into account the domain of discourse. If the domain of discourse in a given text was known beforehand, this information could be exploited to "weight" tags, in other words, to alter the order of semantic tags in the semantic lexical resources for a particular domain. This type of function has successfully been utilized in connection with the EST. In this application-based evaluation, such a function would have

been particularly beneficial when tagging the Finnish Culinary Culture Subset in which the largest number of errors of this error type occurred. By way of illustration, the following words caused a total of 16 errors, since the first semantic tag in the lexicon entry was not the correct tag for the given sense in this test subset: *annos* (1 error), *keittiö* (3 errors), *nauttia* (7 errors), and *resepti* (5 errors). When processing this type of domain-specific text, it would be a practical approach to weight the semantic tags F1 (Food) and F2 (Drinks) to be able to identify the correct senses:

annos	Noun	N5	F1 ("portion")
keittiö	Noun	H2	F1 ("cuisine")
nauttia	Verb	E4.2+	F1/B1 ("to eat")
		F2/B1 ("to drink") A9+	
resepti	Noun	B3/Q1.2	F1 ("recipe")

The second possible approach to resolve the errors caused by the wrong order of senses would be text-based disambiguation which was discussed in connection with procedure number 5 of the EST. This procedure resembles procedure number 4 described above, with the exception that, while in procedure number 4 the weighting is adjusted manually, in this approach, the weighting is decided by the program. This approach, which has not been implemented in the EST yet, is based on the hypothesis formulated by Gale et al. (1992, pp. 233–237). According to their findings, well-written discourses tend to avoid multiple senses of polysemous words. Indeed, their experiments revealed that this tendency was as strong as 98%.

The third possible approach, which could be adopted for resolving the errors of this type, would be local probabilistic disambiguation which was discussed in connection with procedure number 7. According to this procedure, it can be generally supposed that the local

surrounding context determines the correct semantic tag for a given word. The surrounding context can be identified in terms of 1) the words themselves, 2) their POS tags, 3) their semantic tags, or 4) some combination of all three. While procedures number 4 and 5, which were discussed in the previous two paragraphs, are applied to longer stretches of text, procedure number 7 is applied on sentence level.

In fact, a prototype of such a tool was developed in the Benedict project (Löfberg et al., 2004; also see section 1.1). Our aim was to enable context-sensitive dictionary lookups by identifying the very sense of the word which the user is looking for in case there are more senses than one in the dictionary entry. This is a hybrid tool which uses many statistical and rule-based components. The preliminary test results in the Benedict project were encouraging, and similar mechanisms could be used in the future development of the FST software as well. This approach together with other disambiguation mechanisms has also been used in the SAMUELS project to develop a historical semantic tagger which utilizes the *HTOED* as its knowledge base to provide a uniquely fine-grained semantic classification (Alexander et al., 2015).

Finally, it must be noted that the order of senses listed as tags is not always essential. In many applications, such as in information retrieval setting, where only certain features in text need to be recognized, it is not necessary to disambiguate between the senses of ambiguous words in a given context. Instead, it is sufficient that the semantic tag for the relevant sense is included among the semantic tags in the lexicon entry. This was the case, for example, in the Metaphor in End-of-Life Care (MELC) project which investigated the use of violence metaphors for cancer and end-of-life among patients, family carers, and healthcare professionals (Demmen et al, 2015). For this purpose, the project team applied a computer-assisted approach to the analysis of metaphor variation across genres by utilizing a selection of semantic categories of the USAS system which were relevant for the task, such as E3-

("Violent/Angry"), G3 ("Warfare, Defence, and the Army; Weapons"), M4 ("Movement/Transportation: Water"), and W4 ("Weather") (Semino, Hardie, Koller, & Rayson, 2005). Another application of this type could be a semi-automatic Internet monitoring program for which I draft guidelines in section 5.3.2. To evaluate the performance of the FST in its present state for such applications, I calculated another value in addition to the accuracy mentioned above. This value is referred to as "fuzzy accuracy". Fuzzy accuracy has been calculated in the same way as accuracy, with the exception that I have ignored the error type "Wrong Order of Senses", in other words, I have considered the word correctly tagged if any of the semantic tags listed in the lexicon entry is the correct tag. Table 27 below displays the results:

Table 27

Application-Based Evaluation: Fuzzy Accuracy

Test subset	**Fuzzy accuracy(%)**
Helsingin Sanomat (non-fiction)	89.02
Kauniita Valheita (fiction)	88.57
Finnish Culinary Culture (domain-specific)	88.15
Hanna (older Finnish)	86.63
Separate Pool Times for Muslims (Internet)	86.42

When compared to Table 17 in section 4.4.2, which displays the overall results of the application-based evaluation of accuracy, it is evident that the disregard of the order of senses had a significant effect on the results[80].

4.4.3.2.2 Errors caused by the compound engine

Another component in the present FST software which would need substantial improvement is the compound engine described in section 3.3.2. The compound engine is expected to function in the following way. When the FST software discovers a word in the input text which does not exist in the single word lexicon and which is not identified as a constituent of a MWE, it should pass that word to the compound engine. The compound engine should then check if the word is possibly a compound consisting of two words. If this was discovered to be the case, the compound engine should assign the relevant semantic tags for both constituents of the compound. Thereafter, the compound engine should combine the semantic tags of the compound constituents automatically and separate them with a slash in the output. The resulting semantic tags resemble the slash tags which were discussed in section 2.4.1.1.

The compound engine processes many compounds in this way successfully, and it is indeed a useful tool because of the infinite number of possible compounds in the Finnish language, but it does not function reliably. It very often happens that the FST software directly passes compounds to the compound engine, without checking first if the compound is included in the single word lexicon. As a result, the compound engine splits compounds and tags the compound parts separately also in such cases when the compounds do exist in the single word lexicon and thus should not be passed on to the compound engine at all. Instead,

[80] In fact, the percentages would actually be somewhat higher, if such errors were counted including the type "Errors Caused by the Lack of Context Rules" in which the lexicon entry contains the correct semantic tag.

these compounds should be tagged according to the information which is contained in the respective single word lexicon entry. Occasionally, such output produced by the compound engine is nevertheless correct despite this bug. In most cases, however, this results in errors, especially in case of lexicalized compounds (see section 2.5.1.2) in which the meaning of the compound cannot easily be deduced from the sum of the meanings of the compound constituents. Such errors are classified into this error type.

An example of an erroneously processed compound found in the Finnish Culinary Culture Subset is *päiväkoti* ("day-care centre"; literally "dayhome", below in the inessive singular) for which the compound engine produced the following interpretation in the FST output:

pos="Noun/Noun" mwe="com" sem="H4/H1/T1.3" lem="koti/päivä">päiväkodissa

An entry for the compound *päiväkoti* does exist in the single word lexicon:

päiväkoti Noun S8+/S4/M7

Regardless, the compound engine had split this compound and tagged the constituents separately resulting in an error.

The error type "Errors Caused by the Compound Engine" constituted 4.14% of all errors encountered in the application-based evaluation, as is evident from Table 28 below:

Table 28

Application-Based Evaluation: Errors Caused by the Compound Engine

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	Percentage **of all errors**
Number of errors	13	14	30	2	17	76	4.14
Percentage of errors in this error type	17.11	18.42	39.47	2.63	22.37		

By far the largest number of errors occurred in the Finnish Culinary Culture Subset. This was due to the fact that it contained more compounds that the other test subsets. For instance, in all four errors were caused by the erroneously processed compound *raaka-aine* ("raw material"), and three errors were caused by the erroneously processed compound *ruoanlaitto* ("cooking").

The compound engine processed by far the majority of the compounds in the test subsets in such an erroneous manner. Surprisingly, however, a few times it did not split the compound but tagged it successfully according to the information included in the single word lexicon entry. An example of this encountered in the Finnish Culinary Culture Subset is the compound *jälkiruoka* ("dessert") which the software recognized as a single word lexicon entry and, consequently, tagged it correctly as F1:

pos="Noun" mwe="0" sem="F1" lem="jälkiruoka">jälkiruoka

This is a compound consisting of two constituents of which the latter is *ruoka* ("food"). Its lexicon entry contains exactly the same grammatical and semantic information as the lexicon entries for the compounds *pääruoka* ("main course") and *eturuoka* ("starter"), the latter constituent of which is also *ruoka*:

eturuoka	Noun	F1
jälkiruoka	Noun	F1
pääruoka	Noun	F1

Nevertheless, the software passed the compounds *pääruoka* and *eturuoka* on to the compound engine without checking the single word lexicon first. Naturally, this resulted in erroneous outputs as is evident from the following FST outputs[81]:

pos="Noun/Noun" mwe="com" sem="F1/B1 F1/S2 F1/L2 F1/S7.1+/S2" lem="ruoka/pää">pääruoan

pos="Noun/Noun" mwe="com" sem="F1/A5.1+ F1/S7.4+ F1/I1.1+" lem="ruoka/etu">eturuokia

Thus, this is not an error caused by the single word lexicon, but it is related to the software. The bug which causes this must be detected and corrected, since it causes many unnecessary mistakes.

[81] The compound *pääruoka* is in the genitive singular and the compound *eturuoka* is in the partitive plural.

4.4.3.2.3 Errors caused by ellipsis in compound constructions

Sometimes one or more of the compound constituents either at the beginning or at the end of the compound can be left out and replaced by a hyphen for abbreviation purposes (Hakulinen, 2004, p. 420). This phenomenon is referred to as ellipsis, and it was discussed in section 2.5.1.2. An example of an elliptic compound construction encountered in the Finnish Culinary Culture Subset was the construction *maito- ja leipäkauppojen* (literally "of milk and bread shops", here in the genitive plural) which is a truncated form of the words *maitokauppojen ja leipäkauppojen* ("of milk shops and of bread shops"). Since the FST is not able to recognize ellipsis in compound constructions yet, it tags the construction incorrectly as consisting of the words *maito* ("milk") and *leipäkauppa* ("breadshop"). In other words, it misses the second constituent *kauppa* ("shop") of the compound *maitokauppa* ("milkshop") completely by tagging only the first constituent *maito* ("milk"):

pos="Noun" mwe="0" sem="F2 O1.2" lem="maito">maito

pos="_Delimiter" mwe="0" sem="PUNC" lem="-">-

pos="Conjunction" mwe="0" sem="Z5" lem="ja">ja

pos="Noun/Noun" mwe="com" sem="I2.2/F1 I2.2/I1.1 I2.2/H1/F1 I2.2/H1/I1.1"

lem="kauppa/leipä">leipäkauppojen

In all, 0.60% of all errors in the application-based evaluation were caused by this phenomenon, as is evident from Table 29 below:

Table 29

Application-Based Evaluation:

Errors Caused by Ellipsis in Compound Constructions

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	0	0	8	0	3	11	0.60
Percentage of errors in this error type	0.00	0.00	72.73	0.00	27.27		

The FST software should be updated to recognize such elliptic compound constructions and replace them with equivalent non-elliptic compound constructions, which should result in the correct semantic interpretation. That said, this type of errors are neither very common nor are they very serious. Nonetheless, correcting them would make the FST more intelligent.

4.4.3.2.4 Errors caused by wrong semantic tags for ordinal numbers

The error type "Errors Caused by Wrong Semantic Tags for Ordinal Numbers" was caused by the fact that the FST software had tagged ordinal numbers incorrectly. These errors constituted 0.44% of all errors encountered in the application-based evaluation, as Table 30 below illustrates:

Table 30

Application-Based Evaluation:

Errors Caused by Wrong Semantic Tags for Ordinal Numbers

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	3	1	3	1	0	8	0.44
Percentage of errors in this error type	37.50	12.50	37.50	12.50	0.00		

Let us take the ordinal number *ensimmäinen* ("the first") as an example. The FST software had tagged it erroneously as N1 ("Numbers"):

pos="Numeral" mwe="0" sem="N1" lem="ensimmäinen">ensimmäinen

The correct semantic tag, however, would be N4 ("Linear Order") as can be seen from the single word lexicon entry for *ensimmäinen*:

ensimmäinen Numeral N4

Such errors are due to the fact that the TextMorfo component, which carries out the grammatical analysis before the semantic tagging, does not make a distinction between ordinal and cardinal numbers but automatically classifies all numerals as cardinal numbers, which leads to the wrong semantic tag N1 irrespective of differing information in the semantic lexicon entry. This bug should be fixed as well to improve the performance of the FST software. Similarly, the FST software should be taught to recognize digits when they indicate linear order and automatically tag them as N4. Ordinal numbers are indicated by a full stop in Finnish, for example, *1. helmikuuta* ("February 1").

4.4.3.3 Errors related to both the semantic lexical resources and the FST software

The major category "Errors Related to Both the Semantic Lexical Resources and the FST Software" comprises two error types. These are: 1) errors which have been caused by the fact that the FST software does not recognize the auxiliary uses of the verb *olla* ("to be") and 2) errors which have been caused by the fact that there are no context rules implemented in the FST software yet. Solutions for resolving these two error types involve work on both the semantic lexical resources and on the software component.

4.4.3.3.1 Errors caused by the auxiliary verb olla in perfect and pluperfect constructions

The perfect tense in Finnish is formed by using the present tense of the auxiliary verb *olla* ("to be") which is followed by the past participle (e.g. *on leiponut* ("has baked")), and the pluperfect tense is formed by using the past tense of the auxiliary verb *olla* which is followed by the past participle (e.g. *oli leiponut* ("had baked")). Both the verb *olla* and the present or past participle can appear in an inflected form. Unlike with the EST, there are not yet

mechanisms in the FST which are able to recognize cases of the auxiliary use of the verb *olla*. This phenomenon caused 9.37% of all errors encountered in the application-based evaluation, as Table 31 below reveals:

Table 31

Application-Based Evaluation:

Errors Caused by the Auxiliary Verb olla *in Perfect and Pluperfect Constructions*

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	35	70	40	19	8	172	9.37
Percentage of errors in this error type	20.35	40.70	23.26	11.05	4.65		

By far the largest number of errors classified in this error type was detected in the Kauniita Valheita Subset. This was caused by the fact that the subset was taken from the beginning of a novel, and this section contained mostly narration of past events utilizing perfect and pluperfect constructions. Because of the lack of necessary disambiguation mechanisms, the FST software erroneously tagged all the auxiliary uses of the verb olla, for instance, in *oli leikitellyt* ("had played") incorrectly in the following manner:

pos="Verb" mwe="0" sem="A3+ A1.1.1 M6 A8" lem="olla">oli

pos="Verb" mwe="0" sem="E4.1+" lem="leikitellä">leikitellyt

A component which recognizes auxiliary uses of the verb *olla* should be developed in the FST, utilizing the mechanisms of the EST as a model. The correct semantic tag for an auxiliary verb would be Z5 ("Grammatical Bin"). This type of a function would improve the results significantly, since unrecognized auxiliary uses of the verb *olla* were a very common source of error in this evaluation.

4.4.3.3.2 Errors caused by the lack of context rules

The errors of this type were caused by the fact there are no context rules included in the FST yet which would help the program to identify the correct sense of an ambiguous word in a given context. Such context rules have been applied successfully in the EST (see procedure number 6 in section 2.4.1.3).

In cases where a word has more senses than one, the semantic tags for the different senses have been listed in the lexicon entry in perceived frequency order. If the first semantic tag is the relevant tag in the given context, the FST should be able to tag that word correctly. However, if one of the other semantic tags is the relevant tag, the word turns up misinterpreted.

In some cases, the given sense of a word is used only with a particular collocation or with a particular grammatical pattern or inflection. In these cases, it would be possible to make use of the collocational and grammatical information contained in the TextMorfo output in order to write rules resembling the rules which are used in the MWE templates to create context rules for the FST software to enable it to recognize these senses. I have grouped such potential context rule candidates which in this evaluation turned up misinterpreted into this

error type. By comparison, in section 4.4.3.2.1, I discussed the error type "Errors Caused by Wrong Order of Senses". In it I have classified words which have also turned up misinterpreted because of the fact that the first semantic tag in the lexicon entry was not correct. However, they represent cases in which context rules would not be of help, since the given senses do not appear with particular collocations, inflections, or grammatical patterns. Hence, resolving the errors included in the type "Errors Caused by Wrong Order of Senses" requires different types of solutions.

To make the idea clearer, let us look at the noun *luku* as an example. The lexicon entry for this noun contains the following semantic tags:

luku	Noun	N1	Q1.2	Q4.1
	Q3	P1		

The semantic tag N1 stands for the category "Numbers", Q1.2 for "Paper Documents and Writing", Q4.1 for "The Media: Books", Q3 for "Language, Speech, and Grammar", and P1 for "Education in General". However, when this noun is combined with a year with a hyphen, for instance, in the expression *1770-luku* ("the 1770s"), *luku* means a period of time. In such a case, the correct semantic tag would be T1.3. Thus, the noun *luku* used alone would not receive the semantic tag T1.3, but only when it is combined with a year with a hyphen. Since the FST software does not have this information yet, it now tags the expression *1770-luku* as three separate units as follows, resulting in an error:

```
<w pos="Numeral" mwe="0" sem="N1" lem="1770">1770</w>
<w pos="_Delimiter" mwe="0" sem="PUNC" lem="-">-</w>
<w pos="Noun" mwe="0" sem="N1 Q1.2 Q4.1 Q3 P1" lem="luku">luku</w>
```

If a context rule was employed in the FST software, it could enable the FST to tag the entire expression *1770-luku* as well as other similar expressions, such as *2000-luku* and *60-luku*, correctly as T1.3. Such a context rule could be included in a context rule list. Alternatively, this information could be included in an autotagging lexicon, since these are fixed patterns which can have many possible instantiations and which could be tagged effectively through the use of wild cards. The autotagging lexicon included in the EST was discussed in section 2.4.1.2, and guidelines for creating an autotagging lexicon for Finnish will be drafted in section 5.2.3.

The error type "Errors Caused by the Lack of Context Rules" caused 5.29% of all errors encountered in the application-based evaluation, as is evident from Table 32 below:

Table 32

Application-Based Evaluation:

Errors Caused by the Lack of Context Rules

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	21	13	24	17	22	97	5.29
Percentage of errors in this error type	21.65	13.40	24.74	17.53	22.68		

The highest number of these errors occurred in the Finnish Culinary Culture Subset. Much of the text was description of past events and developments, and the noun *luku* combined with a year with a hyphen indicating a period of time appeared a total of 12 times. As could be expected, all these expressions turned up incorrectly tagged.

There were also many other cases in this evaluation in which the correct senses could have been identified through the aid of context rules. By way of illustration, when the verbs *näyttää* (3 different semantic tags listed in the single word lexicon entry) and *vaikuttaa* (5 different semantic tags listed in the single word lexicon entry) are followed by an adjective with the ablative case ending *-lta*/*-ltä*, the sense "to seem" or "to appear" represented by the semantic tag A8 is the correct sense for both of them (e.g. *koira näyttää söpöltä* ("the dog looks cute"); *se vaikuttaa tylsältä* ("it seems boring"))[82].

Yet another common source of error was the adverb *hyvin* which has been assigned the following semantic tags in the single word lexicon entry:

hyvin	Adverb	A5.1+	A13.3

The first semantic tag means "well" (e.g. *kaikki meni hyvin* ("all went well")). However, when *hyvin* is followed by an adjective or an adverb, it is a booster meaning "very" (e.g. *hyvin rauhallinen* ("very peaceful"), *hyvin hiljaa* ("very quietly"). A context rule could enable the FST software to identify the semantic tag A13.3 for the second, booster sense as the correct tag in similar contexts.

There are many opportunities to create context rules which can facilitate the identification of correct semantic tags in cases of ambiguous words. The findings of this evaluation provide many good candidates. Additionally, very useful lists of verb phrases which could also be

[82] There might be some exceptions, but here I concentrate on frequently appearing cases and ignore cases which appear very marginally.

written into context rules can be found, for example, in the book *Tarkista tästä*. *Suomen kielen rektioita suomea vieraana kielenä opiskeleville* ("Check Here. Finnish Verb Rections[83] for a Finnish Language Learner") (Jönsson-Korhola & White, 2002)[84].

4.4.3.4 Other error types

I identified four other error types in the application-based evaluation. These have been grouped into the major category "Other Error Types" and will be discussed in the following subsections.

4.4.3.4.1 Errors caused by TextMorfo

The performance of TextMorfo caused 8.17% of all errors encountered in the application-based evaluation, as Table 33 below shows:

[83] Verb rections refer to the phenomenon in which certain verbs or senses of the verbs have to be used with certain inflections. (As pointed out in section 2.5.1, Finnish predominantly uses inflections where English uses prepositions).

[84] However, it must always be ensured first that the sources used allow copyright-free use of their material.

Table 33

Application-Based Evaluation: Errors Caused by TextMorfo

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	14	29	17	44	46	150	8.17
Percentage of errors in this error type	9.33	19.33	11.33	29.33	30.67		

The largest number of errors occurred in the Hanna Subset and in the Separate Pool Times for Muslims Subset. This result could be expected, since TextMorfo has been developed for the analysis of general modern standard Finnish, while the Hanna Subset contains archaic use of language and the Separate Pool Times for Muslims Subset contains colloquial use of language. The most common error was caused by the fact that TextMorfo had not identified the part of speech correctly. This is exemplified in the sentence *Anna Hannan maata*, *puhui äiti ruokasalista*. ("Allow Hanna to lie down, spoke mother from the dining room."). The FST had processed this sentence in the following way:

pos="Proper" mwe="0" sem="Z1f" lem="Anna">Anna

pos="Proper" mwe="0" sem="Z1f" lem="Hanna">Hannan</

pos="Noun" mwe="0" sem="M7" lem="maa">maata

pos="_Delimiter" mwe="0" sem="PUNC" lem=",">,

pos="Verb" mwe="0" sem="Q2.2 Q2.1" lem="puhua">puhui

pos="Noun" mwe="0" sem="S4f" lem="äiti">äiti

pos="Noun/Noun" mwe="com" sem="H2/F1" lem="sali/ruoka">ruokasalista

pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.

This output contains two disambiguation errors. Firstly, TextMorfo had interpreted the first word in the sentence, *Anna*, as a female proper name (Anna). However, the word in question, *anna*, which was capitalized because of its sentence-initial position, should have been recognized as the imperative mood for the second person singular of the verb *antaa*. This in an ambiguous verb which in this context means "to allow". Secondly, TextMorfo had interpreted the verb *maata* ("to lie down", here in the basic form) as the partive singular of the noun *maa* ("country; land").

In addition, TextMorfo had faced difficulties in recognizing where the sentence ended if a sentence had been placed inside quotation marks. The following example is from the Kauniita Valheita Subset:

"*[…] Äiti on tässä*." *Ensiavussa lääkäri otti lapsen hänen sylistään*.

("'[…] Mum's here'. In the emergency room the doctor took the child from her arms.")

TextMorfo had not recognized the end of the sentence, and it had, therefore, interpreted the word *ensiapu* ("emergency room" in this context) starting the following sentence as a proper noun, since it was capitalized, and not as the common noun *ensiapu* which is included in the single word lexicon with the semantic tag B3. This, naturally, resulted in an error:

pos="Noun" mwe="0" sem="S4f" lem="äiti">Äiti

pos="Verb" mwe="0" sem="A3+ A1.1.1 M6 A8" lem="olla">on

pos="Adverb" mwe="0" sem="M6" lem="tässä">tässä

pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.

pos="Code" mwe="0" sem="Z99" lem=" ">

pos="Proper" mwe="0" sem="Z99" lem="Ensiapu">Ensiavussa

pos="Noun" mwe="0" sem="B3/S2" lem="lääkäri">lääkäri

pos="Verb" mwe="0" sem="M2 X7+ S7.4+ A1.1.1 I1.3 F2 X4.1" lem="ottaa">otti

pos="Noun" mwe="0" sem="T3-/S2" lem="lapsi">lapsen

pos="Pronoun" mwe="0" sem="Z8" lem="hän">hänen

pos="Noun" mwe="0" sem="B1 N3.7/N3.1 N3.4/N3.1" lem="syli">sylistään

pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.

pos="_EndOfSentence" mwe="0" sem="Z99" lem="NULL">NULL

Furthermore, TextMorfo does not recognize and thus cannot process words which are missing from its lexicon. As I noted in section 4.4.3.1.1, if TextMorfo does not recognize a word, this automatically results in the semantic tag Z99 for it, whether the word in question is included in the semantic lexical resources or not. Thus, further expanding of the semantic lexical resources would always necessitate expanding the TextMorfo lexicon simultaneously. In addition, it would be useful to carry out a file comparison between the single word lexicon and the TextMorfo lexicon to make the existing lexicons correspond. On the other hand, if TextMorfo is not developed further, an alternative solution could be to replace TextMorfo completely by another, more accurate and up-to-date morpho-syntactic analyser and parser.

4.4.3.4.2 Errors caused by archaic use of language

This error type includes errors which were caused by the type of language which I considered archaic. In the application-based evaluation, archaic use of language was represented by:

- words which do not belong in modern standard Finnish, such as *karotti*, *polsteri*, *talrikki*, *turrottaa*, and *töyttäys*,[85]
- MWEs which do not belong in modern standard Finnish, such as *herra jesta*, and *mieltä kääntää*[86], and
- archaic spelling variants, such as *ett'*, *jok'ainoa*, *kalliinta*, *peloittaa*, *tyyneenä*, and *väkisenkin*[87].

Not surprisingly, all 65 instances of this error type were found in the Hanna Subset which represented older Finnish text in this evaluation. The classic novel *Hanna* was written in 1886, while, as mentioned in section 4.3.2, the standard Finnish language as it exists today began to establish itself around the 1880s (Häkkinen, 1994, p. 15), and only around the 1920s, the modern spelling norms and grammar were standardized (Pulkkinen, 1972, pp. 57–65). These errors constituted 3.54% of all errors encountered in the application-based evaluation, as can be seen in Table 34 below:

[85] Modern translations: "bowl", "mattress", "plate", "to stare mopingly", "thump".

[86] Modern translations: "good heavens", "to feel disgusted".

[87] Modernized spelling forms with translations: *että* ("that"), *joka ainoa* ("every single"), *kalleinta* ("most expensive/important", here in the partitive singular), *pelottaa* ("to scare"; "to be afraid"), *tyynenä* ("calm", here in the essive singular), *väkisinkin* ("by force", followed by the enclitic particle *-kin* indicating "also").

Table 34

Application-Based Evaluation:

Errors Caused by Archaic Use of Language

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	0	0	0	65	0	65	3.54
Percentage of errors in this error type	0.00	0.00	0.00	100.00	0.00		

It would naturally not be sensible to add either archaic vocabulary or archaic spelling variants to semantic lexical resources such as the FST lexicons which are intended to cover general modern standard Finnish. Nevertheless, despite the archaic use of language, the accuracy obtained in the application-based evaluation with the Hanna Subset was 82.93% (see section 4.4.2). This suggests that the FST could already be of some help in tasks which require automatic semantic analysis of older Finnish text. The results can be further improved by redirecting the FST and the semantic lexical resources to successfully carry out semantic analysis of texts from earlier periods of time as well in the same manner as has been done with the EST, by using Variant Detector (VARD).

VARD has been developed at Lancaster University, and it functions as a pre-processor for text which contains spelling variation. More precisely, utilizing techniques which are

employed in modern spellchecking software, VARD processes spelling variants in texts into an output with modernized forms. This enables the study of historical texts with the same linguistic tools and methods which are used for modern language. VARD can be used both interactively and automatically, and the original variant is retained alongside the modernized variant. The current version of the tool is named VARD 2. VARD was originally developed for the analysis of Early Modern English texts, but lately it has been enabled to process any form of possible spelling variation. It can also be applied to languages other than English by incorporating a new language dictionary and spelling rules into the VARD software. (Lancaster University, 2016). VARD has been adapted and trained for processing, for example, children's language data (Baron & Rayson, 2009), second language learner data (Rayson & Baron, 2011), and SMS data (Tagg et al., 2012). It has also proven to be a very useful tool in the analysis of a corpus of personal Portuguese letters ranging from the 16th to the 20th century (Hendrickx & Marquilhas, 2011).

Unfortunately, VARD is not yet applicable to highly inflectional languages, such as Finnish, which would require lemmatisation as preprocessing. Nevertheless, this might change in the future, since there have been plans to extend the system to allow this (Alistair Baron, personal communication, April 12, 2016). In the meantime, should there be a need to use the FST for the analysis of older Finnish text, a supplementary lexicon containing older Finnish vocabulary could be added in the FST together with mechanisms similar to the mechanisms used in VARD for processing archaic spelling variants. The electronic corpora in the freely accessible online data service Kaino (Kotimaisten kielten keskus, n.d.) could provide useful test material for the development work. Kaino is maintained by the Institute of the Languages in Finland, and its collection contains works starting from the 16th century. The test corpora and the Hanna Subset representing older Finnish text in the evaluations of this thesis were collected from this very source.

4.4.3.4.3 Errors caused by colloquial use of language

This error type includes errors which were caused by the type of language which I considered colloquial. As could be expected, all the 48 instances were found in the Separate Pool Times for Muslims Subset which I collected from the online discussion forum of the yellow press newspaper *Iltalehti* to represent Finnish "Internet language" in the application-based evaluation. These errors constituted 2.62% of all errors encountered in the application-based evaluation, as Table 35 below shows:

Table 35

Application-Based Evaluation:

Errors Caused by Colloquial Use of Language

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	0	0	0	0	48	48	2.62
Percentage of errors in this error type	0.00	0.00	0.00	0.00	100.00		

Colloquial use of language was represented by:

- words which do not belong in standard Finnish, such as *järkätä*, *kantis*, *psori*, *seisokki*, *ulkkis*, and *älämölö*[88],
- MWEs which do not belong in standard Finnish, such as *100 varmaan* and *vetää tumppuun*[89], and
- colloquial spelling variants, such as *voitais* and *wanha*[90].

Despite the differences between the colloquial language use and the standard Finnish language, which has been the focus in the single word lexicon development, the accuracy obtained in the application-based evaluation with the Separate Pool Times for Muslims Subset was 79.46% (see section 4.4.2), which is quite promising and suggests that the FST could already be of help in the analysis of "Internet language" as well. Furthermore, it is possible to improve the results further. As mentioned in the previous subsection, the VARD tool is not only applicable to the analysis of historical text, but it can also be applied to any other form of spelling variation. If a version of VARD which is capable of dealing with Finnish becomes available in the future, it could be trained to process successfully forms of communication which are found in online social media websites, such as Facebook, Twitter, discussion forums, and blogs, as well as in chat and SMS messages. It would also be useful to include in such a lexicon the types of emoticons which express feelings and assign semantic tags for them as well. In the meantime, should there be a need to use the FST for the analysis of Finnish "Internet language", a supplementary lexicon containing colloquial Finnish vocabulary could be added in the FST together with mechanisms similar to the mechanisms used in VARD for processing the colloquial spelling variants. I will return to these issues in

[88] Standard Finnish translations: "organize", "regular customer", "psoriasis", "erection", "foreigner", "fuss".

[89] Standard Finnish translations: "100% sure", "to masturbate".

[90] Standardized spelling forms with translations: *voitaisiin* ("could" in the conditional passive), *vanha* ("old"; w is used for emphasis, since it represents an older spelling of the word).

section 5.3.2 where I suggest ideas for tailoring the Finnish semantic lexical resources for a domain-specific application for Internet monitoring purposes.

4.4.3.4.4 Errors caused by spelling errors in the test subsets

The error type "Spelling Errors in the Test Subset" contains both misspellings and typographical errors. These errors constituted 1.74% of all errors in the application-based evaluation, as Table 36 below displays:

Table 36

Application-Based Evaluation: Spelling Errors in the Test Subset

	Helsingin Sanomat (non-fiction)	**Kauniita Valheita (fiction)**	**Finnish Culinary Culture (domain-specific)**	**Hanna (older Finnish)**	**Separate Pool Times for Muslims (Internet)**	**Total number**	**Percentage of all errors**
Number of errors	2	1	0	1	28	32	1.74
Percentage of errors in this error type	6.25	3.13	0.00	3.13	87.50		

Overall, 28 of the total of 32 errors could be found in the Separate Pool Times for Muslims Subset, which was not surprising, given that it does not contain perfected professional writing but user-created content which is often written in a hurry and not

proofread and spellchecked before sending. A spelling error may also be due to the fact that the writer is not a native speaker of Finnish and does not know the correct spelling.

Spelling errors did not pose a major problem in the application-based evaluation. Nevertheless, having many misspelt and therefore unidentified words does impair the performance of TextMorfo and, consequently, also the performance of the FST. In case the FST was used for text containing a substantial number of spelling errors, such as online discussions in the Internet, it would be beneficial to include a tool in the FST which would preprocess the text and match the erroneous forms with correct forms. VARD, which has also been used for detecting spelling errors in written learner corpora (Rayson & Baron, 2011), or similar mechanisms could be of valuable assistance in this task as well.

4.4.3.5 Error analysis summary

Fourteen different error types were identified among the errors which were encountered in the application-based evaluation. These were further grouped into four major categories.

The major category "Errors Related to the Semantic Lexical Resources" constituted 31.07% of all errors (see Table 18 in section 4.4.2). Even though the Finnish single word lexicon is significantly larger and better at present than it was after the Benedict project when the formative evaluation was carried out, is very useful in terms of lexical coverage, and performs well in the FST software, it would benefit from further improvement. New entries need to be added as well as semantic tags for the missing senses in the existing lexicon entries. Furthermore, the MWE lexicon needs expanding, and its entries need to be written into accurate templates, as is the case with the English counterpart. In section 5.2, I will draft guidelines for the continued development of the Finnish semantic lexical resources as a general language resource.

The major category "Errors Related to the FST Software" constituted 38.20% of all errors (see Table 19 in section 4.4.2), with the largest single error type in it and in the entire evaluation being "Wrong Order of Senses" (33.02%). This error type was caused by the lack of disambiguation mechanisms. To resolve this problem, in section 4.4.3.2.1, I have suggested various approaches based on the solutions developed for the EST. Moreover, the compound engine caused a fair amount of unnecessary errors (4.14%). Addressing the error types of this major category would improve the accuracy of the FST substantially. It would be of uttermost importance to correct the above mentioned two deficiencies in the software, but it would also be beneficial to resolve the errors caused by ellipsis in compound constructions and the errors caused by wrong tags for ordinal numbers.

The major category "Errors Related Both to the Semantic Lexical Resources and to the FST Software" constituted 14.66% of all errors (see Table 20 in section 4.4.2). These errors included the types "Errors Caused by the Auxiliary Verb *olla* in Perfect and Pluperfect Constructions" and "Errors Caused by the Lack of Context Rules". These error types caused a large number of problems which need to be resolved as well. Addressing them would involve work on both the semantic lexical resources and on the software component. Some suggestions for creating context rules have been presented in section 4.4.3.3.2.

The major category "Other Error Types" constituted 16.07% of all errors (see Table 21 in section 4.4.2). TextMorfo needs to be improved and updated, since it caused a large number of errors (8.17%). Alternatively, it could be replaced completely with a more accurate and up-to-date morpho-syntactic analyser and parser. Furthermore, should the FST and the semantic lexicons, which have been developed for the analysis of general modern standard Finnish, be used for text containing archaic or colloquial use of language, the results could be improved significantly by incorporating VARD or similar mechanisms into the FST. VARD or similar

mechanisms could also be very helpful for processing the type of text which contains a considerable number of spelling errors.

4.5 Semantic Labeling Experiment

The third evaluation which I carried out for the new version of the Finnish single word lexicon is referred to as the "semantic labeling experiment". This experiment measures how general native users of Finnish are able to replicate the USAS categorisation which the Finnish semantic lexical resources are based on, and it was carried out in a similar manner as the experiments for Arabic, Chinese, English, Italian, Portuguese, and Urdu, described in El-Haj, Rayson, Piao, and Wattam (forthcoming). The test group for the Finnish language consisted of three native speakers of Finnish who were not familiar with the USAS categorisation in advance of the experiment. They were compensated for their work with movie tickets. The work was done through a user-friendly test interface, where the participants could do everything using mouse clicks except the final phase for each word, in which they copied the final output code to an Excel document.

The participants were each given the same set of 75 sample words which had been randomly selected from the single word lexicon and which all contain humanly assigned USAS semantic category tags. This set of sample words is referred to as the "gold standard". The participants were asked to label each word with a number of USAS semantic category tags which they considered to represent the given word's possible meanings. This task resembles the task which lexicographers have when they compile dictionary entries. The test interface offered a few semantic tag suggestions for each word. These included the semantic tags for the given word from the gold standard in randomized order as well as other random semantic tags from the USAS category system. The suggestions also included semantic tags

for categories which do not exist in the USAS category system at all. The participants were asked to remove the irrelevant semantic tags from this list of suggestions, and they could also add other semantic tags which they considered relevant from a list of all semantic tags used in the USAS category system. They could include as many[91] or as few semantic tags as they felt relevant, and they were also asked to place them in descending order of likelihood. They were provided with a link to the *Kielitoimiston sanakirja*[92] ("The New Dictionary of Modern Finnish") which they could consult, if necessary. They were also free to use any other reference sources.

Three different measures were used as indicators of the quality of the tag assignment. These measures are:

- First tag correct: Whether the first semantic tag selected by the participant matches the first semantic tag in the gold standard.
- Fuzzy: Whether the semantic tags selected by the participant are contained within the gold standard in any order.
- Strict: Whether the semantic tags selected by the participant are the same and appear in the same order as in the gold standard.

[91] The maximum was 10 semantic tags.

[92] This is the only noteworthy monolingual dictionary of Finnish (see section 2.5.3), and, to the best of my knowledge, there is no such dictionary of Finnish in which information about the frequency of the different senses of ambiguous words would be systematically available. Furthermore, there are no large representative corpora for Finnish which could have been utilized for this purpose.

For these measures, Fleiss' Kappa (Fleiss), Krippendorff's alpha (K-alpha), and Observed Agreement (OA) scores were calculated[93]. The results for the three participants are presented in Table 37 below:

Table 37

Semantic Labeling Experiment: Finnish

Measure	**K-alpha**	**Fleiss**	**OA**
First Tag Correct	0.31	0.31	0.81
Fuzzy	0.02	0.01	0.56
Strict	0.12	0.12	0.74

The Fleiss' Kappa and Krippendorff's alpha scores both indicate the reliability of agreement between the three participants in their tasks of assigning semantic category labels to the 75 sample words from the single word lexicon. In general, the participants assigned to the sample words many more semantic tags than what is included in the gold standard, in other words, the corresponding single word lexicon entries. This result was not surprising. Presumably, all the participants used mainly the *Kielitoimiston sanakirja* as their reference source, and the sense distinction in this dictionary is much finer-grained than the sense distinction in the single word lexicon, including also archaic, dialectal, domain-specific, and infrequent senses which the single word lexicon is not intended to cover (see the comparisons of sense distinctions between the *Kielitoimiston sanakirja* and the Finnish single word lexicon in section 3.4 and the discussion about the order of senses in the *Kielitoimiston sanakirja* in section 3.4.1). In addition, it was evident that the fact that the test interface turned all words into lower case had caused some confusion among the participants. The semantic lexicons in

[93] For both Krippendorff's alpha and Fleiss' Kappa, a higher score means better agreement. For example, 1 means perfect agreement, whereas 0 means no more than chance agreement.

the USAS framework are case-sensitive, but because all the test words were in lower case, this led the participants to believe that the case is the opposite, and, therefore, at times they mistook general nouns to be also proper nouns. An example of this is the noun *palmu* ("palm tree"). In addition to the relevant semantic tag L3 ("Plants"), the semantic tag Z1 ("Personal Names") had been suggested as well, since *Palmu* is a Finnish family name.

The observed agreement score indicates the percentage of the judgments on which the three participants agreed in their tasks of assigning semantic tags to the sample words. Even though the participants presumably mainly used the same reference source, their decisions differed relatively much from each other. This shows that it is a complex task to select senses for words and to define their order of likelihood. Indeed, as Kilgarriff (1997, pp. 103) points out in his seminal paper, lexicographers' decisions on whether to "lump" or "split" senses are inevitably subjective, and often the alternative decision would have been equally valid. Furthermore, Véronis (2000) noticed in his experiment that inter-annotator agreement was very low in a straight-forward sense tagging task using a traditional dictionary; for some words the agreement was no better than chance. With the very fine-grained sense distinctions in the *Kielitoimiston sanakirja*, it could be expected that the participants' choices in regard to selecting, lumping, and splitting were quite different from each other. That said, as I noted in section 3.4, the Finnish semantic lexical resources have been compiled by one person only. The reason for this was to ensure that the categorization of the single words and MWEs would be as coherent as possible.

As mentioned at the beginning of this subsection, similar semantic labeling experiments were carried out for Arabic, Chinese, English, Italian, Portuguese, and Urdu (El-Haj et al., forthcoming). In each of these experiments, there were four participants. They were given the same set of 250 sample words in their own language, and they were provided with a number of links to dictionaries, thesauri, and corpora which they could use as reference sources. Even

though the results are not directly comparable to the results for the experiment for Finnish[94], some observations can still be made. The observed agreement score was the highest for Finnish in regard to the First Tag Correct measure, but it was the lowest in regard to the Fuzzy and Strict measures. The Fleiss' Kappa and Krippendorff's alpha scores for the six languages varied relatively much. The results for Finnish followed approximately the same lines, but, in general, they were below the average. Table 38 below displays the results for English:

Table 38

Semantic Labeling Experiment: English

Measure	**K-alpha**	**Fleiss**	**OA**
First Tag Correct	0.36	0.36	0.71
Fuzzy	0.11	0.11	0.58
Strict	0.20	0.20	0.79

These results represent by and large the average among the results for these six languages.

4.6 Chapter Summary

At the beginning of this chapter, I have summarized the formative evaluation carried out at the end of the Benedict project in 2005. Following that, I have reported on two new evaluations subsequent to extending and improving the Finnish single word lexicon during the past years. These are:

[94] The experiment for Finnish was smaller-scale than the experiments for the other languages. In the Finnish experiment, there were only three participants and 75 sample words.

1) The final evaluation which measures the lexical coverage, in other words, the extent of the Finnish single word lexicon. This evaluation has answered **RQ2 (How extensive is the Finnish single word lexicon in terms of lexical coverage?)**.
2) The application-based evaluation which measures the accuracy, in other words, how well the Finnish single word lexicon performs when it is applied in the FST software. This evaluation has answered **RQ3 (How suitable is the Finnish single word lexicon for use in the semantic analysis of Finnish in the FST software?)**.

I have selected the test material for the experiments in the final evaluation and in the application-based evaluation from various sources which reflect such different aspects as genre, domain, and historical period in order to ensure that the results of the evaluations would reflect the overall performance of the Finnish single word lexicon and the FST in practical annotation tasks. The results revealed that the lexical coverage had clearly improved over the Benedict project thanks to the new version of the single word lexicon. The lexical coverage on the test corpora containing general modern standard Finnish ranged from 94.58% to 97.91% which is comparable to the results obtained with the English equivalents and thus indicates that the Finnish single word lexicon indeed covers the majority of core Finnish vocabulary. Furthermore, even though the single word lexicon was developed for the analysis of general modern standard Finnish text, it also performed surprisingly well in the analysis of domain-specific text (95.36%) and older Finnish text (92.11–93.05%) as well as when applied to the analysis of Internet discussions (91.97–94.14%), in which the language often contains different types of colloquialisms as well as spelling errors.

Although the results are not altogether comparable, it was evident that the accuracy had not improved over the accuracy obtained in the Benedict project. The accuracy in the formative evaluation on the subset of Finnish cooking texts was 83.08%, while in the

application-based evaluation the accuracy ranged between 79.51% and 83.48% on the five different test subsets. The results proved that the new, expanded, and improved single word lexicon indeed covers a wider vocabulary than the old version and produces much better results when it is applied in the FST. However, if one wishes to improve the accuracy of the FST, extending the single word lexicon is not alone sufficient for that purpose, but that it would also require extending the MWE lexicon and writing accurate templates for all its entries and, most of all, it would require investing in the development of the other components of the program.

In section 4.4.3, I have provided an analysis of the errors which occurred in the application-based evaluation. I have also suggested ideas for addressing these errors, concentrating primarily on the further improvement of the semantic lexical resources which are the main focus of this thesis. Firstly, the semantic lexical resources need to be expanded by including new entries and missing senses into them. Secondly, all the MWE lexicon entries need to written into accurate templates which would enable the FST to reliably recognize and tag different types of Finnish MWEs. These issues will be discussed in more detail in the following chapter.

I have concluded this chapter by presenting the third evaluation of the new single word lexicon. This semantic labeling experiment measured how general native speakers of Finnish are able to replicate the categorisation of the Finnish single word lexicon. The results showed that the three participants had assigned to the sample words many more semantic tags than what is included in the gold standard, in other words, in the corresponding single word lexicon entries. Furthermore, the results showed that there also was disagreement between the participants in their choice of the semantic tags for the sample words as well as in their choice for their frequency order. This indicates that it is a complex task to select senses for words and to define their order of likelihood.

5 Discussion and Further Development of the Finnish Semantic Lexical Resources

5.1 Introduction

This chapter contains the discussion, reflecting on the main contributions of the thesis presented in chapter three and their evaluation in chapter four. I will begin by drafting guidelines for the continued development of the Finnish semantic lexicons as a general language resource. Consequently, I will discuss the requirements for tailoring the Finnish semantic lexical resources for domain-specific applications. This chapter answers **RQ4 (What resources and methods can be useful for the further development of the Finnish semantic lexical resources, firstly, as a general language resource, and, secondly, when they are applied to new domains?)**.

5.2 Developing the Finnish Semantic Lexicons Further as a General Language Resource

In the following subsections, I will draft guidelines for the continued development of the Finnish semantic lexicons as a general language resource, similar to the resource included in the EST. These guidelines are applicable beyond the work described here, for example, in the development of the semantic lexical resources for semantic taggers in other languages.

The guidelines are for the most part based on the lessons learned from the evaluations presented in the previous chapter. They include ideas for expanding both the single word

lexicon and the MWE lexicon, and I will also suggest a novel solution for writing accurate templates to be able to identify and tag different types of Finnish MWEs. Finally, I will propose the creation of an autotagging lexicon for Finnish.

5.2.1 Improving the single word lexicon

The syntactically and semantically tagged single word lexicon is already very extensive in terms of lexical coverage, as became evident in the final evaluation of lexical coverage (see section 4.3.2). However, since language changes and evolves constantly, it would naturally be necessary to update the lexicon on a regular basis to ensure that all the relevant and current vocabulary is included. In addition to including missing words, it would also be important to add missing senses for the existing lexicon entries.

Expanding the single word lexicon by adding new entries is a well-defined task. It would be practical to start by examining the words which were found to be missing in the final evaluation of lexical coverage (see section 4.3), select those words which would be worthwhile additions to the single word lexicon, and then provide them with relevant syntactic and semantic tags. It would also be useful to make a file comparison between the single word lexicon and the list of 9,996 words found to be the most common in Finnish newspaper texts (Kielipankki, n.d.) to ensure that all the words in the frequency list are already included in the single word lexicon[95]. This frequency list was created in 2004 by CSC (IT Center for Science), and they had used 43,999,826 words of newspaper text from the 1960s as source material[96]. In addition, since the frequency list is based on material which is 50 years old, it would also be sensible to run, for example, recent newspaper text through the

[95] Such a comparison has not been carried out yet, since I encountered the frequency list only after carrying out the evaluations.

[96] A frequency dictionary of Finnish does exist (Saukkonen, Haipus, Niemikorpi, & Sulkala 1979), but it is in the printed form and outdated. Therefore, it would not be useful for these purposes.

FST and search for more new entry candidates to keep the single word lexicon up to date. Furthermore, it would be beneficial to map the USAS semantic categories with the categories used in the Finnish General Upper Ontology (see section 2.5.2.1) and in the Finnish WordNet (see section 2.5.2.2) and collect new entry candidates from these sources as well. This would be a very practical and reliable way to expand the single word lexicon.

The USAS framework is based on the idea of semantic fields, but there are also numerous other techniques which are called semantic annotation or semantic tagging although they are based on different approaches. One example of these different approaches is named entity recognition (e.g. Tjong & De Meulder, 2003; Nadeau & Sekine, 2007) in which names of various entities are labeled. Resources which are used for named entity recognition could be exploited for expanding the coverage of the Z category in the single word lexicon. This top level category includes personal names (Z1), geographical names (Z2), and other proper names (Z3), such as trademarks and names of companies and institutions. By way of illustration, the geographical database GeoNames[97], which is available for download free of charge, could be a beneficial source. In addition, Wikipedia[98] offers various lists which could be utilized for this purpose as well.

It is important to bear in mind that whenever expanding the single word lexicon, the new entries must always be included in the TextMorfo lexicon as well, if these are not included already, to make the two lexicons correspond. If a word is included in the single word lexicon but is missing from the TextMorfo lexicon, this word remains unrecognized (see section 4.4.3.1.1). An alternative option would be to replace TextMorfo completely by another, more accurate and up-to-date morpho-syntactic analyser and parser of Finnish.

Senses which are discovered to be missing must also be included into the existing single word lexicon entries. However, detecting missing senses is more challenging than detecting

[97] For more information, see http://www.geonames.org/.

[98] For more information, see https://fi.wikipedia.org/wiki/Wikipedia:Etusivu.

missing words, since missing senses cannot be searched out automatically, but this task requires manual checking.

The content of the semantic lexical resources is the key issue, not the size. As I pointed out in section 4.3.3.1.1, the main aim in the development of such a general language resource has been to try to incorporate the core vocabulary of Finnish into it. In other words, the purpose is not to include all the vocabulary in the language exhaustively, such as jargon, technical terms, colloquialisms, or very rarely occurring proper nouns or other words; this would result only in an unmanageable lexicon size. This principle applies especially to compounds. Compounding is a very productive means of word formation, and the number of possible compounds is infinite. For this reason, it is practical to include only the relatively frequently appearing compounds as well as lexicalized compounds[99], and the compound engine component described in section 3.3.2 has been developed to process all other possible, less frequently used compounds of a more temporary nature. In regard to adding missing senses to existing lexicon entries, it would be sensible to include all the commonly used senses and, additionally, develop better disambiguation mechanisms in order to be able to choose the correct sense in a given context.

5.2.2 Improving the multiword expression lexicon

Whereas the single word lexicon already has a wide coverage, the present MWE lexicon in turn is still incomplete. For this reason, it was omitted from the evaluations carried out for this thesis. Thus, the logical next step, therefore, would be to start improving the MWE lexicon.

[99] The meanings of lexicalized compounds cannot easily be deduced from the sum of the meanings of the element words (see section 2.5.1.2).

The further development of the MWE lexicon is a much more complex and time-consuming task than the further development of the single word lexicon. Firstly, the MWE lexicon needs to be expanded by adding new entries into it. Secondly, the "quick MWE template solution" (see section 4.2), which was adopted in the Benedict project due to a limited amount of time and which proved to be neither useful nor intelligent, needs to be replaced with a novel system which utilizes more accurate templates. Guidelines for improving the MWE lexicon will be presented in the following subsections.

5.2.2.1 Collecting new entries

The first step in the development of the MWE lexicon involves expanding the size of the MWE lexicon in terms of entries. At the moment, the MWE lexicon contains 6,312 entries which include, for instance, noun and verb phrases, idioms, proverbs, and multiword proper names (see section 3.4.2). By comparison, the English MWE lexicon consists of 18,921 entries (Mudraya et al., 2006). Since Finnish does not use phrasal verbs as much as English and compounds are almost always written as one orthographic word, unlike the case often is for English, it would most likely not be necessary to include equally many entries into the Finnish MWE lexicon[100]. However, it definitely needs substantial expansion, since it still lacks many frequently used MWEs. To begin with, it would be useful to examine the MWEs found to be missing in the application-based evaluation of precision (see section 4.4) and incorporate those MWEs regarded as worthwhile additions into the MWE lexicon. Moreover, new entry candidates could be collected, for example, from the idiom dictionary of Finnish named *Naulan kantaan: Nykysuomen idiomisanakirja* ("To Hit the Nail on the Head. Idiom

[100] It was estimated that 16% of words in English running text are semantic MWEs (Rayson, 2005, p. 4). Unfortunately, corresponding information is not available for Finnish.

Dictionary of Modern Finnish"; Kari, 1993)[101]. Glossaries and lists of Finnish idioms and proverbs could also provide valuable findings. One very beneficial source could be the PhD thesis *Idiomit ja leksikko* ("Idioms and lexicon") written by Marja Nenonen (2002). The appendices of Nenonen's thesis comprise lists of thousands of idioms. Furthermore, the General Upper Ontology (see section 2.5.2.1), the Finnish WordNet (see section 2.5.2.2), and various resources used for named entity recognition, such as GeoNames and Wikipedia, could be exploited for developing not only the single word lexicon but also the MWE lexicon.

Moreover, automatic approaches could be utilized for expanding the coverage of the MWE lexicon. Piao et al. (2005b) have used the EST for identifying such MWEs from corpora which depict single semantic concepts. This approach could be tested for identifying Finnish MWEs as well. In addition, some type of automatic statistical tools could be used for extracting Finnish MWEs from large corpora to find more entry candidates.

5.2.2.2 Creating templates for multiword expressions

The second step in the development of the MWE lexicon involves the creation of a novel system for writing templates to replace the old solution. This will enable one to reliably recognize and tag Finnish MWEs. In this subsection, I will first present the set of wild cards which are used in the MWE templates of the EST, and, thereafter, utilizing them I will draft some example templates for different types of Finnish MWEs.

The information in a Finnish MWE lexicon entry needs to be presented in two parts, in the same way as in the English counterpart. The first part contains a sequence of words and their respective POS tags joined together with various wild cards representing simplified forms of

[101] However, it must always be ensured first that the sources used allow copyright-free use of their material.

regular expressions[102], whereas the second part contains the relevant semantic tags. The POS tags employed come from TextMorfo and are the following: Abbreviation, Adjective, Adverb, Code, Conjunction, Interjection, Noun, Numeral, Preposition, Pronoun, Proper, and Verb. The wild cards used here are the same as the wild cards used in the EST. They are underscore (_), asterisk (*), curly brackets ({}), and slash (/), and they will be clarified in the subsequent paragraphs.

I have distinguished between three different types of Finnish MWEs in regard to writing templates. In the first, most simple MWE type, none of the constituent words can be inflected, no enclitic particles are used[103], and no embedded elements are allowed between the constituent words. Such MWEs are typically fixed expressions, proverbs, and abbreviations, and their templates consist of the constituent words which are joined together with their POS tags by underscores. Table 39 below displays some example templates for such MWEs:

[102] A regular expression is defined by Baker, Hardie, & McEnery (2006, p. 138) as a "type of string that may include special characters (sometimes referred to as "wild cards") that mean the regular expression as a whole will match with more than one string".

[103] In principle, a creative mind would find it possible to add at least some enclitic particles to the end of nearly every word. However, here I concentrate on at least relatively frequently appearing formations and ignore cases which are in principle possible but appear very marginally.

Table 39

Examples of Finnish MWE Templates: MWE Type 1

Finnish MWE template	**Semantic tag/s**	**English translation**
aina_Adverb vain_Adverb	T2++ A13.3	all the more
fil._Abbrev maist._Abbrev	P1/S2	M.Phil
hätä_Noun ei_Verb lue_Verb lakia_Noun	Z4	any port in a storm
kaiken_ Pronoun A_Code ja_Conjunction O_Code	A11.1+++	of utmost importance
kaikkien_Pronoun aikojen_Noun	A13.2	of all time
kyllä_Adverb kai_Adverb	A7-	supposedly
taivas_Noun varjele_Verb	Z4	good heavens

However, there is an alternative, somewhat simpler solution for treating such MWEs. In section 3.4.1, I introduced some frequently co-occurring MWEs which I have included in the single word lexicon instead of the MWE lexicon. These are expressions in which the constituent words cannot be inflected, no enclitic particles are used, and where no embedded elements are allowed between the constituents. In principle, all entries that consist of two or more words with an intervening space between them do belong in the MWE lexicon, but since TextMorfo in the POS tagging phase processes some fixed expressions as single units and then assigns a POS tag to the entire expression[104], it was found practical to treat these as single units in the semantic tagging component as well. For this reason, these have been included in the single word lexicon, and the spaces between the constituents have been replaced by underscores. For instance, the expression *suoraan sanoen* ("to tell you the truth")

[104] These resemble the ditto tags which are used in CLAWS (University Centre for Computer Corpus Research on Language (n.d-c)).

functions as an adverb. Thus it has been assigned the Adverb tag in the TextMorfo lexicon and the semantic tag A5.2+ in the single word lexicon. The resulting entry is thus the following:

suoraan_sanoen	Adverb	A5.2+

The FST recognizes such MWEs as single word lexicon entries without any problems and tags them correctly. For this reason, it would be practical to treat all MWEs of this type in a similar manner when the semantic lexical resources are further developed. Such MWEs which were found to be missing in the application-based evaluation and which could be written as such single word lexicon entries were, for example, *kuin yö ja päivä* ("like chalk and cheese") and *vuosien saatossa* ("over the years"). For them I would suggest the following single word lexicon entries:

kuin_yö_ja_päivä	Adjective	A6.1-
vuosien_saatossa	Adverb	T1.3+

Successful recognition of such MWEs treated as single word lexicon entries would, naturally, also require adding them to the TextMorfo lexicon.

The second MWE type includes MWEs in which one or more constituent words can be inflected and/or it is possible to add enclitic particles to them. This is marked with the asterisk wild card character in the template. Such MWEs are most often noun phrases and multiword proper names. Table 40 below displays some examples of templates for such MWEs:

Table 40

Examples of Finnish MWE Templates: MWE Type 2

Finnish MWE template	**Semantic tag**	**English translation**
absoluuttinen*_Adjective* nollapiste*_Noun*	O4.6	absolute zero
Helsingin_Proper yliopisto*_Noun*	Z3/P1	University of Helsinki
karkeakarvainen*_Adjective* mäyräkoira*_Noun*	L2	wirehaired dachshund
yleisen_Adjective kielitieteen_Noun laitos*_Noun*	P1/Q3	department of general linguistics

Because of the rich morphology and the enclitic particles of the Finnish language, one single MWE template of this type can cover a large number of surface forms. By way of illustration, the MWE *karkeakarvainen mäyräkoira* ("wirehaired dachshund") can be inflected both in singular and plural in all the 15 Finnish cases (see section 2.5.1.1). Both of the MWE constituent words, the adjective *karkeakarvainen* ("wirehaired") and the noun *mäyräkoira* ("dachshund"), take the same endings, since in Finnish attributes always agree with the headword in case and number. Moreover, it is possible to add possessive suffixes and enclitic particles **-***kin* (indicating "too"), **-***ko* (indicating question), and **-***han* (indicating emphasis) to them (see section 2.5.1.1). In all, this results in hundreds of possible surface forms which this one single template should be able to capture. Nevertheless, not all MWEs of this type are equally productive. For example, in the MWEs *Helsingin yliopisto* ("University of Helsinki"), only the second constituent *yliopisto* ("university") can be inflected and take enclitic particles. This is due to the fact that the first constituent is already inflected; in this case, the geographical name *Helsinki* appears in the genitive case (*Helsingin* ("of Helsinki")).

The third MWE type consists of MWEs in which 1) one or more constituent words can be inflected and/or it is possible to add enclitic particles to them and 2) one or more embedded elements can occur between the constituent words. Such MWEs are referred to as discontinuous, and this phenomenon occurs mostly in verb phrases. The POS tags for the possible embedded elements are marked within curly brackets in the MWE template. If there are more possible embedded elements in terms of part of speech than one, their respective POS tags are listed and separated by a slash. This means that any of the items inside the curly brackets can occur in any possible order, and they can also be repeated. In the English MWE lexicon, the number of possible POS tags for the embedded elements has been limited to a maximum of three tags for the sake of simplicity. I would consider this approach practical for the Finnish MWE lexicon as well. Table 41 below displays some examples of templates for such MWEs:

Table 41

Examples of Finnish MWE Templates: MWE Type 3

Finnish MWE template	**Semantic tag**	**English translation**
ajaa*_Verb* { Adjective*/Adverb*/Noun*} parta*_Noun*	B4	to shave
ottaa*_Verb* {Adverb*/ Noun*/Proper*} aurinkoa_Noun	K1/W4	to sunbathe
ottaa*_Verb* {Adverb*/Proper*/Pronoun*} ero*_Noun*	S4/X7-	to divorce
pitää*_Verb* {Adverb*} jalat_Noun {Adverb*} maan_Noun pinnalla_Noun	S1.2.6+	to keep one's feet on the ground

By way of illustration, the MWE template *ottaa*_Verb* {Adverb*/Noun*/Proper*} aurinkoa*_Noun** ("to sunbathe"; literally "to take sun") would capture, among others, the following expressions (embeddings underlined):

- *[minä][105] otan aurinkoa* (literally: "I take sun")
- *[minä] otan tänään aurinkoa* (literally: "I take today sun")
- *[minä] otan rannalla aurinkoa* (literally: "I take on the beach sun")
- *[minä] otan Rivieralla aurinkoa* (literally: "I take at Riviera sun")
- *[minä] otan tänään rannalla aurinkoa* (literally: "I take today on the beach sun")
- *[minä] otan rannalla tänään aurinkoa* (literally: "I take on the beach today sun")

As I noted in section 2.5.1.3 when describing the specific features of the Finnish language, the word order in Finnish is relatively free. Changing the word order, however, often affects the thematic structure resulting in new emphases and nuances even though the core meaning of the sentence remains the same. Furthermore, due to the flexible word order, the adverb *tänään*, for example, would not exclusively appear as an embedded element in the MWE *ottaa aurinkoa*, but it could also appear before or after the MWE, again resulting in a slight change of emphasis as the following examples illustrate:

- *tänään [minä] otan aurinkoa* (literally: "today I take sun")
- *[minä] otan aurinkoa tänään* (literally: "I take sun today")

[105] Personal pronouns are often omitted in the first and second person singular and plural when they appear before the verbs.

Moreover, also due to the flexible word order, the embedded element can sometimes appear in more places than one in a MWE. By way of illustration, the template *pitää*_Verb {Adverb} jalat* {Adverb} maan_Noun pinnalla_Noun* ("to keep one's feet on the ground") listed among the examples in Table 41 allows an embedded adverb in two different places. Thus, this template would capture both of the following MWEs:

- *pidin aina jalat maan pinnalla* (literally: "I kept always the feet on the ground")
- *pidin jalat aina maan pinnalla* (literally: "I kept the feet always on the ground")

And, naturally, this template would also capture, for instance, the following expressions which include embeddings in two different places:

- *pidin aina tiukasti jalat maan pinnalla* (literally: "I kept always firmly the feet on the ground")
- *pidin jalat aina tiukasti maan pinnalla* (literally: "I kept the feet always firmly on the ground")
- *pidin aina jalat tiukasti maan pinnalla* (literally: "I kept always the feet firmly on the ground")

The templates for this third MWE type are the most productive. If all the possible inflections and their combinations were taken into account as well as all potential enclitic particles, the number of different MWEs captured with such templates would be in the hundreds, if not in the thousands. Furthermore, if all imaginable embedded elements in their basic and inflected forms were taken into account, the number would further increase enormously.

5.2.3 Creating an autotagging lexicon

In addition to developing the Finnish semantic lexical resources by improving the lexicons for single words and MWEs, I also suggest the creation of a small autotagging lexicon similar to the autotagging lexicon included in the EST. An autotagging lexicon consists of fixed patterns which can have many possible instantiations. Such expressions can be tagged effectively through the use of wild cards. For example, the following autotagging lexicon entry would tag all combinations of numbers and the abbreviation *dl*, such as *½ dl* and *5 dl*, as N3.4 ("Measurement: Volume").

\S+dl[106]	Abbrev	N3.4

The autotagging lexicon could contain corresponding entries, for example, for the types of items presented in Table 42:

[106] "\S+" is a regular expression that matches any single string.

Table 42

Suggestions for Finnish Autotagging Lexicon Entries

Types	**Semantic tag**	**Examples**
Currencies	I1	euro, punta, dollari, rupla[107], €, £, $
Volume	N3.4	m^3, l, dl, ml, cl
Weight	N3.5	kg, g, mg
Area	N3.6	ha[108], m^2
Length and speed	N3.8	km, m, cm, mm, km/s, m/s
Amounts	N5	%, ‰, kpl[109]
Temperature	O4.6	°C
Time	T1	klo[110]
Period of time	T1.3	v, kk, h, min, s

A novel idea for the development of the autotagging lexicon arose in connection with the error type "Errors Caused by the Lack of Context Rules" in the application-based evaluation. In section 4.4.3.3.2, I noted that if a context rule was employed in the FST software, it would enable the FST to tag, for instance, the expressions *1770-luku* ("1770s") and *60-luku* ("60s") together with other similar expressions correctly as T1.3. Alternatively, this could be carried out with the aid of the following autotagging lexicon entry:

Numeral+-luku Noun T1.3

[107] Translations: euro, pound, dollar, rouble.

[108] Translation: hectare.

[109] Translation: pc. (abbreviation for "piece").

[110] This is the abbreviated form of the word *kello* ("clock"). It appears in constructions such as *klo 10* ("10 o'clock").

The absence of an autotagging lexicon does not cause any errors to the tagged output. However, the English autotagging lexicon has been found very practical for improving the results, and I believe that such a mechanism would be a useful addition to the FST as well to make the program more intelligent and would remove the need to spell out many patterns in full.

5.3 Tailoring the Finnish Semantic Lexical Resources for Domain-Specific Applications

In section 5.2, I suggested ideas for the continued development of the Finnish semantic lexicons as a general language resource. It was evident that the most acute task in the further development of the semantic lexical resources would be, firstly, to expand the MWE lexicon and, secondly, to write templates for all its entries. Needless to say, this would involve a considerable amount of time and effort, but it would be essential for applications in which all single words and MWEs in a text need to be recognized. Naturally, the single word lexicon would also benefit from expansion, although it is presently, as the final evaluation of lexical coverage proved, already in a much more mature state than the MWE lexicon.

However, the Finnish semantic lexical resources could already be put to practical use more speedily and easily. Much less work would be required if the semantic lexical resources were tailored for a specific purpose to deal with only one particular domain or task. In such a case, only the relevant single words and MWEs would need to be recognized and thus included in the semantic lexical resources rather than any single word or a MWE which could be considered as belonging to general standard modern Finnish; the latter would be the case in the development of a general language resource. The Finnish semantic lexical resources already have a good coverage, and if a particular application requires, they can be expanded

with relevant vocabulary from the given field. Moreover, if a particular application requires, it is possible to create new semantic categories and subdivide the existing categories further, thanks to the flexibility of the USAS category system.

In section 2.3, I discussed different types of semantic ontologies. The USAS category system, which is a complete conceptual system, represents the conceptual analysis method which was reviewed in section 2.3.1. However, the development of domain-specific applications comes closer to the content analysis method which was reviewed in section 2.3.2. The selection of certain single words, MWEs, and semantic tags forms a type of a "content analysis dictionary" which is tailored for a particular task (cf. the General Inquirer, section 2.3.2.1).

In the following subsections, I will briefly envisage how to tailor the Finnish semantic lexical resources for specific domains and tasks, using named entity recognition, Internet content monitoring, psychological profiling, and sentiment analysis as example cases. These tasks require differing levels of adaptation; some tasks require more changes in the semantic lexicons, whereas some tasks require the identification of useful existing semantic field information and less new entries in the semantic lexicons.

5.3.1 Named Entity Recognition

After submitting this thesis for examination, I began expanding the semantic categories Z1 ("Personal Names", including the semantic tag Z1 for family names, Z1f for female names, and Z1m for male names) and Z2 ("Geographical Names") with the intention to test the FST for named entity recognition tasks and to compare these results to the results obtained with FiNER, a standard rule-based named entity tagger for Finnish (Silfverberg, 2015). For this task, I utilized word lists provided to me by my colleague Kimmo Kettunen (National Library

of Finland) who had collected them from freely available sources, such as the resources provided by the Institute for the Languages of Finland, the National Land Survey of Finland, and Wikipedia. These word lists contained Finnish and Swedish first names and family names[111] as well as names of cities, towns, villages, and neighbourhoods in Finland.

I carried out the lexicon expansion in the following manner, utilizing Microsoft Excel. Firstly, I added the grammatical tag "Proper" and the relevant semantic tag, in other words, Z1, Z1f, Z1m, or Z2, to all the words in the above mentioned lists. Secondly, I combined the resulting lists and the existing Finnish single word lexicon[112] and sorted these entries in alphabetical order. Thirdly, I worked through the resulting document by searching for duplicates among the proper names. It is very common in Finnish that a geographical name is also a personal name (e.g. *Jurva* and *Sirkka*), and sometimes it can be some other proper name as well representing the semantic category Z3 (e.g. *Anttila*, the name a Finnish department store chain which recently went bankrupt). Furthermore, some female and male names are also family names (e.g. *Ahti* and *Helmi*). In cases of such ambiguous proper names, I combined the semantic tags and arranged them in perceived frequency order in the lexicon entry, for instance:

Ahti	Proper	Z1	Z1m	
Anttila	Proper	Z1	Z2	Z3
Helmi	Proper	Z1f	Z1	
Jurva	Proper	Z2	Z1	
Sirkka	Proper	Z1f	Z2	

[111] Swedish is the second official language in Finland, and, therefore, Swedish personal names are common in Finland.

[112] Finnish personal names and names of cities, towns, villages, and neighbourhoods are generally written as one orthographic word.

Consequently, the number of words which have been assigned the semantic tag Z1, Z1f, or Z1m as the first semantic tag in the single word lexicon entry has now increased from 3,850 entries to 17,677 entries, and the number of words which have been assigned the semantic tag Z2 as the first semantic tag in the single word lexicon entry has increased from 4,215 entries to 19,665 entries.

In the evaluation (Kettunen & Löfberg, forthcoming), two different datasets were used as test material:

1) Digitized Finnish historical newspaper collection Digi. This OCRed newspaper collection contains 1,960,921 pages of newspaper material published between the years 1771 and 1910, both in Finnish and Swedish. Only the Finnish material was used in our evaluation. The collection has a great number of OCR errors, and the estimated word level correctness is about 70–75%. The NER evaluation collection of Digi consists of 75,931 words. (Kettunen, Mäkelä, Kuokkala, Ruokolainen, & Niemi, 2016; Kettunen & Pääkkönen, 2016)
2) Modern Finnish technology news. This NER evaluation collection consists of 31,000 words[113] in 240 articles published in the technology and business-oriented online newspaper *Digitoday*[114]. In our evaluation, we used 64% of this data.

The results showed that the FST performed for the most part as well as FiNER with persons and locations in modern data. With historical data, the FST performed a little worse with persons than FiNER, whereas with locations both taggers performed equally. Corporations were not evaluated in the historical data, but in *Digitoday's* data FiNER

[113] Punctuation marks are included in the word count.

[114] *Digitoday* was merged to a Finnish tabloid newspaper named *Iltasanomat* in 2016, and it is now available at http://www.is.fi/digitoday/.

performed clearly better with corporations than the FST. The outcome of this evaluation was promising, firstly, considering that at this point the Finnish single word lexicon had only been expanded with Finnish and Swedish personal names and Finnish geographical names, and, secondly, considering that, unlike FiNER, the FST does not yet include any mechanisms to disambiguate between ambiguous names.

Thus, this domain-specific task for the FST and the Finnish semantic lexical resources requires most of all extending the lexical coverage by adding new entries to the lexicons and new senses to the existing lexicon entries, where the proper names are ambiguous. Now that the single word lexicon has a good coverage of Finnish and Swedish personal names and names of cities, towns, villages, and neighbourhoods in Finland, I will next expand it, as well as the MWE lexicon, with foreign personal and geographical names, and, consecutively, with names of both Finnish and foreign companies, organizations, institutions, and trademarks, following the same procedures as described in this subsection. I also intend to add Finnish geographical names which are still missing, such as names of rivers, lakes, and fells. The resulting lexicons will be made available open source for research under the Creative Commons license at the USAS website[115].

5.3.2 Semi-automatic Internet content monitoring

The FST and its semantic lexicons could also be reoriented for developing a semi-automatic Internet content monitoring program which could be used for browsing quickly and efficiently through Internet text in order to detect certain predefined characteristics of speech. To achieve this, it would first be necessary to study the type of speech which is to be identified. In the second phase, the relevant features would be recorded in the semantic lexical

[115] http://ucrel.lancs.ac.uk/usas/

resources, and, subsequently, the semantic tagger would be incorporated within a search engine component to locate them. Thus, this task is a combination of extending the coverage of the semantic lexical resources as well as of identifying and using existing relevant semantic category tags.

The monitoring program would not be fully automatic, but it would rather be a semi-automatic tool for pre-processing text. If the monitoring program discovered questionable content, it would alert a human supervisor, such as a moderator of a discussion forum, and point him to the relevant spot on the website. This person could check to see if there is reason for concern and then intervene, if considered necessary. This type of an arrangement could be seen as a sensible division of work between the human being and the computer. The human being would write the rules according to which the computer would do all the hard and monotonous work, while the human being would still be in charge of the end result by manually checking the possible findings of the computer. Beneficiaries of the system which is proposed here could include, for example, publishers, media companies, and Internet operators who maintain social media websites, the police and other law enforcement agencies, as well as healthcare authorities and organizations[116].

5.3.2.1 Internet content monitoring program for detecting hate speech targeted at immigrants

In the following subsections, I will draft guidelines for the development of an Internet content monitoring program utilizing the Finnish semantic lexical resources. The example

[116] Should such applications be developed, ethical and legal concerns need to be taken into account throughout the development process.

case is hate speech targeted at immigrants which is a common problem among user-created content in Finnish social media websites.

5.3.2.1.1 Studying the speech of the selected target group

The first step in the creation of the proposed Internet content monitoring program would be to study the relevant features which need to be recognized. Various sources of information could be useful for gathering the necessary background information. A particularly valuable source of information would be the logging of messages which have been deemed unacceptable earlier by the moderators and which, therefore, have been deleted from the website. Many websites also provide a possibility for their users to report inappropriate messages. These messages could contain very useful material as well. Furthermore, it would be helpful to consult specialists, literature, and studies in the given field.

The lexicon construction would not occur only at the initial stages of the development process, but a developer, such as a linguist, would examine incoming messages, comments, and blog postings at regular intervals and update and expand the lexicons to improve the system. I believe that the most practical method for the updating and expanding would be to examine the material which the monitoring program identifies as possibly alarming and sends over to the moderators to check manually. In case this material indeed contains unacceptable content, it is likely that in the vicinity of the hits identified by the monitoring program there are also other single words and MWEs which could provide useful clues for the program but which are not recorded in it yet. Constant lexicon development is also important because of the fact that, as Warner and Hirschberg (2012, p. 21) point out, to evade an automatic moderation system, writers may try to obscure the words in their questionable text with, for example, intentional misspellings and expanded spelling, in other words, they may separate

the characters by spaces or punctuation marks. Such endeavours must be recognized, and, subsequently, common intentional misspellings must be included in the semantic lexicons and the monitoring program must be trained to recognize words even if they include expanded spelling.

Hate speech targeted at immigrants could be identified in Internet text by following the procedures described below.

1) Find hits for certain relevant single words and MWEs irrespective of the semantic category they fall into. For example, the following could be relevant:

- *mutakuono*, *mutiainen*, *rättipää*, *sompanssi*, *neekeri*, and *ählämi* (derogatory words used for referring to non-white people),
- *hyysäri* (a derogatory word for a person who takes care of immigrants and worries about their living conditions) and *hyysätä* (the corresponding verb),
- *rotu* ("race"), *rodullinen*, and *rodullisesti* (the corresponding adjective and adverb),
- *suvakki* (a derogatory word for a person who advocates racial tolerance),
- *pohjasakka* ("scum"),
- *väriviallinen* (literally "colour defective"),
- *terroristi* ("terrorist"),
- *raiskata* ("to rape"), *raiskaaja* ("rapist"),
- *patriootti* ("patriot"), *isänmaa* ("home country"),
- *apina* ("monkey"), *eläin* ("animal"), *elukka* (colloquial form of *eläin*),
- *monikulttuurisuus* ("multiculturality"), and
- *etninen puhdistus* ("ethnic cleansing").

Perhaps also some proper names, which, according to my observations, often appear in connection with this type of writing, could provide useful clues for detecting questionable content. Examples of such names are Hitler, Breivik (Anders Behring Breivik, a Norwegian who committed the Norway 2011 attacks), and Halla-aho (Jussi Halla-aho, a Finnish politician and a present member of the European Parliament who has widely criticized the Finnish immigration policy and who was convicted of disturbing religious worship).

Even though the single words and MWEs listed above could be of interest here irrespective of the semantic category they fall into, they should somehow be labeled together as alarming and thus important patterns for the program to recognize. I would suggest grouping all of them under the same semantic tag for this particular application. Such a semantic tag could be, for example, A15-/E3-, in which the semantic tag A15- signifies risk and danger, while the semantic tag E3- signifies violence and anger. The tag A15-/E3- would be suitable for this purpose, since none of the entries in the present semantic lexical resources has been assigned this tag. An alternative solution would be to establish an entirely new semantic category for these single words and MWEs or subdivide further an existing category to allow an even finer-grained taxonomy.

2) Look for certain semantic tags

In addition to looking for certain predefined single words and MWEs within text, it would also be possible to make use of semantic field information to find relevant results. Such semantic tags which could be of interest in the proposed application and could provide useful clues for revealing the sentiments behind the text include, for example, the following:

- E3- (e.g. *kiduttaa* ("to torture"), *pistää vihaksi* ("to make someone angry")),

- L1- (e.g. *massamurha* ("mass murder"), *ottaa hengiltä* ("to kill")),
- G3 (e.g. *ampua* ("to shoot"), *Molotovin cocktail* ("Molotov cocktail")),
- S9 (e.g. *islam* ("Islam"), *kristitty* ("Christian")),
- Z2/S2 (e.g. *kurdi* ("Kurd"), *maahanmuuttaja* ("immigrant")), and
- X7+ (e.g. *aikoa* ("to intend"), *suunnitella* ("to plan"), *haluta* ("to want")).

5.3.2.1.2 Combining the relevant hits

The Internet content monitoring program could monitor the incoming messages in real-time. Such messages in which the program does not detect any questionable content could be passed on directly to the website. In turn, such messages in which the program discovers questionable content could be directed to the moderators for manual checking. In addition to the chronological order, it would also be practical to be able to organize these messages automatically in the order of urgency. This order could be determined with the aid of some statistical methods. Statistical methods could also be utilized for weighting the hits identified by the program to be able to detect the most questionable spots in the texts. The usefulness of weighting and the best solutions for carrying it out can be worked out only by experimenting with different approaches on test material. I believe that successful implementation of weighting mechanisms would significantly further improve the performance of the program and eliminate the possibility of false alarms. In addition, machine learning techniques could be applied to combine the potential features and learn which were most productive for that specific task.

5.3.2.1.3 Adapting the semantic lexical resources and the FST for "Internet language"

Yet one more issue to consider in the development of such applications is the often quite informal nature of "Internet language" which includes various colloquialisms in terms of vocabulary, spelling, and grammar. Even though the semantic lexical resources have been developed for the analysis of standard Finnish, in the final evaluation (see section 4.3.2) the lexical coverage on the test corpora including texts collected from online discussions ranged between 91.97% and 94.14%. Thus, the results were surprisingly good and proved that the Finnish semantic lexical resources could already be applied for the analysis of more informal writing contained on the Internet. Furthermore, in section 4.4.3.4.3 I brought up the possibility of further improving the results by training the program to cope with the features which differentiate "Internet language" from standard Finnish. This could be done, for example, by incorporating an additional semantic lexicon containing colloquial vocabulary and emoticons into the FST as well as VARD or similar mechanisms to help to deal with the spelling variation. Finally, spelling errors, which often appear in such text type, could also be addressed by using VARD or similar mechanisms, as was suggested in section 4.4.3.4.4.

5.3.2.2 Other Internet content monitoring applications

The guidelines described above could also be utilized in the development of other Internet content monitoring applications, such as for detecting people with suicidal ideation. To be able to recognize the relevant features, for instance, a corpus containing messages from online discussion forums and blog entries related to the given topic could be collected and analyzed, and literature and previous studies could also provide useful background knowledge. In addition, it would be helpful to consult specialists in the given field. Gathering the relevant

expressions is a challenging task; the search could not be limited to some obvious single words or MWEs, such as *itsemurha* ("suicide") or *tappaa itsensä* ("to kill onself"), but it must go much deeper, since these feelings can be expressed in various ways and much less explicitly, as will become evident from the examples below. For instance, the studies carried out by Ollikainen (1994) and Utriainen and Honkasalo (1996), which deal with suicide notes left by people who had taken their lives, could be very beneficial. I believe that the writings in suicide notes reflect the feelings of despairing people with suicidal ideation also in earlier phases of the suicide process and could provide valuable background information for the proposed application. For example, the following single words and MWEs, which are already included as entries in the Finnish semantic lexical resources, appeared frequently in the suicide notes discussed in Ollikainen (1994) and Utriainen and Honkasalo (1996): *hyvästi* ("goodbye"), *kiitos* ("thank you"), *umpikuja* ("dead end"), *yksin* ("alone"), *yksinäinen* ("lonely"), *antaa anteeksi* ("to forgive"), *antaa periksi* ("to give up"), *mennä pieleen* ("to flop"), and *päättää elämänsä* ("to end one's life"). Moreover, I would suggest writing MWE templates, for instance, for the following expressions, which were also mentioned in these two studies, and adding them to the semantic lexical resources as relevant patterns to be recognized:

- *ei jaksaa [enää]*[117] ("not to be able to take it [anymore]"),
- *olla [aivan/täysin] lopussa* ("to be [completely] finished"),
- *ei haluta [enää] elää* ("to have no desire for life [anymore]"),
- *olla liian heikko* ("to be too weak"),
- *ei olla päämäärää* ("to have no goal"),
- *liikaa paineita* ("too much pressure"),

[117] Square brackets indicate an optional element.

- *ei pystyä syömään/nukkumaan* (“not to be able to eat/sleep”), and
- *viimeinen toive* (“last wish”).

These single words and MWEs could be grouped, for example, under the semantic tag A15-/L1-, in which the semantic tag A15- signifies risk and danger and L1- signifies death and dying. The tag would be suitable for this purpose, since none of the entries in the present semantic lexical resources has been assigned this tag. Alternatively, a new semantic category could be created or an existing semantic category could be subdivided.

In addition to looking for certain predefined single words and MWEs within text to detect suicidal ideation, it would also be possible to utilize semantic field information to find relevant hits. The most obvious semantic tag to look for would doubtlessly be L1-. Examples of entries tagged as L1- in the existing semantic lexical resources are *hirttäytyä* (“to hang oneself”), *hukuttautua* (“to drown oneself”), *itsemurha* (“suicide”), *kuoliaaksi* (“to death”), *ampua kuula kalloonsa* (“to shoot a bullet into one’s head”), *ikuinen uni* (“eternal sleep”), and *nukkua pois* (“to pass away”, literally “to sleep away”). There are, moreover, many other semantic tags which could also be of interest in such an application and provide useful clues, for instance:

- A1.4- (e.g. *huono-onninen* (“unlucky”), *tuhoon tuomittu* (“doomed”)),
- A5.1- (e.g. *huono* (“bad”), *kurja* (“lousy”)),
- A11.1- (e.g. *mitätön* (“insignificant”), *samantekevä* (“unimportant”)),
- A12- (e.g. *kriisi* (“crisis”), *ongelma* (“problem”), *vaikeus* (“difficulty”)),
- B2- (e.g. *sairas* (“ill”), *särky* (“ache”), *lopen uupunut* (“totally exhausted”)),
- E4.1- (e.g. *itkeä* (“to cry”), *katua* (“to regret”), *masentunut* (“depressed”)),
- E4.2- (e.g. *pettymys* (“disappointment”), *tuskastunut* (“anguished”)),

- E5- (e.g. *hirvittää* (“to be terrified”), *järkytys* (“shock”), *pelko* (“fear”)),
- E6- (e.g. *ahdistunut* (“distressed”), *huolissaan* (“worried”), *stressi* (“stress”)),
- F2++ (e.g. *juopotella* (“to booze”), *kaatokänni* (“bender”)),
- F3 (e.g. *heroiini* (“heroin”), *polttaa pilveä* (”to smoke pot”)),
- G2.2- (e.g. *epäreilu* (“unfair”), *häpeällinen* (“shameful”), *vääryys* (“injustice”)),
- H4- (e.g. *asunnoton* (“roofless”), *koditon* (“homeless”)),
- I1.1- (e.g. *köyhä* (“poor”), *rahaton* (“penniless”)),
- I1.2 (e.g. *rästi* (“arrears”), *velkaantua* (“to get into debt”)),
- I3.1- (e.g. *työtön* (“unemployed”), *saada potkut* (“to get the sack”)),
- O4.2- (e.g. *nuhruinen* (“shabby”), *ruma* (“ugly”)),
- S1.1.3- (e.g. *syrjäytyä* (“to become marginalized”), *jättäytyä pois* (“to drop out”)).
- S1.2.5- (e.g. *avuton* (“helpless”), *heikko* (“weak”)),
- S1.2.6- (e.g. *järjetön* (“absurd”), *naiivi* (“naive”), *typerä* (“stupid”)),
- S4- (e.g. *avioero* (“divorce”), *äiditön* (“motherless”), *lapseton* (“childless”)),
- S5- (e.g. *yksin* (“alone”), *yksinäisyys* (“loneliness”)),
- S7.1- (e.g. *alistua* (“to capitulate”), *nöyrtyä* (“to humble oneself”), *tappio* (“defeat”)),
- S7.2- (e.g. *epäkunnioittava* (“disrespectful”), *nöyryytys* (“humiliation”)),
- T2- (e.g. *jäähyväiset* (“farewell”), *loppu* (”end”)),
- X5.2- (e.g. *kyllästynyt* (“bored”), *tylsä* (“tedious”)),
- X7- (e.g. *ei-toivottu* (“unwanted”), *hyljeksitty* (“rejected”)),
- X9.1- (e.g. *kyvytön* (“incapable”), *neuvoton* (“lost”), *tyhmä* (“stupid”)), and
- X9.2- (e.g. *epäonnistunut* (“unsuccessful”), *moka* (“blunder”)).

Following corresponding procedures, it would also be possible to tailor the FST and its semantic lexical resources, for example, for detecting:

- rape threats,
- paedophiles,
- hate speech characteristic of violent offenders, such as school shooters,
- hate speech targeted at sexual or other minorities, and
- cyberbullying, cyberharassment, and cyberstalking.

5.3.3 Psychological Profiling

The FST and its semantic lexicons could be tailored for the purposes of psychological profiling as well. In fact, the EST has already been tested for this task; it was used together with the Dictionary of Affect in Language (Whissell & Dewson, 1986) in a study by Hancock et al. (2013) which examined the features of crime narratives provided by psychopathic homicide offenders[118]. The results revealed that the psychopathic homicide offenders of the test group described their crimes, powerful emotional events, in an idiosyncratic manner. Their narratives contained an increased number of cause and effect statements, with a relatively high number of subordinating conjunctions, and a great number of references to basic physiological and self-preservation needs, such as eating, drinking, and money. They were less emotional and less positive, and they showed an emotional detachment in terms of a higher use of the past tense. High number of disfluencies indicated that the task of delineating one's crime is cognitively challenging, and the increased use of past tense and fewer present tense verbs indicated that they wanted to distance themselves from the murders. (Hancock et al., 2013, pp. 110–111)

If the FST and its semantic lexicons were used for similar purposes, it might be useful to expand their affective vocabulary; this would potentially obviate the need to use a

[118] Note also that LWIC (see section 2.3.2.2) has been used extensively for psychological profiling.

complementary dictionary of affect in language. For instance, the following semantic categories contain affective vocabulary which could be beneficial for this purpose:

E2	Liking	E6	Worry, Concern/Confident
E3	Calm/Violent/Angry	S7.2	Respect
E4.1	Happy/Sad: Happy	X5.2	Interest/Boredom/Excited/
E4.2	Happy/Sad: Contentment		Energetic
E5	Fear/Bravery/Shock		

All these categories utilize plus and minus markers to indicate a positive or a negative position on a semantic scale. Thus, to create different types of psychological profiles, the relevant combinations of semantic categories need to be recognized and, if necessary, expanded with relevant new entries for the task at hand. It would also be beneficial to train such an application to cope with the features of colloquial Finnish, for example, by incorporating an additional semantic lexicon containing colloquial vocabulary and emoticons into the FST as well as VARD or similar mechanisms to help to deal with the spelling variation. Possible spelling errors could also be addressed by using VARD or similar mechanisms (see section 5.3.2.1.3).

5.3.4 Sentiment Analysis

Finally, the FST and its semantic lexicons could also be redirected as a tool for sentiment analysis, for example, to help to analyze opinions expressed in online communication. It is exceedingly popular to write customer reviews and blog postings which express opinions on products and services, and since customer feedback on the Internet influences other customers' decisions, such user-created content has become an important source of

information for businesses when they develop marketing and product development plans (Lee, Jeong, & Lee, 2008, p. 230). The EST has already been tested for this purpose, in the classification of short text comments by sentiment in VoiceYourView (Simm et al, 2010). The results were promising, but the EST needs to be combined with other techniques in order to provide good sentiment classification.

The USAS category system contains a wealth of categories including sentiment-bearing nouns, verbs, adjectives, and adverbs as well as MWEs which could be useful in the development of sentiment analysis applications, for example:

A1.2	Suitability	E5	Fear/Bravery/Shock
A1.5.2	Usefulness	E6	Worry, Concern/Confident
A1.9	Avoiding	G2.2	General Ethics
A5.1	Evaluation: Good/Bad	I1.3	Money: Price
A5.3	Evaluation: Accuracy	O4.2	Judgement of Appearance
A5.4	Evaluation: Authenticity	S1.2.1	Approachability and Friendliness
A6.2	Comparing: Usual/Unusual	S1.2.4	Politeness
A6.3	Comparing: Variety	S1.2.5	Toughness: Strong/Weak
A11.1	Importance: Important	S1.2.6	Sensible
A11.2	Importance: Noticeability	S7.2	Respect
A12	Easy/Difficult	S8	Helping/Hindering
A15	Safety/Danger	X5.2	Interest/Boredom/Excited/
E2	Liking		Energetic
E3	Calm/Violent/Angry	X7	Wanting; Planning; Choosing
E4.1	Happy/Sad: Happy	X9.1	Ability: Ability, Intelligence
E4.2	Happy/Sad: Contentment	X9.2	Ability: Success and Failure

All these categories except A1.9 utilize plus and minus markers to indicate a positive or a negative position on a semantic scale. Furthermore, the following semantic categories could provide beneficial information:

A13.2	Degree: Maximizers	A13.6	Degree: Diminishers
A13.3	Degree: Boosters	A13.7	Degree: Minimizers
A13.4	Degree: Approximators	Z6	Negative
A13.5	Degree: Compromisers		

The semantic categories A13.2–A13.6 express degree, and thus they either intensify or downtone a sentiment which the single word or the MWE they relate to expresses (e.g. *äärettömän hyvä* ("extremely good"), *erittäin hyvä* ("very good"), and *melko hyvä* ("rather good")). The semantic tag Z6, in turn, negates the sentiment related to the single word or the MWE which it is associated with (e.g. *se ei ole hyvä* ("it is not good")).

The Finnish semantic lexicons already include a great number of entries in the above mentioned semantic categories; thus, this task would involve mostly using existing relevant semantic categories for analysis. If necessary, the semantic lexicons could be further expanded with vocabulary which is relevant for the task at hand as well as with emoticons. Furthermore, it would also be useful to train the program to cope with the features of colloquial Finnish and possible spelling errors in a similar way as was suggested in section 5.3.2.1.3.

5.4 Chapter Summary

In this chapter, I have drafted guidelines for the further development of the Finnish semantic lexical resources which are the main focus of this thesis. Thus, this chapter has

answered **RQ4 (What resources and methods can be useful for the further development of the Finnish semantic lexical resources, firstly, as a general language resource, and, secondly, when they are applied to new domains?)**.

I have started by suggesting ideas for expanding and updating both the single word lexicon and the MWE lexicon as general language resources. The single word lexicon, however, is already mature and has a good coverage, as the final evaluation of lexical coverage proved, but the MWE lexicon, in turn, is still a "preliminary version" which will become much more useful when it is expanded and when accurate templates are written for all its entries. Hence, I have suggested a solution for writing accurate templates to be able to reliably recognize and tag the different types of Finnish MWEs. I have also suggested the creation of a small autotagging lexicon similar to the autotagging lexicon included in the EST. An autotagging lexicon consists of fixed patterns which can have many possible instantiations. These expressions can be tagged effectively through the use of wild cards.

Expanding the MWE lexicon and writing templates for all its entries would be a key step in the further development of the semantic lexical resources as a general language resource. Carrying out this task would require a large investment of time and effort, but it would be essential if the semantic lexical resources were applied in the type of text analysis in which all single words and MWEs in text need to be recognized. However, in applications in which only certain relevant single words and MWEs would need to be recognized, the Finnish semantic lexical resources could already be put to practical use quite speedily and easily. I have briefly envisaged how to tailor the Finnish semantic lexical resources for specific domains and tasks, using named entity recognition, Internet content monitoring, psychological profiling, and sentiment analysis as example cases. These are tasks which require differing levels of adaptation in terms of extending the coverage of the semantic lexicons and utilizing the existing semantic categories.

6 Conclusions and Future Work

This final chapter provides the conclusions of this thesis. I will begin by summarizing the thesis and, thereafter, I will consider how it answers the research questions which I posed in chapter one. Subsequently, I will discuss the limitations of this work and will then proceed to look at the novel contributions which this thesis makes to the field. Finally, I will suggest ideas for further work and research.

6.1 Summary of the Work

In chapter one, I provided an introduction to this thesis by describing the context, objectives, and organization of the thesis as well as its significance. The overall objective of this thesis was to contribute to the development of semantic lexical resources for the Finnish language.

In chapter two, I established the background for this thesis. I began by defining the most important related concepts and, thereafter, reviewed some examples of semantic ontologies which represent the conceptual analysis method and the content analysis method. In addition, I briefly discussed some other, less related systems relying on semantic ontologies. Subsequently, I presented the UCREL Semantic Analysis System (USAS) which is another semantic ontology representing the conceptual analysis method. The first semantic tagger developed in the USAS framework, the EST, and the semantic lexical resources which it relies on have functioned as a model for the development of the Finnish counterparts. In addition, I introduced briefly other extensions to the USAS framework to which both the EST and the FST belong and which has during the past few years evolved into a multilingual semantic annotation system. I concluded the chapter with a brief account of the key points on

the Finnish language. In particular, I concentrated on those specific grammatical features of Finnish which have had an effect on the development of the FST, both in terms of the semantic lexical resources and of the software. These features are: rich morphology, productive use of compounding, and relatively free word order. I also briefly reviewed previous work and research on the creation of related lexical resources for Finnish. These are quite different from the semantic lexical resources discussed in this thesis. Firstly, the semantic lexicons developed for the FST use semantic fields as the organizing principle, while the other lexicons are built applying other organizing principles. Secondly, the semantic lexicons for the FST are intended for full text analysis and thus contain entries representing all parts of speech, whereas the other lexicons contain entries representing a limited set of parts of speech.

In chapter three, I detailed the development and the structure of the Finnish semantic lexical resources as well as the principles and practices which I followed in their creation. I built the semantic lexical resources for Finnish from scratch and made them compatible with the English semantic lexical resources, addressing the differences between the two languages. In addition, I clarified how these resources differ from each other both in terms of content and construction. Even though the Finnish semantic lexical resources are the main focus of this thesis, I also provided a brief summary of the development and the structure of the software component. This was necessary, because semantic lexical resources such as ours cannot be developed in isolation, but the software in which they will be applied needs to be taken into account in many respects all through the development process. With regard to the software component, it needed some modification, and the encoding system was changed to enable the program to process Finnish and the other new languages which were later incorporated in the USAS framework. In contrast, the semantic categories, which were originally created for the semantic analysis of English, did not need any modification at all, but they were found to be

entirely suitable for use in the semantic analysis of Finnish as well. An illustration of the output produced by the FST concluded this chapter.

In chapter four, I first summarized briefly the results of the formative evaluation which was carried out at the end of the Benedict project. Thereafter, I reported two new evaluations subsequent to extending and improving the Finnish semantic lexical resources after the Benedict project. Firstly, I carried out the final evaluation to measure the lexical coverage, in other words, the extent of the Finnish single word lexicon. Secondly, I carried out the application-based evaluation to measure the accuracy, in other words, how well the Finnish single word lexicon performs in the FST software. I described the test material which I had selected from various sources representing different genres, domains, and historical periods in order to ensure that the results of the evaluation would reflect the overall performance of the Finnish semantic lexicons in practical annotation tasks. The results revealed that the lexical coverage had clearly improved over the Benedict project, thanks to the new version of the single word lexicon; it indeed now covers the majority of the core Finnish vocabulary. The accuracy, however, had not improved since the Benedict project. The single word lexicon performed very well when applied in the FST, but it was evident that extending the single word lexicon is not alone sufficient to improve accuracy, but that it would also require further development of the MWE lexicon and, most of all, it would require investing in the development of the other components of the program. Consequently, I provided an analysis of the errors, and on the basis of it I suggested ideas for addressing the problems. I concluded the chapter by presenting the third evaluation of the new single word lexicon. This was referred to as the "semantic labeling experiment", and it measured how general native speakers of Finnish are able to replicate the categorisation of the Finnish single word lexicon. The results showed that the three participants had assigned to the sample words many more semantic tags than what is included in the gold standard, in other words, in the corresponding single word

lexicon entries. The results also showed that there was disagreement between the participants in their choice of the semantic tags for the sample words as well as in their choice for their frequency order. This indicates that it is a complex task to select senses for words and to define their order of likelihood.

In chapter five, I drafted guidelines for the future development of the Finnish semantic lexical resources. I proposed two different approaches for realizing this task. The first approach would be to continue to develop the semantic lexicons as a general language resource similar to the resource included in the EST. The single word lexicon and especially the MWE lexicon would benefit from expansion and updating as language changes over time, but it would be even more necessary to create accurate templates for all MWE lexicon entries to be able to reliably recognize and tag different types of Finnish MWEs. For this reason, I suggested solutions for carrying out this task. Furthermore, I proposed the creation of a small autotagging lexicon, similar to the autotagging lexicon included in the EST, to be able to effectively tag fixed patterns which can have many possible instantiations, such as measures. The second approach for the future development of the Finnish semantic lexical resources would be to tailor them for a specific purpose to deal with only one particular domain or task. I concluded the chapter by envisaging the creation of such applications, using named entity recogition, Internet content monitoring, psychological profiling, and sentiment analysis as example cases.

6.2 Research Questions Revisited

The main objective of this thesis has been the development of the Finnish semantic lexical resources; they function as the dictionary which the FST relies on. In order to meet this overall objective, I undertook an investigation as to whether and how it was possible to create

resources for Finnish which are compatible with the existing English semantic lexical resources while addressing the differences between the two languages.

Related to the main objective of this thesis, I addressed the four research questions, which I posed in chapter one, in the following way:

> **RQ1: What do the Finnish semantic lexical resources consist of, what type of principles and practices have been followed in their creation, and how do these resources differ from their English counterparts both in terms of content and construction?**

This research question was answered in chapter three.

I built the Finnish semantic lexical resources from scratch but applied the same semantic tagset as has been applied for the English counterparts and followed the same practices and principles. Furthermore, similarly to the English counterparts, the aim in the development of the Finnish semantic lexical resources was to build them primarily into a general language resource. With regard to the differences between the two languages, they were studied and addressed throughout the development process by designing the Finnish semantic lexical resources to meet the needs of semantic analysis of Finnish. As is the case with the English counterparts, the Finnish semantic lexical resources also consist of two separate lexicons: one consisting of single words and one consisting of MWEs. These lexicons were introduced in sections 3.4.1 and 3.4.2. At present, the single word lexicon contains 45,781 entries and the MWE lexicon contains 6,113 entries[119].

There are three significant differences between the semantic lexical resources for Finnish and English. Firstly, the English semantic lexicons contain both basic forms of words as well

[119] The single word lexicon was expanded for named entity purposes after the submission of this thesis (see section 5.3.1). The focus of lexicons described here is general language.

as their inflectional variants, whereas the Finnish counterparts consist of basic forms only. This is due to the fact that at the initial phase of the EST construction, the developers had no reliable automatic English lemmatiser available, and, therefore, they also had to include the inflected forms. For Finnish this approach would have been totally impossible due to its highly inflectional and agglutinative nature. Thus, the FST uses the Finnish morpho-syntactic analyser and parser TextMorfo for preprocessing to reduce Finnish words to basic forms, and only after that are these basic forms compared to the semantic lexicon entries which are also in basic form. Secondly, some Finnish MWEs were included in the single word lexicon and not in the MWE lexicon; such a phenomenon does not occur in the English semantic lexical resources. The reason for this was that because in the POS tagging phase TextMorfo processes some Finnish fixed expressions as single units and then assigns a POS tag to the entire expression, it was practical to treat these as single units in the semantic tagging component as well. Thirdly, there was no need to try to include all possible Finnish compounds exhaustively into the semantic lexical resources, since a component called the compound engine was developed in the FST software to process such compounds which are missing from the single word lexicon. The demand for such a function arose from the fact that in Finnish compounding is a very productive means of word formation, and trying to include all possible combinations of words would result in an unmanageable lexicon size. As a consequence, only the most commonly occurring compounds as well as lexicalized compounds are included in the single word lexicon, and all other possible, less frequently used compounds of a more temporary nature are handled by the compound engine. With the aid of TextMorfo and the compound engine, it is thus possible to recognize a vast number of different word forms in Finnish text.

Overall, there is a symbiotic relationship between the semantic lexicons and the software. Such lexicons cannot be developed in isolation but the software in which they will be applied needs to be taken into consideration in many respects throughout the development process.

RQ2: How extensive is the Finnish single word lexicon in terms of lexical coverage?

This research question was answered in section 4.3.

I showed in the final evaluation that the single word lexicon already has a very large lexical coverage. It was measured to range from 94.58% to 97.91% on the five different test corpora which contained general modern standard Finnish. The results are comparable to the results obtained with the English semantic lexical resources and thus indicate that the Finnish single word lexicon indeed covers the majority of core Finnish vocabulary. Furthermore, even though the Finnish single word lexicon was developed for the analysis of general modern standard Finnish text, it also performed surprisingly well in the analysis of domain-specific text (95.36%) and older Finnish text (92.11% to 93.05%) as well as when applied to the analysis of Internet discussions (91.97% to 94.14%) which often contain colloquialisms and spelling errors. This was shown to be comparable to the English semantic lexical resources where the coverage was calculated at 97.59% for general modern standard English, 95.38% for domain-specific text, and 94.40% for historical text.

RQ3: How suitable is the Finnish single word lexicon for use in the semantic analysis of Finnish in the FST software?

This research question was answered in section 4.4.

It was evident from the application-based evaluation that the accuracy had not improved over the accuracy obtained in the formative evaluation at the end of the Benedict project, even though the results of these two evaluations are not directly comparable. The amount of errors caused by missing single words was considerably smaller now than at the end of the Benedict project, thanks to the extended and updated version of the single word lexicon; thus, it has been shown that the lexicon performed well and has improved in suitability. However, other components of the software and insufficient disambiguation mechanisms caused various problems which need to be addressed to improve the accuracy. Obviously, further software development is outside the scope of this thesis. Moreover, the MWE lexicon needs expansion, and accurate templates need to be written for all its entries.

RQ4: What resources and methods can be useful for the further development of the Finnish semantic lexical resources, firstly, as a general language resource, and, secondly, when they are applied to new domains?

This research question was answered in chapter five.

To improve the Finnish semantic lexical resources as a general language resource, I suggested the following. Firstly, both the single word lexicon and the MWE lexicon need to be expanded and updated by adding new entries and by adding possible missing senses to the existing entries. I recommended various sources which could be helpful for this purpose. Secondly, all the MWE lexicon entries need to be written into accurate templates which would enable the FST to reliably recognize and tag different types of Finnish MWEs. I drafted guidelines for creating such templates and also proposed the creation of a small autotagging lexicon.

Another, considerably faster and easier way to put the Finnish semantic lexical resources to practical use would be to tailor them for a specific purpose to deal with only one particular domain or task. I drafted guidelines for tailoring the Finnish semantic lexical resources for specific domains and tasks, using named entity recognition, Internet content monitoring, psychological profiling, and sentiment analysis as example cases. These tasks require differing levels of adaptation; some tasks require more changes in the semantic lexicons, whereas some tasks require the identification of useful existing categories and less new entries in the semantic lexicons.

6.3 Limitations of the Work

The limitations of this work are most of all related to the reference sources which have been available for the development and for the evaluation of the Finnish semantic lexical resources.

There is no up-to-date, comprehensive frequency dictionary of Finnish which I could have used to ensure that all the relevant vocabulary is included in the semantic lexicons. Neither is there a dictionary in which the senses in the dictionary entries would have been arranged in frequency order. Such a dictionary would have been very helpful in choosing the order of the semantic tags in the semantic lexicon entries. Overall, as noted in section 2.5.3, Finnish lexicography is not as advanced as for other languages and does not offer a wealth of options. For this reason, I based my decisions for the most part on the electronic version of the *Kielitoimiston sanakirja* ("The New Dictionary of Modern Finnish", 2008), and before its publication I used the electronic version of the *Gummeruksen uusi suomen kielen sanakirja* ("The Gummerus New Dictionary of the Finnish Language", 1998). Additionally, I based my decisions on my intuition and on my work experience as a lexicographer.

There are no large balanced representative corpora available for Finnish, like, for example, the British National Corpus[120] and the Lancaster-Oslo/Bergen Corpus[121] are for English. Moreover, the selection of readily available free downloadable corpora containing clean data for Finnish is very limited because of copyright issues, especially with respect to modern Finnish. For this reason, I compiled the test corpora which I used in the evaluations of this thesis and chose various texts which largely reflected the text types included in the Lancaster-Oslo/Bergen Corpus. Hopefully in the future it is possible to test the Finnish semantic lexical resources with larger and better corpora.

The semantic lexicons are static and represent the language as it is now. Thus, over time as language changes, they will become less suitable. To counter this, the methods and guidelines I proposed in this thesis for further expanding and improving the semantic lexicons will deal with this limitation going forwards. In fact, I have already started this work and will discuss this in more detail in section 6.5.

The methods proposed in this thesis for creating semantic lexical resources might be applicable beyond this work only to European languages, given the nature of the software framework, existing corpus resources, and the grammar and syntax of those other languages. However, it has been proven already that this is not the case; based on the work for Finnish, the methods have now been tested to languages beyond Europe, such as Arabic, Urdu, and Chinese.

6.4 Novel Contributions

This thesis is an original contribution to the growing body of knowledge in the development of semantic analysis systems and, in particular, in the development of large

[120] For more information, see http://www.natcorp.ox.ac.uk/.

[121] For more information, see http://www.hit.uib.no/icame/lob/lob-dir.htm#lob4.

semantic lexical resources. The Finnish semantic lexicons, which were described in chapter three, are the first large-scale general purpose lexicons created for Finnish which are structured according to semantic field classifications. Thus, the thesis fills the gap in research by introducing these unique and valuable lexical resources for work and research involving the Finnish language. These semantic lexicons enable automatic semantic field analysis of large corpus data and, as such, can be very useful in various applications for NLP and corpus linguistics. They also offer various possibilities to develop multilingual applications utilizing the equivalent semantic taggers and semantic lexicons for the thirteen languages which belong in the USAS framework. The Finnish single word lexicon already has a good coverage, as became evident in the final evaluation presented in section 4.3, and the MWE lexicon is a good "preliminary version" which will become much more useful when it is expanded and when accurate templates are written for all its entries according to the suggestions presented in section 5.2.2.2. The single word lexicon is now available open source for research under the Creative Commons license at the USAS website[122].

In section 5.2, I drafted guidelines for the further development of the Finnish semantic lexicons as a general language resource. These guidelines include suggestions for expanding both lexicons, for writing accurate templates for different types of Finnish MWEs, and for creating a Finnish autotagging lexicon. The guidelines are applicable beyond the work described here and can be utilized, for example, in the development of the semantic lexical resources for semantic taggers in other languages.

In section 5.3, I drafted guidelines for utilizing and tailoring the Finnish semantic lexical resources for domain-specific applications. Firstly, I described how the Finnish semantic lexical resources can benefit named entity recognition and reported the expansion of the single word lexicon with personal and geographical names to meet these needs. Secondly, I

[122] For more information, see http://ucrel.lancs.ac.uk/usas/.

envisaged a semi-automatic Internet content monitoring program. Such an application could benefit the work of moderators in various social media websites and websites of newspapers and magazines by helping to detect different types of hate speech which is a severe and constantly growing problem in Finland. The proposed application could also be of help to the police and other law enforcement agencies as well as to healthcare authorities and organizations for detecting people with suicidal ideation, paedophiles, rape threats, cyberbullying, cyberharassment, and cyberstalking. Overall, thanks to the flexibility of the USAS category system, the FST and its semantic lexical resources can easily be adapted to suit different contexts and domain-specific applications, even though they have originally been created for the analysis of general language.

In chapter four, I presented three different types of quantitative evaluations which can be used for the evaluation of semantic lexical resources. Two of these are corpus-based methods which combine both theory and practical application. The third is an evaluation of whether general native users of a language can easily replicate the semantic categorisation used in the semantic lexical resources.

The FST was the first non-English version of the EST, and it facilitated further experiments with the semantic category system. The experiences gained of both the lexicon and the software development of the FST have already been very useful when the USAS framework has been expanded to new languages, especially to languages which, like Finnish, are highly inflectional. This is the first PhD thesis which investigates the creation of semantic lexical resources for a semantic tagger belonging in the USAS framework. The insights and conclusions presented here will further benefit the development of equivalent lexicons for other languages, and I hope that this work will also encourage other researchers to investigate the topic in their theses. Finally, I hope that this thesis encourages various types of semantic applications for Finnish, both monolingual and multilingual.

6.5 Future Directions

The overall aim in the development process of the FST and its semantic lexical resources was to adapt the semantic analysis system originally developed for English to meet the needs of semantic analysis of Finnish. This setting offers various interesting possibilities for future work and research.

The EST has been used successfully in many corpus linguistics and NLP applications which were listed in section 2.4.1. It would now be possible to apply the equivalent FST for similar purposes. Of these, I would find sentiment analysis particularly interesting, and I would also like to continue this work in the context of developing the type of an Internet content monitoring program which was suggested in section 5.3.2. I believe that there is a dire social need for such applications. The work which I started to expand the semantic lexicons for NER purposes will continue, and in the following phase I will add foreign personal names and geographical names as well as names of both Finnish and foreign companies, organizations, institutions, and trademarks. I will make the resulting NER lexicons available open source for research under the Creative Commons license at the USAS website. Furthermore, I have already added new entries representing general modern standard Finnish to the semantic lexicons and edited some of the existing entries following the guidelines proposed in this thesis. I will continue with this work as well.

The USAS framework has now been extended for Czech, Chinese, Dutch, French, Italian, Malay, Portuguese, Russian, Spanish, Urdu, and Welsh as well[123]. As a result of these efforts, we now have at our disposal a package of equivalent semantic taggers based on equivalent semantic lexicons which are capable of processing all these languages. The equivalent structure enables the development of multilingual applications, since the semantic tagset acts as a kind of a "meta-dictionary" or "lingua franca" between the languages. This would make it

[123] There are plans to extend the USAS framework next for Arabic, Norwegian, and Swedish.

possible to use these semantic taggers, for example, for the purposes of machine translation and crosslingual plagiarism detection. Moreover, it would be intriguing to apply the semantic taggers for cross-lingual information extraction. In fact, in March 2016, the BBC organized a multilingual NewsHACK event themed "Multilingual Journalism: Tools for Future News" in which they offered an opportunity for teams of language technology researchers to work with their own tools with multilingual data from the BBC's connected studio. "Team 1" from Lancaster University was represented by Paul Rayson, Scott Piao, and Hugo Sanjurjo González. They used the USAS semantic taggers for English, Chinese, and Spanish and built a prototype tool named "Multilingual Reality Check" to bridge related news stories across these languages. As a result, journalists can simply click on news stories in the system, and the system will show them related articles in the other languages, ranked in order of relevance. (ESRC Centre for Corpus Approaches to Social Science, 2016) A similar application including the FST and many more languages might be found very useful in Finland as well.

Two bilingual applications utilizing the USAS semantic taggers already exist. The first bilingual application was the context-sensitive dictionary search tool for English and Finnish which we developed in the Benedict project. The second bilingual application was the automatic semantic assistance tool for translators developed in the ASSIST project which utilized the English and Russian Semantic Taggers. It would now be possible to try the USAS semantic taggers in such applications between the many more language pairs.

Appendix A: USAS Semantic Tagset in English

A GENERAL & ABSTRACT TERMS
A1 General
A1.1.1 General Actions, Making, etc.
A1.1.2 Damaging and Destroying
A1.2 Suitability
A1.3 Caution
A1.4 Chance, Luck
A1.5 Use
A1.5.1 Using
A1.5.2 Usefulness
A1.6 Physical/Mental
A1.7 Constraint
A1.8 Inclusion/Exclusion
A1.9 Avoiding
A2 Affect
A2.1 Affect: Modify, Change
A2.2 Affect: Cause, Connected
A3 Being
A4 Classification
A4.1 Generally Kinds, Groups, Examples
A4.2 Particular/General; Detail
A5 Evaluation
A5.1 Evaluation: Good/Bad
A5.2 Evaluation: True/False
A5.3 Evaluation: Accuracy
A5.4 Evaluation: Authenticity
A6 Comparing
A6.1 Comparing: Similar/Different
A6.2 Comparing: Usual/Unusual
A6.3 Comparing: Variety
A7 Definite (+ Modals)
A8 Seem/Appear
A9 Getting and Giving; Possession
A10 Open/Closed; Hiding/Hidden; Finding; Showing
A11 Importance
A11.1 Importance: Important
A11.2 Importance: Noticeability
A12 Easy/Difficult
A13 Degree
A13.1 Degree: Non-Specific
A13.2 Degree: Maximizers
A13.3 Degree: Boosters
A13.4 Degree: Approximators
A13.5 Degree: Compromisers
A13.6 Degree: Diminishers
A13.7 Degree: Minimizers
A14 Exclusivizers/Particularizers
A15 Safety/Danger

B THE BODY & THE INDIVIDUAL
B1 Anatomy and Physiology
B2 Health and Disease
B3 Medicines and Medical Treatment
B4 Cleaning and Personal Care
B5 Clothes and Personal Belongings

C ARTS & CRAFTS
C1 Arts and Crafts

E EMOTIONAL ACTIONS, STATES, & PROCESSES
E1 General
E2 Liking
E3 Calm/Violent/Angry
E4 Happy/Sad
E4.1 Happy/Sad: Happy
E4.2 Happy/Sad: Contentment
E5 Fear/Bravery/Shock
E6 Worry, Concern/Confident

F FOOD & FARMING
F1 Food
F2 Drinks
F3 Cigarettes and Drugs
F4 Farming and Horticulture

G GOVERNMENT & THE PUBLIC DOMAIN
G1 Government, Politics, and Elections
G1.1 Government etc.
G1.2 Politics
G2 Crime, Law, and Order
G2.1 Crime, Law, and Order: Law and Order
G2.2 General Ethics
G3 Warfare, Defence, and the Army; Weapons

H ARCHITECTURE, BUILDINGS, HOUSES, & THE HOME
H1 Architecture, Kinds of Houses and Buildings
H2 Parts of Buildings
H3 Areas Around or Near Houses
H4 Residence
H5 Furniture and Household Fittings

I MONEY & COMMERCE
I1 Money Generally
I1.1 Money: Affluence
I1.2 Money: Debts
I1.3 Money: Price
I2 Business
I2.1 Business: Generally
I2.2 Business: Selling
I3 Work and Employment
I3.1 Work and Employment: Generally
I3.2 Work and Employment: Professionalism
I4 Industry

K ENTERTAINMENT, SPORTS, & GAMES
K1 Entertainment Generally
K2 Music and Related Activities
K3 Recorded Sound etc.
K4 Drama, the Theatre, and Show Business
K5 Sports and Games Generally
K5.1 Sports
K5.2 Games
K6 Children's Games and Toys

L LIFE & LIVING THINGS
L1 Life and Living Things
L2 Living Creatures Generally
L3 Plants

M MOVEMENT, LOCATION, TRAVEL, & TRANSPORT
M1 Moving, Coming, and Going
M2 Putting, Taking, Pulling, Pushing, Transporting, etc.
M3 Movement/Transportation: Land
M4 Movement/Transportation: Water
M5 Movement/Transportation: Air
M6 Location and Direction
M7 Places
M8 Remaining/Stationary

N NUMBERS & MEASUREMENT
N1 Numbers
N2 Mathematics
N3 Measurement
N3.1 Measurement: General
N3.2 Measurement: Size
N3.3 Measurement: Distance
N3.4 Measurement: Volume
N3.5 Measurement: Weight
N3.6 Measurement: Area
N3.7 Measurement: Length and Height
N3.8 Measurement: Speed
N4 Linear Order
N5 Quantities
N5.1 Entirety; Maximum
N5.2 Exceeding; Waste
N6 Frequency etc.

O SUBSTANCES, MATERIALS, OBJECTS, & EQUIPMENT
O1 Substances and Materials Generally
O1.1 Substances and Materials Generally: Solid
O1.2 Substances and Materials Generally: Liquid
O1.3 Substances and Materials Generally: Gas
O2 Objects Generally
O3 Electricity and Electrical Equipment
O4 Physical Attributes
O4.1 General Appearance and Physical Properties
O4.2 Judgement of Appearance
O4.3 Colour and Colour Patterns
O4.4 Shape
O4.5 Texture
O4.6 Temperature

P EDUCATION
P1 Education in General

Q LINGUISTIC ACTIONS, STATES, & PROCESSES
Q1 Communication
Q1.1 Communication in General
Q1.2 Paper Documents and Writing
Q1.3 Telecommunications
Q2 Speech Acts
Q2.1 Speech etc.: Communicative
Q2.2 Speech Acts
Q3 Language, Speech, and Grammar
Q4 The Media
Q4.1 The Media: Books
Q4.2 The Media: Newspapers etc.
Q4.3 The Media: TV, Radio, and Cinema

S SOCIAL ACTIONS, STATES, & PROCESSES
S1 Social Actions, States, and Processes
S1.1 Social Actions, States, and Processes
S1.1.1 General
S1.1.2 Reciprocity
S1.1.3 Participation
S1.1.4 Deserve etc.
S1.2 Personality Traits
S1.2.1 Approachability and Friendliness
S1.2.2 Avarice
S1.2.3 Egoism
S1.2.4 Politeness
S1.2.5 Toughness: Strong/Weak
S1.2.6 Sensible
S2 People
S2.1 People: Female
S2.2 People: Male
S3 Relationship
S3.1 Relationship: General
S3.2 Relationship: Intimate/Sexual
S4 Kin
S5 Groups and Affiliation
S6 Obligation and Necessity
S7 Power Relationship
S7.1 Power, Organizing
S7.2 Respect
S7.3 Competition
S7.4 Permission
S8 Helping/Hindering
S9 Religion and the Supernatural

T TIME
T1 Time
T1.1 Time: General
T1.1.1 Time: General: Past
T1.1.2 Time: General: Present; Simultaneous
T1.1.3 Time: General: Future
T1.2 Time: Momentary
T1.3 Time: Period
T2 Time: Beginning and Ending
T3 Time: Old, New, and Young; Age
T4 Time: Early/Late

W THE WORLD & OUR ENVIRONMENT
W1 The Universe
W2 Light
W3 Geographical Terms
W4 Weather
W5 Green Issues

X PSYCHOLOGICAL ACTIONS, STATES, & PROCESSES
X1 General
X2 Mental Actions and Processes
X2.1 Thought, Belief
X2.2 Knowledge
X2.3 Learn
X2.4 Investigate, Examine, Test, Search
X2.5 Understand
X2.6 Expect
X3 Sensory
X3.1 Sensory: Taste
X3.2 Sensory: Sound
X3.3 Sensory: Touch
X3.4 Sensory: Sight
X3.5 Sensory: Smell
X4 Mental Object
X4.1 Mental Object: Conceptual Object
X4.2 Mental Object: Means, Method
X5 Attention
X5.1 Attention
X5.2 Interest/Boredom/Excited/Energetic
X6 Deciding
X7 Wanting; Planning; Choosing
X8 Trying
X9 Ability
X9.1 Ability: Ability, Intelligence
X9.2 Ability: Success and Failure

Y SCIENCE & TECHNOLOGY
Y1 Science and Technology in General
Y2 Information Technology and Computing

Z NAMES & GRAMMATICAL WORDS
Z0 Unmatched Proper Noun
Z1 Personal Names
Z2 Geographical Names
Z3 Other Proper Names
Z4 Discourse Bin
Z5 Grammatical Bin
Z6 Negative
Z7 If
Z8 Pronouns etc.
Z9 Trash Can
Z99 Unmatched

Appendix B: USAS Semantic Tagset in Finnish

A YLEISET & ABSTRAKTIT SANAT
A1Yleiset sanat
A1.1.1 Toiminta ja tekeminen
A1.1.2 Vahingoittaminen ja tuhoaminen
A1.2 Soveltuvuus
A1.3 Varovaisuus
A1.4 Sattuma ja tuuri
A1.5 Käyttö
A1.5.1 Käyttäminen
A1.5.2 Hyödyllisyys
A1.6 Aineellisuus/käsitteellisyys
A1.7 Rajoittaminen
A1.8 Mukaan ottaminen / pois jättäminen
A1.9 Välttäminen
A2 Vaikutus
A2.1 Vaikutus: muuttaminen ja muuttuminen
A2.2 Vaikutus: syy ja seuraus
A3 Olemassaolo
A4 Luokitteleminen
A4.1 Laji, tyyppi ja esimerkki
A4.2 Erityisyys/yleisluontoisuus; yksityiskohtaisuus
A5 Arvioiminen
A5.1 Arvioiminen: hyvä/huono
A5.2 Arvioiminen: tosi/epätosi
A5.3 Arvioiminen: virheettömyys ja tarkkuus
A5.4 Arvioiminen: aitous
A6 Vertaileminen
A6.1 Vertaileminen: samanlainen/erilainen
A6.2 Vertaileminen: tavallinen/epätavallinen
A6.3 Vertaileminen: monipuolisuus ja vaihtelevuus
A7 Mahdollisuus ja välttämättömyys
A8 Vaikutelma
A9 Saaminen ja antaminen; omistaminen
A10 Avoin/suljettu; piilottaminen, löytäminen ja näyttäminen
A11 Tärkeys
A11.1 Tärkeys: tärkeä
A11.2 Tärkeys: huomattava
A12 Helppous/vaikeus
A13 Aste
A13.1 Aste (yleiskategoria)
A13.2 Aste: maksimoiminen
A13.3 Aste: vahvistaminen
A13.4 Aste: likimääräisyys
A13.5 Aste: suhteellisuus
A13.6 Aste: osittaisuus
A13.7 Aste: minimoiminen
A14 Rajaaminen/täsmentäminen
A15 Turvallisuus/vaarallisuus
B KEHO & IHMINEN
B1 Anatomia ja fysiologia
B2 Terveys ja sairaus
B3 Lääkkeet ja sairaanhoito
B4 Siivous ja henkilökohtainen hygienia
B5 Vaatteet ja henkilökohtaiset tavarat
C TAIDE & KÄSITYÖ
C1 Taide ja käsityö
E TUNNE-ELÄMÄ & MIELENTILAT
E1 Tunne-elämä ja mielentilat (yleiskategoria)
E2 Mieltymys
E3 Rauhallisuus / väkivaltaisuus, vihaisuus
E4 Onnellisuus/surullisuus
E4.1 Onnellisuus/surullisuus: onnellisuus
E4.2 Onnellisuus/surullisuus: tyytyväisyys
E5 Pelko, järkytys / rohkeus
E6 Huoli/huolettomuus
F RAVINTO & MAATALOUS
F1 Ruoka
F2 Juoma
F3 Tupakka ja huumeet
F4 Maatalous ja puutarhanhoito
G HALLINTO, POLITIIKKA & LAKI
G1 Hallinto, politiikka ja vaalit
G1.1 Hallinto
G1.2 Politiikka
G2 Rikos, laki ja järjestys
G2.1 Rikos, laki ja järjestys: laki ja järjestys
G2.2 Etiikka
G3 Sodankäynti, puolustus ja armeija; aseet
H ARKKITEHTUURI, RAKENNUKSET & KOTI
H1 Arkkitehtuuri ja rakennukset
H2 Rakennusten osat
H3 Rakennusten lähialueet
H4 Asuminen ja oleskelu
H5 Huonekalut ja sisustus
I RAHA & LIIKETOIMINTA
I1 Raha (yleiskategoria)
I1.1 Raha: varakkuus
I1.2 Raha: velka
I1.3 Raha: hinta
I2 Liiketoiminta
I2.1 Liiketoiminta (yleiskategoria)
I2.2 Liiketoiminta: myyminen
I3 Työ ja työllisyys
I3.1 Työ ja työllisyys (yleiskategoria)
I3.2 Työ ja työllisyys: ammattimaisuus
I4 Teollisuus
K VIIHDE, URHEILU & PELIT
K1 Viihde (yleiskategoria)
K2 Musiikki
K3 Musiikin tallentaminen
K4 Näyttämötaide ja viihdeteollisuus
K5 Urheilu ja pelit
K5.1 Urheilu
K5.2 Pelit
K6 Lasten leikit ja lelut
L ELÄMÄ & ELOLLINEN LUONTO
L1 Elämä
L2 Eläimet
L3 Kasvit
M LIIKKUMINEN, SIJAINTI, MATKAILU & KULJETUS
M1 Liikkuminen, tuleminen ja meneminen
M2 Laittaminen, ottaminen, vetäminen, työntäminen ja kuljettaminen
M3 Liikkuminen/kuljettaminen maalla
M4 Liikkuminen/kuljettaminen vedessä
M5 Liikkuminen/kuljettaminen ilmassa
M6 Sijainti ja suunta
M7 Paikat ja alueet
M8 Pysyminen/liikkumattomuus
N NUMEROT & MITTAAMINEN
N1 Numerot
N2 Matematiikka
N3 Mittaaminen
N3.1 Mittaaminen (yleiskategoria)
N3.2 Mittaaminen: koko
N3.3 Mittaaminen: etäisyys
N3.4 Mittaaminen: tilavuus
N3.5 Mittaaminen: paino
N3.6 Mittaaminen: pinta-ala
N3.7 Mittaaminen: pituus ja korkeus
N3.8 Mittaaminen: nopeus
N4 Järjestys
N5 Määrä
N5.1 Kokonaisuus ja enimmäismäärä
N5.2 Liiallisuus ja ylimääräisyys; jätteet
N6 Yleisyys
O AINEET, ESINEET & TARVIKKEET
O1 Aineet (yleiskategoria)
O1.1 Kiinteät aineet
O1.2 Nestemäiset aineet
O1.3 Kaasumaiset aineet
O2 Esineet (yleiskategoria)
O3 Sähkö ja sähkölaitteet
O4 Fyysiset ominaisuudet
O4.1 Ulkoasu ja fyysiset ominaisuudet
O4.2 Ulkoasun arviointi
O4.3 Värit ja kuviot
O4.4 Muoto
O4.5 Tuntu ja rakenne
O4.6 Lämpötila
P KOULUTUS
P1 Koulutus
Q KIELELLISET TOIMINNOT & PROSESSIT
Q1 Kommunikaatio
Q1.1 Kommunikaatio (yleiskategoria)
Q1.2 Kirjalliset dokumentit ja kirjoittaminen
Q1.3 Televiestintä
Q2 Puheaktit
Q2.1 Kommunikatiiviset puheaktit
Q2.2 Puheaktit (yleiskategoria)
Q3 Kieli, puhe ja kielioppi
Q4 Viestimet
Q4.1 Viestimet: kirjat
Q4.2 Viestimet: lehdistö
Q4.3 Viestimet: TV, radio ja elokuvat
S SOSIAALISET TOIMINNOT & PROSESSIT
S1 Sosiaaliset toiminnot ja prosessit
S1.1 Sosiaaliset toiminnot ja prosessit
S1.1.1 Sosiaaliset toiminnot ja prosessit (yleiskategoria)
S1.1.2 Vastavuoroisuus
S1.1.3 Osallistuminen
S1.1.4 Ansaitseminen
S1.2 Luonteenpiirteet
S1.2.1 Ystävällisyys
S1.2.2 Ahneus
S1.2.3 Itsekkyys
S1.2.4 Kohteliaisuus
S1.2.5 Kovuus: vahva/heikko
S1.2.6 Järkevyys
S2 Ihmiset
S2.1 Naiset
S2.2 Miehet
S3 Sosiaaliset suhteet
S3.1 Sosiaaliset suhteet (yleiskategoria)
S3.2 Sosiaaliset suhteet: intiimit suhteet ja seksi
S4 Perhe ja suku
S5 Ryhmät ja kytkökset
S6 Velvollisuus ja välttämättömyys
S7 Valtasuhteet
S7.1 Valta ja järjesteleminen
S7.2 Kunnioittaminen
S7.3 Kilpaileminen
S7.4 Salliminen
S8 Auttaminen/estäminen
S9 Uskonto ja yliluonnollinen
T AIKA
T1 Aika
T1.1 Aika (yleiskategoria)
T1.1.1 Aika: mennyt aika
T1.1.2 Aika: nykyaika ja samanaikaisuus
T1.1.3 Aika: tulevaisuus
T1.2 Aika: hetki
T1.3 Aika: ajanjakso
T2 Aika: alkaminen ja loppuminen
T3 Aika: vanha / uusi ja nuori; ikä
T4 Aika: aikainen/myöhäinen
W MAAILMA & YMPÄRISTÖ
W1 Maailmankaikkeus
W2 Valo
W3 Maantiede
W4 Sää
W5 Ympäristöasiat
X PSYKOLOGISET TOIMINNOT, TILAT & PROSESSIT
X1 Psykologiset toiminnot, tilat ja prosessit (yleiskategoria)
X2 Mielen toiminnot ja prosessit
X2.1 Ajatteleminen ja uskominen
X2.2 Tietäminen
X2.3 Oppiminen
X2.4 Tutkiminen, testaaminen ja etsiminen
X2.5 Ymmärtäminen
X2.6 Odottaminen ja ennakoiminen
X3 Aistit
X3.1 Aistit: maku
X3.2 Aistit: kuulo
X3.3 Aistit: tunto
X3.4 Aistit: näkö
X3.5 Aistit: haju
X4 Mielen sisällöt
X4.1 Mielen sisällöt: käsitteelliset sisällöt
X4.2 Mielen sisällöt: keinot ja menetelmät
X5 Huomio
X5.1 Huomioiminen
X5.2 Kiinnostuneisuus, innostuneisuus ja energisyys / ikävystyneisyys
X6 Päättäminen
X7 Haluaminen, suunnitteleminen ja valitseminen
X8 Yrittäminen
X9 Taito
X9.1 Taito: kyvykkyys ja älykkyys
X9.2 Taito: onnistuminen ja epäonnistuminen
Y TIEDE & TEKNOLOGIA
Y1 Tiede ja teknologia (yleiskategoria)
Y2 Tietotekniikka
Z NIMET & KIELIOPILLISET SANAT
Z0 Sanastoista puuttuvat erisnimet
Z1 Henkilönimet
Z2 Maantieteelliset nimet
Z3 Muut erisnimet
Z4 Huudahdukset, fraasit jne.
Z5 Funktiosanat
Z6 Kieltoa ilmaisevat sanat
Z7 Ehdollisuutta ilmaisevat sanat
Z8 Pronominit jne.
Z9 Roskakori
Z99 Sanastoista puuttuvat sanat

Appendix C: Semantic Categories of the USAS Tagset with Prototypical Examples from the Finnish Semantic Lexical Resources

This appendix is an adaptation of Introduction to the USAS Category System written by Dawn Archer, Andrew Wilson and Paul Rayson (2002). It displays the 21 top level semantic categories as well as their 232 subcategories with various prototypical Finnish language examples of both single words and MWEs. A prototypical example here means that the entry has been considered unambiguous and to belong into that category only, in other words, the entry has been assigned only the semantic tag in question. The pluses and minuses indicate antonymous pairs or a positive or negative position on a semantic scale, and the markers "f" and "m" indicate females and males respectively. For more information on the semantic tagset and on the principles and the practices followed in the lexicon construction, see section 3.4.

A GENERAL & ABSTRACT TERMS

A1 GENERAL

Entries are sub-classified into the following:

A1.1.1 GENERAL ACTIONS, MAKING, ETC.

Abstract terms relating to an activity/action, a characteristic/feature, a construction/craft, and/or the action of constructing/crafting

Prototypical examples:

askare, automatisoida, esikäsittely, hääräämineninen, kivetä, laatia, manuaalinen, poraus, soseuttaa, suorite, tekeillä, toiminnallinen, työstö

painaa hommia, suora toiminta, ei tehdä elettäkään (-), tumput suorina (-)

A1.1.2 DAMAGING AND DESTROYING

Abstract terms depicting damage/destruction/demolition/pollution etc.

Prototypical examples:

epäkuntoinen, halkeama, kolaroida, laho, pilalla, rikki, ränsistynyt, tuhoisa, vaurio, viallisuus

hajottaa alkutekijöihinsä, pohjaan palanut

A1.2 SUITABILITY

Abstract terms relating to appropriateness, suitability, aptness, etc.

Prototypical examples:

asianmukaisesti (+), salonkikelpoisuus (+), relevantti (+), soveltua (+), asiattomasti (-), epätarkoituksenmukainen (-), kelvottomuus (-), vaalikelvoton (-)

kuin luotu (+), olla omiaan (+)

A1.3 CAUTION

Abstract terms relating to vigilance/care/prudence or the lack of

Prototypical examples:

harkiten (+), huolellinen (+), turvatoimi (+), varmuuden_vuoksi (+), varoa (+), huolimaton (-), impulsiivisuus (-), varomattomuus (-)

kieli keskellä suuta (+), pitää varansa (+)

A1.4 CHANCE, LUCK

Abstract terms depicting likelihood/probability/providence or the lack of

Prototypical examples:

kohtalo, sallimus, sattumalta, hyväonninen (+), onnenpotku (+), välttyä (+), huono-osainen (-), ikävä_kyllä (-), suuronnettomuus (-), tapaturmaisesti (-)

pitää peukkua, sattuman kaupalla, moukan tuuri (+), käydä kalpaten (-)

A1.5 USE

Entries are sub-classified into the following:

A1.5.1 USING

Abstract terms denoting use or the lack of

Prototypical examples:

ergonominen, hyödyntää, käytettävyys, käyttöönotto, soveltaa, käyttämättömyys (-), lepotilainen (-), pois_käytöstä (-), säästellä (-)

käyttää hyödyksi, tyhjän panttina (-)

A1.5.2 USEFULNESS

Abstract terms denoting usefulness or the lack of

Prototypical examples:

funktionaalinen (+), hyödyttää (+), monikäyttöisyys (+), yleishyödyllinen (+), hyödyttömyys (-), käyttökelvoton (-)

kuin taivaan lahja (+), maksaa vaivan (+), joutaa kaatopaikalle (-)

A1.6 PHYSICAL/MENTAL

Abstract terms denoting (level of) practicality/abstraction

Prototypical examples:

abstraktio, aineeton, konkretisoida, käytännönläheisesti, maallinen, materialismi, teoreettisuus

A1.7 CONSTRAINT

Abstract terms denoting (level of) restriction/autonomy

Prototypical examples:

ansa (+), kahle (+), karanteeni (+), lukittua (+), telkeäminen (+), irti (-), kaoottinen (-), rajoituksettomasti (-), ryöstäytyä (-)

ottaa koppi (+), päästä pakoon (-)

A1.8 INCLUSION/EXCLUSION

Abstract terms denoting (level of) inclusion/exclusion

Prototypical examples:

ainesosa (+), koostua (+), lukeutua (+), mukaan_lukien (+), oheis (+), sisällyttää (+), poissa_laskuista (-)

vetää mukaan (+), joutua hyllylle (-)

A1.9 AVOIDING

Abstract terms denoting (level of) avoidance/evasion etc.

Prototypical examples:

kaihtaminen, karttaa, pinnata, välttely

antaa asian olla, pitää näppinsä erossa

A2 AFFECT

Entries are sub-classified into the following:

A2.1 AFFECT: MODIFY, CHANGE

Abstract terms denoting (propensity for) change

Prototypical examples:

evoluutio (+), modifiointi (+), mukautua (+), muuntautumis (+), säädettävä (+), ennallaan (-), koskematon (-), muuttumaton (-)

ei viru eikä vanu (-)

A2.2 AFFECT: CAUSE, CONNECTED

Abstract terms denoting causal relationship or the lack of

Prototypical examples:

aiheuttaa, ansiosta, kausaliteetti, koitua, peruste, riippuen, tämän_takia, vuoksi, välillisesti

alku ja juuri, antaa aihetta

A3 BEING

Abstract terms relating to being/existing

Prototypical examples:

eksistenssi (+), esiintymä (+), läsnä (+), olemassa_oleva (+)

olla maisemissa (+), ei mailla eikä halmeilla (-)

A4 CLASSIFICATION

Entries are sub-classified into the following:

A4.1 GENERALLY KINDS, GROUPS, EXAMPLES

Abstract terms denoting types, groups, examples

Prototypical examples:

alakohtainen, erittely, lajitella, leimallinen, näytekappale, versio

A4.2 PARTICULAR/GENERAL; DETAIL

Abstract terms denoting (level of) generality/detail

Prototypical examples:

erikoistua (+), erityispiirre (+), määrätynlainen (+), ominais (+), tunnusomaisesti (+), yksityiskohta (+), geneerisyys (-), yleisesti_ottaen (-), yleisluonteinen (-)

kyseessä oleva (+), mennä asiaan (+)

A5 EVALUATION

Entries are sub-classified into the following:

A5.1 EVALUATION: GOOD/BAD

Evaluative terms depicting quality

Prototypical examples:

evaluointi, keskiarvo, pisteyttää, tasoinen, mallikelpoinen (+), sujuvasti (+), entistä_paremmin (++), vertaansa_vailla (+++), epäkohta (-), kehno (-), välttävästi (-), huonommuus (--), katastrofaalinen (---)

erota edukseen (+), lyödä laudalta (++), olla omaa luokkaansa (+++), ei olla kaksinen (-), heikoin lenkki (---)

A5.2 EVALUATION: TRUE/FALSE

Evaluative terms depicting truth

Prototypical examples:

oikeassa (+), suoraan_sanoen (+), tosiasiallinen (+), totuus (+), verifioida (+), epärehellinen (-), harhauttaa (-), illuusio (-), puuta_heinää (-), valheellisesti (-)

käsi sydämellä (+), pitää paikkansa (+), tekaista omasta päästään (-), valkoinen valhe (-)

A5.3 EVALUATION: ACCURACY

Evaluative terms depicting accuracy

Prototypical examples:

eksaktisti (+), oikeaan_osuva (+), sananmukainen (+), tarkkaan_ottaen (+), täsmennys (+), erehdys (-), haksahtaa (-), silmämääräinen (-), virhepäätelmä (-)

naulan kantaan (+), olla oikeilla jäljillä (+), asian vierestä (-)

A5.4 EVALUATION: AUTHENTICITY

Evaluative terms depicting authenticity

Prototypical examples:

aitous (+), alkuperäiskappale (+), autenttinen (+), keinotekoinen (-), teennäisesti (-), väärentäminen (-)

olla oma itsensä (+), koreilla lainahöyhenissä (-)

A6 COMPARING

General comparative terms

Prototypical example:

verrattuna

A6.1 COMPARING: SIMILAR/DIFFERENT

Comparative terms denoting similarity/difference

Prototypical examples:

aivan_kuten (+), analogia (+), matkia (+), sama (+), simulointi (+), identtinen (+++), erilainen (-), eritä (-), käänteis (-), sen_sijaan (-), vaihtoehto (-)

olla samaa mieltä (+), ottaa mallia (+), olla eroa kuin yöllä ja päivällä (-)

A6.2 COMPARING: USUAL/UNUSUAL

Comparative terms denoting (level) of anomaly

Prototypical examples:

elämäntapa (+), perinnäis (+), tavallinen (+), valtavirta (+), eksentrinen (-), kummallisesti (-), omaperäisyys (-), tavanomaisesta_poikkeava (-)

mahtua viisitoista tusinaan (+), olla tapana (+), ei kasvaa joka oksalla (-)

A6.3 COMPARING: VARIETY

Comparative terms denoting (level of) variety

Prototypical examples:

heterogeeninen (+), kaikenlainen (+), kokooma (+), lajitelma (+), monimuotoisesti (+), valikoima (+), yhtä_ja_toista (+), monotonia (-)

olla valinnan varaa (+), ties mitä (+)

A7 DEFINITE

Abstract terms of modality (possibility, necessity, certainty, etc.)

Prototypical examples:

arvatenkin (+), lienee (+), potentiaalinen (+), selviö (+), tae (+), empiä (-), epätodennäköisesti (-), kai (-), kiistanalainen (-)

kaikin mokomin (+), mennä takuuseen (+), olla vaakalaudalla (-)

A8 SEEM/APPEAR

Abstract terms relating to appearance/impression

Prototypical examples:

ilmenemä, ns., nähtävästi, näköis, oloinen, päällisin_puolin, yleisvaikutelma

näillä näkymin

A9 GETTING, GIVING; POSSESSION

Abstract terms relating to allocating/relinquishing/acquiring/receiving etc.

haalia (+), hallussapito (+), ikioma (+), saatavissa (+), tarjolle (+), hävikki (-), lahjoitus (-), tuliaiset (-)

pitää hyvänään (+), jäädä paitsi (-)

A10 OPEN/CLOSED; HIDING/HIDDEN; FINDING; SHOWING

Abstract terms relating to (level of) openness/concealment/exposure etc.

Prototypical examples:

altistus (+), avonainen (+), julki (+), näyttäytyä (+), päivänselvästi (+), ulkomuoto (+), anonyymi (-), kadota (-), kätkö (-), piilottelu (-), salaa (-)

ihmisten ilmoilla (+), käydä ilmi (+), luurankoja kaapissa (-)

A11 IMPORTANCE

Entries are sub-classified into the following:

A11.1 IMPORTANCE: IMPORTANT

Abstract terms denoting importance/significance

Prototypical examples:

etusija (+), merkittävyys (+), painoarvo (+), kaikki_kaikessa (+++), samantekevä (-), triviaali (-), tyhjänpäiväisesti (-), vähämerkityksisyys (-)

olla sydämen asia (+), kaiken A ja O (+++), ei pitää minään (-), rikka rokassa (-)

A11.2 IMPORTANCE: NOTICEABILITY

Abstract terms denoting noticeability/markedness

Prototypical examples:

jälki (+), kouriintuntuva (+), tehoste (+)

herättää huomiota (+), merkille pantava (+), pitää matalaa profiilia (-)

A12 EASY/DIFFICULT

Abstract terms denoting (level of) difficulty

Prototypical examples:

alkeet (+), havainnollistaa (+), helppo (+), kansantajuinen (+), leikiten (+), haastava (-), kantapään_kautta (-), ongelmallinen (-), pulma (-), työläs (-)

helppo nakki (+), päästä helpolla (+), joutua koville (-), olla pulassa (-)

A13 DEGREE

Prototypical example:

suhteellinen

A13.1 DEGREE: NON-SPECIFIC

Non-specific terms of degree (e.g. intensifiers)

Prototypical examples:

edes, jopa, joskus_jopa, parahiksi

A13.2 DEGREE: MAXIMIZERS

Intensifiers that amplify to the upper extreme

Prototypical examples:

ennen_kaikkea, kaikkein, kertakaikkisen, perin_pohjin, täysin, äärimmäisen

kaikkien aikojen, koko sydämestään, pohjamutia myöten

A13.3 DEGREE: BOOSTERS

Intensifiers that amplify to a high degree (but not the upper extreme)

Prototypical examples:

aikamoinen, entisestään, erittäin, etupäässä, monin_verroin, suuressa_määrin, varsin

kahta kauheammin

A13.4 DEGREE: APPROXIMATORS

Downtowners that express an approximation

Prototypical examples:

keskimäärin, liki, melkein, n., suunnilleen, suurin_piirtein

niillä main

A13.5 DEGREE: COMPROMISERS

Downtowners that express an assumed norm or call into question the appropriacy of X

Prototypical examples:

jokseenkin, jonkinlainen, melko_lailla, suhteellisen

ei enempää eikä vähempää, niin ja näin, siinä ja siinä

A13.6 DEGREE: DIMINISHERS

Downtowners that express only part of the potential force of X or seek to imply that the force of X is limited in some way

Prototypical examples:

himpun_verran, hiukan, osittain

A13.7 DEGREE: MINIMIZERS

Downtowners that imply that the force of X is limited in a maximal way

Prototypical examples:

hädin_tuskin, juuri_ja_juuri

niukin naukin, olla hilkulla

A14 EXCLUSIVIZERS/PARTICULARIZERS

Focusing subjuncts that draw attention to/focus upon X

Prototypical examples:

ainostaan, kyseinen, nimenomaisesti, varsinkin, varta_vasten

A15 SAFETY/DANGER

Abstract terms relating to (the level) of safety/danger

Prototypical examples:

myrkyttömyys (+), riskittömästi (+), turvallinen (+), hengenvaara (-), riski (-), turvaton (-), uhanalainen (-), vaarantua (-)

ehjin nahoin (+), panna vaakalaudalle (-), piru merrassa (-)

B THE BODY & THE INDIVIDUAL

B1 ANATOMY AND PHYSIOLOGY

Terms relating to the (human) body and bodily processes

Prototypical examples:

aineenvaihdunnallinen, aivastelu, DNA, elimistö, hereillä, kaljuuntua, kardio, kuukautiset, nieleskellä, pikkuaivot, välikorva, yöuni, äidinmaito

biologinen kello, olla raskaana

B2 HEALTH AND DISEASE

Terms relating to the (state of the) physical condition

Prototypical examples:

terveydellinen, hyvinvointi (+), virkistäytyä (+), aivoinfarkti (-), allerginen (-), ALS (-), nyrjähtää (-), pökerryksissä (-), sairastaa (-), tulehdus (-), vammais (-)

elämänsä kunnossa (+), terve kuin pukki (+), Alzheimerin tauti (-), antaa ylen (-), hepatiitti B (-)

B3 MEDICINES AND MEDICAL TREATMENT

Terms relating to medication/medical treatment

Prototypical examples:

anestesia, antibiootti, EKG, geriatria, homeopaattinen, kirurgisesti, korvatipat, kuntouttaa, polikliininen, stetoskooppi, tekohampaat, vasektomia, verenkuva

suusta suuhun -menetelmä, pehmeä hoitomuoto

B4 CLEANING AND PERSONAL CARE

Terms relating to domestic/personal hygiene

Prototypical examples:

aurinkovoide, hammasharja, huuhteluaine, hygieeninen, kampaus, kodinhoito, maskara, moppaus, peseytyä, pyykki, taloustyöt, siivoamaton (-)

after shave, käydä kylvyssä

B5 CLOTHES AND PERSONAL BELONGINGS

Terms relating to clothes and other personal belongings

Prototypical examples:

alushousut, asukokonaisuus, hameenhelma, hupullinen, kengät_jalassa, käsine, puvustaa, päärme, sandaali, solmioneula, univormupukuinen, alastomuus (-)

stay up -sukat, vetää housut jalkaan, ilkosen alasti (-), paljain päin (-)

C ARTS & CRAFTS

C1 ARTS AND CRAFTS

Terms relating to artistic/creative activities

Prototypical examples:

etsaus, graafinen, juliste, keramiikka, kudonnais, luolamaalaus, maalipinta, negatiivi, ornamentti, taideteollisesti, taiteellisuus, valokuvaaminen, veistos

art deco, luova toiminta

E EMOTIONAL ACTIONS, STATES, & PROCESSES

E1 EMOTIONAL ACTIONS, STATES, & PROCESSES: GENERAL

General terms depicting emotional actions, states, and processes

Prototypical examples:

emootio, ilmapiiri, intuitiivisesti, subjektiivinen, ilmeettömästi (-), tunteeton (-)

E2 LIKING

Terms depicting fondness/affection/partiality/attachment or the lack of

Prototypical examples:

fanittaa (+), hellyttävä (+), mielellään (+), suosio (+), tervetullut (+), kaikkein_mieluiten (+++), antipatia (-), inhota (-), kaunaisesti (-), paheksunta (-)

olla lähellä sydäntä (+), ihastua ikihyviksi (+++), ei sietää silmissään (-), pitkin hampain (-)

E3 CALM/VIOLENT/ANGRY

Terms depicting (level of) serenity/composure/anger/violence

Prototypical examples:

hillitysti (+), levollisuus (+), malttaa (+), sopu (+), tyyni (+), aggressio (-), ahdistella (-), järjestyshäiriö (-), raivoissaan (-), rettelöidä (-)

ottaa rennosti (+), pitää pää kylmänä (+), nyrkit pystyssä (-), polttaa päreensä (-)

E4 HAPPY/SAD

Entries are sub-classified into the following:

E4.1 HAPPY/SAD: HAPPY

Terms depicting (level of) happiness

Prototypical examples:

euforisesti (+), ilo (+), pelleillä (+), ikionnellinen (+++), epätoivoinen (-), ikävöidä (-), murheellisuus (-), sydäntä_särkevästi (-)

kevättä rinnassa (+), itku kurkussa (-)

E4.2 HAPPY/SAD: CONTENTMENT

Terms depicting (level of) contentment

Prototypical examples:

myhäileminen (+), nautinto (+), tyytyväisesti (+), frustraatio (-), harmillinen (-), pettymys (-), tuskastua (-)

hykerrellä käsiään (+), olla mielissään (+), kurkkua myöten täynnä (-)

E5 FEAR/BRAVERY/SHOCK

Terms relating to (level of) trepidation/courage/surprise etc.

Prototypical examples:

ennakkoluuloton (+), rohjeta (+), sankarillinen (+), seikkailunhaluisesti (+), urotyö (+), uskalias (+), hirvittää (-), kammottava (-), pelokas (-), säikähtää (-)

rohkea rokan syö (+), selkäpiitä karmiva (-), saada sätky (-)

E6 WORRY, CONCERN, CONFIDENT

Terms relating to (level of) apprehension/confidence etc.

Prototypical examples:

huoleti (+), jännityksetön (+), luottavaisuus (+), epäilyttää (-), heikkohermoinen (-), huolissaan (-), stressi (-)

olla kuin Ellun kanat (+), perhosia vatsassa (-), poissa tolalta (-)

F FOOD & FARMING

F1 FOOD

Terms relating to food and food preparation

Prototypical examples:

aamiais, alkuruoka, borssi, feta, hirvenliha, illastaa, kokata, kulinaarinen, kypsentää, maissijauho, ravitsemuksellisesti, ruokinta, vegetaarinen, paasto (-)

chili con carne, köyhät ritarit, seisova pöytä, nähdä nälkää (-)

F2 DRINKS

Terms relating to drinks and drinking

Prototypical examples:

aamuryyppy, appelsiinimehu, cocktail, dekantoida, espresso, piimä, teenjuonti, juopua (++)

musta ryssä, kulauttaa kurkkuunsa, kupponen kuumaa, viinaan menevä (++), olla kuivin suin (-)

F3 CIGARETTES AND DRUGS

Terms relating to cigarettes and (non-medicinal) drugs, including the effects of

Prototypical examples:

heroiini, huumeinen, nikotiini, nuuska, päihde, sikari_suussa, tupakoida

olla pilvessä, panna tupakaksi

F4 FARMING AND HORTICULTURE

Terms relating to agriculture and horticulture

Prototypical examples:

hajakylvö, karjankasvatus, kompostoida, maalais, maataloudellisesti, metsittää, niitto, paimentolais, pienviljelmä, puutarhanhoidollinen, tarhata

ajaa karjaa, olla oraalla

G GOVERNMENT & THE PUBLIC DOMAIN

G1 GOVERNMENT, POLITICS, AND ELECTIONS

Entries are sub-classified into the following:

G1.1 GOVERNMENT ETC.

Terms relating to government and governmental activities

Prototypical examples:

aluehallinto, byrokraattisesti, G7, parlamentarismi, täysistunto, valtiollinen, viranomais

julkinen sektori, tulopoliittinen kokonaisratkaisu

G1.2 POLITICS

Terms relating to politics and political activities

Prototypical examples:

demokratia, esivaali, fasistisesti, glasnost, kansanvaltainen, keskustalais, lobata, poliittisesti, sosialismi, vihr, äänestää, epäpoliittinen (-)

G2 CRIME, LAW, AND ORDER

Entries are sub-classified into the following:

G2.1 CRIME, LAW, AND ORDER: LAW AND ORDER

Terms relating to crime/criminal activities and the legal system

Prototypical examples:

alibi, hovioikeus, KRP, lakisääteisesti, petossyyte, rangaista, ratsata, tutkintavankeus, lainmukainen (+), syytön (+), ilkivaltainen (-), laittomasti (-), pirattikopio (-), salakuljettaa (-)

nostaa kanne, saada ehdollista, harmaa talous (-)

G2.2 GENERAL ETHICS

Terms relating to moral principles/accepted moral practices or the lack of

Prototypical examples:

etiikka, humaani (+), hyve (+), kunniallisesti (+), oikeudenmukainen (+), vilpitön (+), epäreilusti (-), juonitella (-), puolueellinen (-), rietas (-)

pitää sanansa (+), puhdas kuin pulmunen (+), punoa juonia (-), reittä pitkin (-)

G3 WARFARE, DEFENCE, AND THE ARMY; WEAPONS

Terms relating to national security/the armed forces/combat etc.

Prototypical examples:

ammuskella, aseellinen, aseistautua, atomipommi, ilmaisku, maailmansota, militaarisuus, sotilaallisesti, aselepo (-), aseistariisunta (-), rauhanturvaaminen (-)

avata tuli, käydä sotaa, aseeton palvelus (-), haudata sotakirves (-)

H ARCHITECTURE, BUILDINGS, HOUSES, & THE HOME

H1 ARCHITECTURE AND KINDS OF HOUSES AND BUILDINGS

Terms relating to buildings/habitats of various kinds and their construction

Prototypical examples:

arkkitehtoninen, huoneisto, kartano, korjausrakentaminen, LVI, muurata, rakennustaiteellisesti, rakennuttaa, riemukaari, ulkohuone

H2 PARTS OF BUILDINGS

Terms relating to parts of buildings

Prototypical examples:

auditorio, aula, hissikuilu, ikkunalauta, ikkunallinen, julkisivu, kattotiili, kuisti, kupoli, makuusali, odotushuone, ovi, parveke, sviitti, vesikatto, vessa, yläkerta

H3 AREAS AROUND OR NEAR HOUSES

Terms relating to areas around or near houses/buildings

Prototypical examples:

aukio, haja-asutusalue, kadunkulma, kortteli, kävelytie, patio, pihakiveys, puistoalue, puutarha, suojatie, takapiha

H4 RESIDENCE

Terms relating to habitation/occupancy or the lack of

Prototypical examples:

asua, asuttaminen, henkikirjoituspaikka, huusholli, majoitus, oleskeleminen, yöpyä, asunnottomuus (-), koditon (-)

asettua taloksi, katto pään päälle

H5 FURNITURE AND HOUSEHOLD FITTINGS

Terms relating to furniture and fittings used within the home/buildings

Prototypical examples:

allaskaappi, divaani, höyhentyyny, istumapaikka, kalustettu, lauteet, liinavaatteet, reunuslista, rullaverho, tapetoida, uunikinnas, kalustamaton (-)

I MONEY & COMMERCE

I1 MONEY GENERALLY

Terms relating to money generally

Prototypical examples:

annuiteetti, budjetoida, dollari, ecu, euro, FIM, käteis, lantti, monetaarinen, pankkitoiminta, rahapoliittisesti, rahoittaa, sekki

I1.1 MONEY: AFFLUENCE

Terms relating to (level of) wealth/prosperity

Prototypical examples:

elanto, investoida, hyväpalkkaisuus (+), liikevoitto (+), vakavarainen (+), vauraasti (+), kerjäläis (-), maksukyvyttömyys (-), pienituloinen (-), tyhjin_käsin (-), rutiköyhä (---)

elää herroiksi (+), lyödä leiville (+), elää kädestä suuhun (-), matti kukkarossa (-)

I1.2 MONEY: DEBTS

Terms relating to (level of) debt

Prototypical examples:

erääntyä, muistutuslasku, rästi, tilinylitys, vekseli, velallis, konkurssi (+), vararikkoinen (+)

höylätä muovirahaa, tehdä velkaa

I1.3 MONEY: PRICE

Terms relating to cost/worth/value (includes invoicing procedures)

Prototypical examples:

hinnoitella, hintainen, kustannus, lunnas, postimaksu, rahastaa, tariffi, hintava (+), kalliisti (+), alehinta (-), ilmaiseksi (-), taloudellisuus (-)

käydä kalliiksi (+), maksaa maltaita (+)

I2 BUSINESS

Entries are sub-classified into the following:

I2.1 BUSINESS: GENERALLY

Terms relating to business generally

Prototypical examples:

alihankinta, franchising, kaupallisesti, liike-elämä, liiketoiminnallinen, maailmankauppa, pörssi, taloudellis, yrittäjyys, ei-kaupallinen (-)

käydä kauppaa

I2.2 BUSINESS: SELLING

Terms relating to trading/retail

Prototypical examples:

huutokauppa, mainonta, myydä, myyjäis, osto, vuokrata, vähittäiskauppa, myymätön (-)

olla kaupan

I3 WORK AND EMPLOYMENT

Entries are sub-classified into the following:

I3.1 WORK AND EMPLOYMENT: GENERALLY

Terms relating to work and employment generally

Prototypical examples:

ammatillisesti, ansiotyö, elinkeino, palkata, työllistyä, työllisyys, työperäinen, virkaura, lakkolais (-), pitkäaikaistyöttömyys (-)

harjoittaa ammattia, käydä työssä, saada potkut (-), sanoutua irti (-)

I3.2 WORK AND EMPLOYMENT: PROFESSIONALISM

Terms relating to (level of) professionalism

Prototypical examples:

ammattikokemus (+), ammattitaitoisesti (+), liikemiesmäisyys (+), amatöörimäisesti (-), ammattitaidoton (-), epäpätevyys (-)

I4 INDUSTRY

Terms relating to industry (including types of)

Prototypical examples:

koneenrakennus, puunjalostus, rikastamo, tehdasmainen, tekstiiliteollisuus, teollisesti

K ENTERTAINMENT, SPORTS, & GAMES

K1 ENTERTAINMENT GENERALLY

Terms relating to entertainment generally

Prototypical examples:

cancan, harraste, karnevaali, lomailla, naamiaiset, partiolais, penkkiurheilu, sirkusnäytäntö, tanhu, tanssittaa, telttailu, ulkoilla, yöelämä

after ski, break dance, panna jalalla koreasti

K2 MUSIC AND RELATED ACTIVITIES

Terms relating to music and related activities

Prototypical examples:

aaria, alttoviulu, diskomusiikki, juomalaulu, kuudestoistaosanuotti, musiikillinen, musikaalisesti, sormiharjoitus, svengata, tahtipuikko, urut, ääniala

avata ääni, kevyt musiikki

K3 RECORDED SOUND ETC.

Terms relating to recorded sound / sound recording

Prototypical examples:

hifi, LP, magnetofoni, MP3, samplaaminen, stereofonisesti, äänittää

K4 DRAMA, THE THEATRE, AND SHOW-BUSINESS

Terms relating to drama, the performing arts, etc.

Prototypical examples:

baletti, dramaturgia, kuvaelma, kesäteatteri, koreografinen, lausunta, pantomiimi, teatteritaide, varietee

drag show, stand up

K5 SPORTS AND GAMES GENERALLY

Entries are sub-classified into the following:

K5.1 SPORTS

Terms relating to sporting activities

Prototypical examples:

aerobic, aitoa, alppiyhdistetty, ankkuriosuus, F1, hiekkaeste, hirvenhiihto, hölkkä, joogata, jäärata-ajo, keilailu, sauvakävely, yösuunnistus

Cooperin testi, vetää leukoja

K5.2 GAMES

Terms relating to games and other leisure activities

Prototypical examples:

biljardi, bingo, kädenvääntö, minigolf, pokeri, saappaanheitto, selviytymispeli

heittää tikkaa, kaataa valtilla

K6 CHILDREN'S GAMES AND TOYS

Terms relating to children's games and toys

Prototypical examples:

domino, hippa, keinu, kumiankka, mollamaija, puujalat, tinasotilas, tutti

hypätä narua

L LIFE & LIVING THINGS

L1 LIFE AND LIVING THINGS

Terms relating to life and death

Prototypical examples:

biologinen, luonnonvaraisuus, henkiinjääminen (+), edesmennyt (-), hengenlähtö (-), hirttäytyä (-), kuoliaaksi (-), kuolleisuus (-), menehtyä (-), sukupuutto (-)

elävien kirjoissa (+), selvitä hengissä (+), joutua hirteen (-), saada surmansa (-)

L2 LIVING CREATURES GENERALLY

Terms relating to living creatures (e.g. non-human)

Prototypical examples:

ahma, alkueläin, dalmatiankoira, dinosaurus, evä, kissanpentu, koiransukuinen, päätäi, soidin, emakko (f), naarasleijona (f), koiraspuolinen (m), sonni (m)

turkkilainen angora, karkeakarvainen mäyräkoira

L3 PLANTS

Terms relating to plants and plant-life

Prototypical examples:

ahkeraliisa, ainavihanta, apilankukka, emi, floora, hieskoivu, juurikas, kasvitieteellisesti, kukkais, köynnösruusu, levä, sappitatti

jalo lehtipuu

M MOVEMENT, LOCATION, TRAVEL, & TRANSPORT

M1 MOVING, COMING AND GOING

Terms depicting movement (towards and away from X)

Prototypical examples:

eräretki, hortoilu, jaloitteleminen, kuljeksia, kuperkeikka, liirto, matkustaa, motoriikka, pakolais, räpäyttää, siirtolais

käydä pitkäkseen, ottaa jalat alleen

M2 PUTTING, TAKING, PULLING, PUSHING, TRANSPORTING, ETC.

Terms depicting putting/taking/pulling/pushing movements/activities

Prototypical examples:

evakuoiminen, huolita, kuljetus, kärrätä, lastaaminen, matkarahti, siirtely, vinssata, viskellä

M3 MOVEMENT/TRANSPORTATION: LAND

Terms depicting means of transport / ways of transporting and/or travelling on land

Prototypical examples:

ahkio, ajoneuvo, autoilla, jalkaisin, joukkoliikenne, kuorma-auto, kyyditys, maanteitse, moottorikelkka, pikajuna, polkupyörä, potkulauta, taksikyyti, traktori

kevyt liikenne, olla ratissa

M4 MOVEMENT/TRANSPORTATION: WATER

Terms depicting means of transport / ways of transporting and/or travelling by water

Prototypical examples:

höyrylaiva, kirkkovene, krooli, melominen, merenkulku, polskuttaa, ponttonilautta, uitto, veneillä, vesiteitse

heittää talviturkki, lähteä merille, uida koiraa

M5 MOVEMENT/TRANSPORTATION: AIR

Terms depicting means of transport / ways of transporting and/or travelling by air

Prototypical examples:

avaruussukkula, helikopteri, ilmailu, ilmateitse, laskuvarjo, rahtikone, vesitaso

M6 LOCATION AND DIRECTION

Terms depicting position of/point of reference for X

Prototypical examples:

alapuolinen, eteenpäin, Etelä-, itä, itäinen, luona, navigoida, paikan_päällä, pituussuuntaan, sijainti, tuolta_puolen, ulompi, ulos, ääri

näillä tienoin, rapakon takana, ristiin rastiin

M7 PLACES

Terms depicting geographical/conceptual spaces

Prototypical examples:

alueellinen, kansainvälisesti, kaupunkilais, kirkonkylä, kotimainen, kunnallinen, länsivallat, mantere, osavaltio, rantatontti, reviiri, ydinkeskusta

kolmas maailma, tuhansien järvien maa

M8 REMAINING/STATIONARY

Terms depicting the various stages of inactivity (stopping/loitering/immobility etc.)

Prototypical examples:

kelliä, liikkumaton, makoilu, seisoskella

niille jalansijoilleen, olla paikallaan

N NUMBERS & MEASUREMENT

N1 NUMBERS

Number terms (e.g. cardinal, ordinal, fraction, etc.)

Prototypical examples:

desimaaliluku, kahdeksasosa, kaksitoista, kolmannes, kymmenluku, nro, nolla, numeroida, parisataa, puolitoista, triljoona

leipurin tusina, pitää lukua

N2 MATHEMATICS

Mathematical terms

Prototypical examples:

derivaatta, geometrinen, hajonta, jaollinen, laskeskella, laskuoppi, matemaattisesti, puolittua, tilastollisesti, vähennyslasku

jaoton luku, laskea päässään

N3 MEASUREMENT

Prototypical example:

jakauma

N3.1 MEASUREMENT: GENERAL

General measurements

Prototypical examples:

hehto, karaatti, kvantitatiivisesti, mitoittaa, oktaaniluku, perusmitta

N3.2 MEASUREMENT: SIZE

Terms of measurement relating to size

Prototypical examples:

kokonumero, kokoinen, kookas (+), kukkura (+), reilunpuoleinen (+), jätti (+++), pien (-), minikokoinen (---), pikkuriikkinen (---)

häviävän pieni (---)

N3.3 MEASUREMENT: DISTANCE

Terms of measurement relating to distance

Prototypical examples:

ylettyä, etäällä (+), kauko (+), syrjäinen (+), kauimpana (+++), kintereillä (-), lähelle (-), ulottuvilla (-), vieri_vieressä (-), lähemmäksi (--)

niin kauas kuin silmä kantaa (+), kivenheiton päässä (-), käden ulottuvilla (-)

N3.4 MEASUREMENT: VOLUME

Terms of measurement relating to volume

Prototypical examples:

litrainen, paksuinen, tilavuus, pullea (+), pyylevyys (+)

N3.5 MEASUREMENT: WEIGHT

Terms of measurement relating to weight

Prototypical examples:

kiloinen, punnitus, taara, tyhjäpaino, alipainoinen (-)

kuollutta painoa

N3.6 MEASUREMENT: AREA

Terms of measurement relating to area

Prototypical examples:

pinta-ala, laajuinen (+)

N3.7 MEASUREMENT: LENGTH AND HEIGHT

Terms of measurement relating to length and height

Prototypical examples:

korkuinen, levyinen, syvyinen, ympärysmitta, leveähkö (+), pitkänhuiskea (+), lyhyenläntä (-), lyhytkasvuinen (-)

N3.8 MEASUREMENT: SPEED

Terms of measurement relating to speed

Prototypical examples:

tempo, tuntinopeus, kiriä (+), niin_pian_kuin_mahdollista (+), pika (+), ajan_mittaan (-), hidastella (-), kiireetön (-), vähitellen (-)

alta aikayksikön (+), pitää kiirettä (+), päivä kerrallaan (-), vähin erin (-)

N4 LINEAR ORDER

Terms relating to linear movement/order, sequencing, etc.

Prototypical examples:

aluksi, alustava, edellis, ennakkoon, ex, hännänhuippu, jatkumo, kukin_vuorollaan, kuudeskymmenes, peräkkäisyys, seuraavaksi, sittemmin

ensi alkuun, loppujen lopuksi, vuoron perään

N5 QUANTITIES

Terms depicting quantities

Prototypical examples:

annostus, henkeä_koti, kahvikupillinen, kpl, prosenttisesti, moni (+), enemmistö (+++), jokunen (-), jonkin_verran (-), ripaus (-), ainoa (---)

per nenä, koko joukko (+), käydä vähiin (-), ainoa lajiaan (---)

N5.1 ENTIRETY; MAXIMUM

Terms depicting maximal/maximum quantities

Prototypical examples:

enimmäis (+), maksimi (+), täysi (+), alaraja (-), kaistale (-), minimoida (-), osittainen (-), puolikas (-)

hela hoito (+), olla tupaten täynnä (+)

N5.2 EXCEEDING; WASTE

Terms depicting excessive/wasteful quantities

Prototypical examples:

haaskata (+), jäljellä (+), jäämistö (+), kohtuuton (+), liika (+), suma (+), yliampuva (+)

heittää hukkaan (+), mennä harakoille (+)

N6 FREQUENCY ETC.

Terms relating to frequency/rate of recurrence

Prototypical examples:

frekvenssi, kuukausittainen, toisinaan, monesti (+), toistua (+), tuon_tuostakin (+), alituinen (+++), aika_ajoin (-), harvoin (-), satunnainen (-)

harva se päivä (+), moneen otteeseen (+)

O SUBSTANCES, MATERIALS, OBJECTS, & EQUIPMENT

O1 SUBSTANCES AND MATERIALS GENERALLY

Terms relating to substances and materials generally

Prototypical examples:

atomi, C-vitamiini, jodi, kemiallinen, kivennäis, kofeiini, lisäaine, raaka-aine, suolahappo, yhdiste

O1.1 SUBSTANCES AND MATERIALS GENERALLY: SOLID

Terms depicting solid substances/materials

Prototypical examples:

akryyli, duffelikangas, fajanssi, graniitti, ihra, jääpuikko, keraaminen, kitti, kupari, noki, teräs

O1.2 SUBSTANCES AND MATERIALS GENERALLY: LIQUID

Terms depicting liquid substances (other than drinks) and wetness

Prototypical examples:

diesel, etikkaliemi, juomavesi, laimennusaine, liottaminen, nestemäinen, pellavaöljy, kuivattu (-)

O1.3 SUBSTANCES AND MATERIALS GENERALLY: GAS

Terms depicting / relating to gases

Prototypical examples:

aerosoli, happi, hengitysilma, hiilimonoksidi, häkä, ilokaasu, otsoni, propaani, savukaasu, vesihöyry

O2 OBJECTS GENERALLY

Terms relating to objects generally

Prototypical examples:

ampulli, dieselmoottori, esineistö, höylä, kaakeli, kannu, kulutushyödyke, plo, säiliö, vesijohto

kimpsut ja kampsut

O3 ELECTRICITY AND ELECTRICAL EQUIPMENT

Terms relating to electricity and electrical equipment

Prototypical examples:

antenni, anturi, elektroniikka, kaikuluotain, muuntaja, pistorasia, reaktori, sähkökäyttöinen, tasavirta

kytkeä virta

O4 PHYSICAL ATTRIBUTES

Entries are sub-classified into the following:

O4.1 GENERAL APPEARANCE AND PHYSICAL PROPERTIES

Terms relating to general appearance / physical properties

Prototypical examples:

alava, helmeillä, kotikutoinen, layout, rakenteellisesti

O4.2 JUDGEMENT OF APPEARANCE

Descriptive terms relating to the appearance/look of X

Prototypical examples:

eleganssi (+), komea (+), luksus (+), sopusuhtainen (+), viehättävä (+), hyrskyn_myrskyn (-), nuhruisuus (-), poissa_muodista (-), sekasotku (-)

miellyttää silmää (+), hävityksen kauhistus (-), olla kuin pommin jäljiltä (-)

O4.3 COLOUR AND COLOUR PATTERNS

Terms depicting colour and other visual attributes

Prototypical examples:

harmaa, kellertää, kuviollinen, monivärisyys, raidoitus, sinivalkoinen, värjätä

O4.4 SHAPE

Terms relating to shapes / the shape of X

Prototypical examples:

kaareutua, kupera, köyristää, loiva, muotoinen, soikeus, suorakulmainen

O4.5 TEXTURE

Terms depicting texture

Prototypical examples:

huokoinen, jauhemaisuus, karvainen, kimmoisa, murea, pehmoinen, tahmeus

O4.6 TEMPERATURE

Terms relating to temperature

Prototypical examples:

ilmastoida, lämpötila, terminen, avotuli (+), esilämmitys (+), kiehauttaa (+), metsäpalo (+), paahde (+), jäässä (-), pakastaa (-), paleltuminen (-), viluinen (-)

absoluuttinen nollapiste, panna tulelle (+), olla kananlihalla (-)

P EDUCATION

P1 EDUCATION IN GENERAL

Terms relating to education in general

Prototypical examples:

aikuiskoulutus, akatemia, apulaisprofessuuri, esiopetus, koululais, opiskella, oppimäärä, pedagogisesti, tieteidenvälinen, yliopistollinen, yo, oppimaton (-)

avoin yliopisto, cum laude

Q LINGUISTIC ACTIONS, STATES, & PROCESSES

Q1 COMMUNICATION

Entries are sub-classified into the following:

Q1.1 COMMUNICATION IN GENERAL

Terms relating to communication in general

Prototypical examples:

aineisto, ekspressiivinen, gallup, ilmaista, info, kaavio, kommunikointi, propaganda, sanoma, slogan, symbolisesti, tiedote, viestittää, yhteydenotto

tehdä selkoa

Q1.2 PAPER DOCUMENTS AND WRITING

Terms relating to written communication (including writing/printing implements and documentation)

Prototypical examples:

adressi, allekirjoitus, asiakirja, fontti, hieroglyfi, kirjeitse, kuponki, kutsukortti, muistio, pöytäkirja, sivuinen, typografia, kirjoittamaton (-)

kirjata ylös, panna nimensä alle

Q1.3 TELECOMMUNICATIONS

Terms relating to telecommunications

Prototypical examples:

faksata, puhelimitse, puhelinkeskus, tekstiviesti, tukiasema, älypuhelin

kuuma linja, saada langan päähän

Q2 SPEECH ACTS

Entries are sub-classified into the following:

Q2.1 SPEECH ETC.: COMMUNICATIVE

Terms relating to spoken communication

Prototypical examples:

haastattelu, juoruilu, keskustella, konsultoida, kysely, neuvottelu, palaute, vuoropuhelu, tuppisuu (-)

ajatuksien vaihtaminen, puhua ympäri

Q2.2 SPEECH ACTS

Speech acts terms

Prototypical examples:

itseilmaisu, jokeltaa, kirkaista, retorinen, saaga, suullisesti, sanaton (-),

ajatella ääneen, suun soittaminen, leikkiä mykkäkoulua (-)

Q3 LANGUAGE, SPEECH, AND GRAMMAR

Terms relating to language (including linguistic/grammatical terms)

Prototypical examples:

aakkos, adjektiivi, erikoiskieli, eufemismi, huutomerkki, kansanruno, kirjain, lyyrisesti, monikollinen, puhekielisesti, riimittää, semantiikka, slangi, suomentaa

epäsuora sanajärjestys, lingua franca

Q4 THE MEDIA

General terms relating to the media

Prototypical examples:

joukkoviestin, julkaista, kustannustoiminta, media, painos, tiedotusväline

julkinen sana

Q4.1 THE MEDIA: BOOKS

Media terms relating to (types of) books and their production

Prototypical examples:

bibliografia, esipuhe, hakuteos, kaunokirjallisuus, kirjapainotaito, lukemisto, manga, opaskirja, painotuote, pokkari

kirjallinen lähde

Q4.2 THE MEDIA: NEWSPAPERS ETC.

Media terms relating to (types of) newspapers and their production

Prototypical examples:

aikakauslehti, iltapäivälehti, journalistinen, pääkirjoitus, rubriikki, tabloidi

keltainen lehdistö

Q4.3 THE MEDIA: TV, RADIO, AND CINEMA

Media terms specifically relating to TV, radio, and the cinema

Prototypical examples:

animaatio, asiaohjelma, filmatisoida, kotivideo, kuunnelma, kuvaruutu, mykkäfilmi, radioida, saippuaooppera, TV, valkokangas

film noir, lyhyet aallot

S SOCIAL ACTIONS, STATES, & PROCESSES

Entries are sub-classified into the following:

S1.1.1 SOCIAL ACTIONS, STATES, AND PROCESSES: GENERAL

Terms relating to social actions, state, and processes in general

Prototypical examples:

emännöidä, ihmistenvälinen, kohdella, kädenpuristus, käytös, seremoniallinen, sivilisaatio, sosiologisesti, visiitti

istua iltaa, pitää yhteyttä

S1.1.2 RECIPROCITY

Terms relating to the exchange of X / lack of exchange between (groups of) people

Prototypical examples:

interaktiivisesti (+), keskinäinen (+), molemminpuolisuus (+), toinen_toisestaan (+), uskollinen (+), epälojaali (-)

kristillinen tasajako (+), olla sujut (+), saada lämmintä kättä (-)

S1.1.3 PARTICIPATION

Terms relating to participation/involvement or the lack of

Prototypical examples:

huippukokous (+), illanistujaiset (+), kokoontua (+), osallistua (+), sessio (+), symposium (+), väliintulo (+), boikotointi (-)

lähteä leikkiin mukaan (+), loistaa poissaolollaan (-), ottaa etäisyyttä (-)

S1.1.4 DESERVE ETC.

Terms relating to entitlement/eligibility/merit etc.

Prototypical examples:

ansaitusti (+), perusoikeus (+)

S1.2 PERSONALITY TRAITS

Terms depicting personality traits/characteristics

Prototypical examples:

maneeri, oikullisuus, temperamentti

S1.2.1 APPROACHABILITY AND FRIENDLINESS

Terms depicting (level of) approachability/friendliness

Prototypical examples:

ekstrovertti (+), ihmisrakas (+), luontevasti (+), vieraanvaraisuus (+), epäsosiaalisuus (-), pahansuopa (-), tylysti (-)

hieroa sovintoa (+), ottaa avosylin vastaan (+), kuin seipään niellyt (-), sydän kivestä (-)

S1.2.2 AVARICE

Terms depicting (level of) avarice/generosity

Prototypical examples:

ahne (+), itsekkäästi (+), kateus (+), mustasukkainen (+), avokätisesti (-), epäitsekkyys (-), pyyteetön (-), ystävänpalvelus (-)

ajaa omaa etuaan (+), dollarinkuvat silmissä (+), hellittää kukkaron nyörejä (-)

S1.2.3 EGOISM

Terms depicting (level of) egoism

Prototypical examples:

itsetunto, minuus, itsetietoinen (+), leuhkasti (+), mahtailla (+), turhamainen (+), ylvästely (+), kainosti (-), ujostella (-)

ajaa omaa etuaan (+), olla olevinaan (+), häntä koipien välissä (-), niellä ylpeytensä (-)

S1.2.4 POLITENESS

Terms depicting (level of) politeness

Prototypical examples:

käyttäytymissääntö, anteeksipyyntö (+), kiitos (+), kohtelias (+), hävytön (-), kärkevä (-), rienata (-), sarkastinen (-), töykeys (-)

antaa anteeksi (+), olla paha suustaan (-)

S1.2.5 TOUGHNESS: STRONG/WEAK

Terms depicting (level of) strength/weakness

Prototypical examples:

sisukkaasti (+), voimakastahtoisuus (+), haavoittuva (-), heiveröinen (-), vastustuskyvyttömyys (-)

olla kanttia (+), kuin kala kuivalla maalla (-)

S1.2.6 SENSIBLE

Terms depicting (level of) sensibleness/absurdity

Prototypical examples:

johdonmukainen (+), järkevästi (+), kohtuus (+), täyspäinen (+), vastuuntunto (+), absurdi (-), edesvastuuton (-), naiivius (-), typeryys (-), törttöillä (-)

kohtuus kaikessa (+), olla maalaisjärkeä (+), olla pää pilvissä (-)

S2 PEOPLE

Terms indicating that particular words relate to / denote people

Prototypical examples:

henkilöllisyys, ihmiskunta, kansoittaa, lähimmäinen, sukupuoli, yksilö, yksityishenkilö

Homo sapiens

S2.1 PEOPLE: FEMALE

Terms relating to females

Prototypical examples:

naisellinen, rouvas, tyttömäisesti, epänaisellinen (-), kimma (f), matami (f), neiti (f), poikamiestyttö (f), rouvashenkilö (f)

kauniimpi sukupuoli

S2.2 PEOPLE: MALE

Terms relating to males

Prototypical examples:

miehekkyys, poikamainen, epämiehekkäästi (-), jätkä (m), kundi (m), Mr (m), ukkeli (m), äijä (m)

S3 RELATIONSHIP

Entries are sub-classified into the following:

S3.1 RELATIONSHIP: GENERAL

Terms relating to relationships in general

Prototypical examples:

ihmissuhde, kumppanuus, ystävyyssuhde

hieroa tuttavuutta, pitää seuraa, panna välit poikki (-)

S3.2 RELATIONSHIP: INTIMATE/SEXUAL

Terms relating to relationships that are intimate and/or sexual or to a person's sexual orientation.

Prototypical examples:

eroottis, esileikki, flirttailu, halaileminen, heteroseksuaalisuus, lempi, masturboida, sadomasokismi, seksuaalisesti, suukko, treffit, selibaatti (-)

käydä naisissa, lihan himot

S4 KIN

Terms relating to relationships between family members / familiars

Prototypical examples:

adoptio, avioliitto, isyys, kosia, lapsenlapsi, leski, lähiomainen, naittaa, perikunta, yksinhuoltaja, äidillisesti, avioero (-), naimaton (-), orpo (-), anoppi (f), isoäiti (f), kummityttö (f), eno (m), isäpuoli (m), sulhanen (m), veljekset (m)

mennä kihloihin, saada rukkaset (-)

S5 GROUPS AND AFFILIATION

Terms relating to groups / the level of association / affiliation between groups

Prototypical examples:

fuusio (+), heimo (+), kollektiivisesti (+), lauma (+), liittolais (+), yhdistys (+), yhteinen (+), itsenäisyys (-), omin_voimin (-), riippumaton (-), yksinään (-)

kuulua yhteen (+), yksissä neuvoin (+), omin päin (-), ylhäinen yksinäisyys (-)

S6 OBLIGATION AND NECESSITY

Terms depicting (level of) obligation/necessity

Prototypical examples:

ehto (+), pakko (+), sitovasti (+), tarpeen_tullen (+), tarvita (+), velvoittaa (+), harkinnanvaraisesti (-), omavalintainen (-), tarpeettomuus (-), vapaaehtois (-)

kantaa vastuu (+), ota tai jätä (+)

S7 POWER RELATIONSHIP

Entries are sub-classified into the following:

S7.1 POWER, ORGANIZING

Terms depicting power/authority/influence and organisation/administration

Prototypical examples:

hierarkia, status, arvovaltainen (+), diktatuuri (+), niskan_päällä (+), organisoida (+), päätösvalta (+), alaisuus (-), epäitsenäinen (-), kukistua (-), nöyrästi (-), totella (-)

alistaa valtaan (+), panna päiväjärjestykseen (+), olla tossun alla (-)

S7.2 RESPECT

Terms depicting (level of) respect/deference/reverence

Prototypical examples:

arvostaa (+), ihaileva (+), kunnioitettava (+), perinteikkäästi (+), häpäistä (-), ivallinen (-), nöyryytys (-), parodiointi (-), sarkastisesti (-)

antaa arvoa (+), kunnioittaa hetken hiljaisuudella (+), katsoa kuin halpaa makkaraa (-), pitää pilkkanaan (-)

S7.3 COMPETITION

Terms depicting competition/rivalry/contest or the lack of

Prototypical examples:

alkuerä, arvokisa, kaksintaistelu, lopputurnaus, semifinaali, kilpailla (+), kilpailuhenkinen (+), kilvoittelu (+)

mitellä voimiaan (+)

S7.4 PERMISSION

Terms depicting (level of) permission/consent/authorisation

Prototypical examples:

erioikeus (+), myöntävästi (+), oikeutettu (+), sallia (+), suostumus (+), valtuuttaa (+), evätä (-), kielto (-), suvaitsematon (-), tabu (-)

antaa hyväksymys (+), voimassa oleva (+), julistaa pannaan (-), omin luvin (-)

S8 HELPING/HINDERING

Terms depicting (level of) help/hindrance

Prototypical examples:

avulias (+), edesauttaa (+), opastus (+), taustatuki (+), yhteistyöhaluisesti (+), ennaltaehkäisy (-), estää (-), tottelematon (-), vastarinta (-)

pelastava enkeli (+), pitää huolta (+), jättää omilleen (-), panna kapuloita rattaisiin (-)

S9 RELIGION AND THE SUPERNATURAL

Terms relating to religions and the supernatural

Prototypical examples:

astrologia, evankelinen, keijukais, körttiläis, magia, manala, paavius, pyhittää, ristiäiset, sielunvaellus, taikauskoisesti, tarot, ufo, epäpyhä (-)

bar mitsva, Isä meidän -rukous, musta magia, new age, päästä taivaaseen

T TIME

T1 TIME

Prototypical examples:

ajankäyttö, ennen_pitkää, kellonaika, klo, koskaan, päivämäärä, tasatunti

Greenwichin aika, jälkeen Kristuksen

T1.1 TIME: GENERAL

Terms relating to time in general

Entries are sub-classified into the following:

T1.1.1 TIME: GENERAL: PAST

General terms relating to a past (period/point in) time

Prototypical examples:

aika_päiviä_sitten, arkeologinen, edellisvuotinen, eilen, historiallisesti, menneisyys, retro, takavuosi, tertiäärikausi, tähänastinen, viikinkiaika

ennen muinoin, historian siipien havina

T1.1.2 TIME: GENERAL: PRESENT; SIMULTANEOUS

General terms relating to a present (period/point in) time

Prototypical examples:

ajankohtainen, ajantasaisuus, juuri_nyt, kerralla, nyky, paraikaa, samaan_aikaan, toistaiseksi, vireillä, eri_aikaan (-)

aikaansa seuraava, ajan henki, näillä minuuteilla

T1.1.3 TIME: GENERAL: FUTURE

General terms relating to a future (period/point in) time

Prototypical examples:

futuristinen, huomenna, loppuelämä, lähitulevaisuus, tästä_lähin

hamassa tulevaisuudessa, joku kaunis päivä

T1.2 TIME: MOMENTARY

Terms relating to a momentary/transitory (period/point in) time

Prototypical examples:

ajoittua, hetki, niihin_aikoihin, nukkumaanmenoaika, päivätä, tuolloin

puolilta öin, sillä erää

T1.3 TIME: PERIOD

Terms relating to a specific period of time

Prototypical examples:

ajanjakso, arki-ilta, kalenterikuukausi, kesä, kesäkuinen, ma, väliaikaisesti, öisin, aamusta_iltaan (+), kauan (+), monivuotinen (+), pidemmäksi_aikaa (++), kauimmin (+++), hetkellisyys (-), lyhytaikainen (-), ohimenevästi (-)

viikon sisällä, hyvän aikaa (+), iän kaiken (+++), lyhyt aikaväli (-)

T2 TIME: BEGINNING AND ENDING

Terms depicting commencement/completion

Prototypical examples:

alkaa (+), lähtökohta (+), viritä (+), keskeneräinen (++), meneillään (++), pysyvä (+++), vakinaisuus (+++), ehtyä (-), jäähyväis (-), katkos (-), keskeytyä (-), päättyminen (-), raueta (-), valmiiksi (-)

panna vireille (+), sanoista tekoihin (+), jättää kesken (-)

T3 TIME: OLD, NEW, AND YOUNG; AGE

Terms relating to age/maturity

Prototypical examples:

ikä, kahdeksankuinen, kolmikymppinen, aikuinen (+), eläkeläis (+), täysi-ikäisyys (+), ikääntyä (++), ikivanha (+++), alaikäinen (-), innovaatio (-), lapsuus (-), moderni (-), uudis (-), alkuperäinen (---), upouusi (---)

kypsä ikä (+), elämän ilta (+), vanha kuin taivas (+++), aikaansa edellä (-), olla lapsenkengissä (-)

T4 TIME: EARLY/LATE

Terms relating to well-timedness

Prototypical examples:

ennenaikaisesti (+), ennättää (+), hyvissä ajoin (+), aikaisintaan (+++), iltamyöhä (-), myöhästellä (-), viive (-), myöhemmin (--), viimeistään (---)

kukonlaulun aikaan (+), olla etuajassa (+), viime tipassa (-)

W THE WORLD & OUR ENVIRONMENT

W1 THE UNIVERSE

Terms relating to the universe/cosmos

Prototypical examples:

auringonpimennys, ilmakehä, puolikuu, revontuli, taivaankappale, universumi

musta aukko, punainen planeetta

W2 LIGHT

Terms related to light

Prototypical examples:

aamurusko, hohtaa, kiiltävä, kohdevalo, kynttilänvalo, päivänvalo, varjoisa (-)

pilkkosen pimeä (-)

W3 GEOGRAPHICAL TERMS

Geographical terms

Prototypical examples:

aallokko, aarniometsä, atolli, hiekkaranta, hiidenkirnu, joki, jäälautta, kelottua, maantieteellisesti, makeavetinen, metsäinen, penkere, vulkaaninen, öljylähde

ahventen valtakunta, luonnon helma

W4 WEATHER

Terms relating to the climate / weather conditions

Prototypical examples:

helleaalto, ikirouta, ilmastollisesti, kaatosade, keli, kumpupilvi, meteorologia, salamoida, sääolot, tsunami, ukkos, usvainen

sataa kaatamalla

W5 GREEN ISSUES

Environmental terms

Prototypical examples:

eko, ekologinen, kasvihuoneilmiö, luomu, luonnonsuojelu, ongelmajäte, saastua, saastuttaa, ympäristövaikutus, öljylautta

X PSYCHOLOGICAL ACTIONS, STATES, & PROCESSES

X1 PSYCHOLOGICAL ACTIONS, STATES, AND PROCESSES: GENERAL

General terms relating to psychological actions, states, and processes

Prototypical examples:

alitajunta, ego, hengenlaatu, mentaliteetti, mieliala, psyykkinen

X2 MENTAL ACTIONS AND PROCESSES

Terms relating to mental actions and processes in general

Prototypical examples:

hypnotisoida, transsi, unelmoida

X2.1 THOUGHT, BELIEF

Terms relating to reasoning/thinking and level of belief/scepticism

Prototypical examples:

ajatelma, filosofisesti, harkinta, ideoida, järkeillä, kognitiivinen, luulo, mielipide, mietiskellä, otaksua, epäluuloinen (-), skeptisesti (-)

hautoa mielessään, saada päähänpisto

X2.2 KNOWLEDGE

Terms relating to (level) of knowledge/perception/retrospection

Prototypical examples:

maine (+), mieleenpainuvasti (+), tietoinen (+), tietotaito (+), ikimuistoinen (+++), epätietoisesti (-), perehtymätön (-), tuntemattomuus (-), unohdus (-)

tuntea kuin omat taskunsa (+), ei olla aavistustakaan (-)

X2.3 LEARN

Terms relating to (level of) learning/mastery/deduction/realisation

Prototypical examples:

oppia (+), opetteleminen (+), sisäistää (+), sivistyä (+), kovapäinen (-), oppimiskyvytön (-), sivistymättömästi (-)

ottaa opikseen (+), päästä selvyyteen (+)

X2.4 INVESTIGATE, EXAMINE, TEST, SEARCH

Terms relating to investigation/examination

Prototypical examples:

analysoida, etsiskely, heuristiikka, jäljittäminen, katsastaa, kyselytutkimus, monitoroida, selata, suuretsintä, testata, tulikoe, tutkimustyö

etsiä käsiinsä, päästä jäljille, tunnustella kepillä jäätä

X2.5 UNDERSTAND

Terms depicting (level of) understanding/comprehension

Prototypical examples:

ahaa-elämys (+), empaattinen (+), hahmottaa (+), oivaltaa (+), ymmärrettävä (+), arvoituksellinen (-), epälooginen (-), hämmästyä (-), ihmeissään (-)

päästä jyvälle (+), mennä laskut sekaisin (-), yli ymmärryksen (-)

X2.6 EXPECT

Terms depicting (level of) expectation

Prototypical examples:

aavistaa (+), ennakoida (+), ennuste (+), odotetusti (+), skenaario (+), aavistamaton (-), odottamattomuus (-), yllätys (-), äkkiarvaamatta (-)

elätellä toivoa (+), kuin salama kirkkaalta taivaalta (-), lyödä ällikällä (-)

X3 SENSORY

Prototypical examples:

aistiminen, havainnoida, havaintokyky, tuntuinen, vaistota, havaintokykyinen (+), alitajuinen (-)

X3.1 SENSORY: TASTE

Sensory terms relating to taste

Prototypical examples:

jälkimaku, kitkerä, koemaisto, makeahko, makuinen, suolainen, maittava (+)

X3.2 SENSORY: SOUND

Sensory terms relating to sound

Prototypical examples:

akustiikka, auditiivinen, kaiku, kuulo, surista, ulina, ääntely, korvinkuultava (+), kovaääninen (+), täyttä_kurkkua (+++), hiirenhiljaa (-), mykkä (-), vaientaa (-)

kantautua korviin, kurkku suorana (+++), ei pihaustakaan (-)

X3.3 SENSORY: TOUCH

Sensory terms relating to touch

Prototypical examples:

hamuilla, hipelöiminen, käpälöidä, näpräys, sively, taputella

X3.4 SENSORY: SIGHT

Sensory terms relating to sight

Prototypical examples:

bongaaminen, katsella, näköhavainto, silmäys, visuaalisesti, näkymätön (-), silmät_kiinni (-)

mittailla katseella, vaihtaa katseita

X3.5 SENSORY: SMELL

Sensory terms relating to smell

Prototypical examples:

haistella, hajustettu, hajuaisti, katku, löyhkätä, parfymointi, hajustamaton (-)

X4 MENTAL OBJECT

Entries are sub-classified into the following:

X4.1 MENTAL OBJECT: CONCEPTUAL OBJECT

Terms depicting conceptual objects / objects of the mind (e.g. ideas/concepts)

Prototypical examples:

aatteellinen, aihepiiri, ajatusmaailma, asenne, idea, kriteeri, maailmankatsomuksellisesti, näkökulma, premissi, puheenaihe, seikka

olla kyse

X4.2 MENTAL OBJECT: MEANS, METHOD

Terms relating to mental practises/procedures/resources/techniques

Prototypical examples:

jollakin_tavalla, juju, menetelmä, metodinen, niksi, taktikoida, tekotapa, systemaattisesti

jollakin ilveellä

X5 ATTENTION

Entries are sub-classified into the following:

X5.1 ATTENTION

Terms relating to the (level of) attention

Prototypical examples:

huomiokyky, tarkkaavaisesti (+), uppoutua (+), häiriötekijä (-), mietteissään (-),

poissaoleva (-)

kiinnittää huomiota (+), muissa maailmoissa (-)

X5.2 INTEREST/BOREDOM/EXCITED/ENERGETIC

Terms depicting (level of) interest/energy/boredom etc.

Prototypical examples:

ahkerasti (+), elinvoimaisuus (+), innostus (+), kiinnostaa (+), virike (+), fanatismi (+++), ikävystyä (-), laiska (-), vastahakoisesti (-), välinpitämätön (-)

elämää sykkivä (+), panna tuulemaan (+), maata kuin härski silli (-)

X6 DECIDING

Terms relating to decisions / decision making or the lack of

Prototypical examples:

johtopäätös (+), määrätietoinen (+), nyrkkisääntö (+), päätöksenteko (+), päättämättömyys (-)

pitää päänsä (+), olla kahden vaiheilla (-)

X7 WANTING; PLANNING; CHOOSING

Terms depicting (level of) desire/aspiration

Prototypical examples:

aie (+), haave (+), kunnianhimoinen (+), kysyntä (+), pyrkiä (+), tahallaan (+), toiveikkaasti (+), valita (+), ei-toivottu (-), haluton (-), lempata (-), spontaani (-)

haikailla perään (+), sosiaalinen tilaus (+), ei olla aikomustakaan (-), hetken mielijohteesta (-)

X8 TRYING

Terms depicting (level of) effort/resolution

Prototypical examples:

kokeilla (+), pitkäjännitteinen (+), ponnistelu (+), sinnikkyys (+), sinnitellä (+)

antaa kaikkensa (+), nähdä vaivaa (+)

X9 ABILITY

Entries are sub-classified into the following:

X9.1 ABILITY: ABILITY, INTELLIGENCE

Terms depicting (level of) ability/intelligence

Prototypical examples:

yleistieto, älykkyysosamäärä, ansioitua (+), asiantuntemus (+), kyetä (+), ovelasti (+), suorituskyky (+), huippulahjakas (+++), alokasmainen (-), juntti (-), kyvyttömästi (-), taitamattomuus (-)

olla rahkeita (+), aukko sivistyksessä (-), peukalo keskellä kämmentä (-)

X9.2 ABILITY: SUCCESS AND FAILURE

Terms depicting (level of) success/failure

Prototypical examples:

aikaansaannos, lopputulos, läpimurto (+), pärjätä (+), saavutus (+), tuloksekas (+), epäonnisesti (-), fiasko (-), kariutua (-), menestyksetön (-), tyriä (-)

lyödä itsensä läpi (+), ylittää itsensä (+), mennä pieleen (-), vetää vesiperä (-)

Y SCIENCE & TECHNOLOGY

Y1 SCIENCE AND TECHNOLOGY IN GENERAL

Terms relating to science and technology

Prototypical examples:

alipaine, astronautiikka, biokemiallinen, luonnontiede, mekaanis, optiikka, radioaktiivinen, säteilyttää, teknisesti, valo-oppi

Y2 INFORMATION TECHNOLOGY AND COMPUTING

Terms relating to information technology and computing

Prototypical examples:

alihakemisto, anonyymipalvelin, formatoida, gigatavu, hakukone, HTML, hubi, muistitikku, PDA, tekoäly, verkkoasiointi, virustorjunta, WWW

luonnollisen kielen käsittely, optinen lukija

Z NAMES & GRAMMATICAL WORDS

Z0 UNMATCHED PROPER NOUN

(Not in use in the Finnish semantic lexical resources)

Z1 PERSONAL NAMES

Nouns that distinguish/identify an individual

Prototypical examples:

Ahonen, al-Husseini, Derjabin, Forsblom, Järvinen, Picasso, Condoleezza (f), Johanna (f), Meryl (f), Aleksanteri (m), Gennadi (m), Osama (m), Veikko (m)

Z2 GEOGRAPHICAL NAMES

Nouns that distinguish/identify a specific place

Prototypical examples:

Aasia, Atlasvuoret, Enontekiö, espanjalais, Filippiinit, Hollywood, Hämeenkatu, Kanta-Häme, Madagaskar, Niili, orientaalinen, Pallastunturi, pohjalais, Wales

Addis Abeba, Apenniinien niemimaa, Brittiläinen Kolumbia, Englannin kanaali, Gran Canaria, Hämeen lääni, Skotlannin ylämaat

Z3 OTHER PROPER NAMES

Nouns that distinguish/identify a product, company, etc.

Prototypical examples:

Ajax, Benetton, Etteplan, Fazer, RAY, Ritz, Samsung, Vattenfall, Xerox, YTV

Alma Media, Euroopan avaruusjärjestö, Musta Pörssi, Rank Xerox, Stora Enso

Z4 DISCOURSE BIN

Discourse markers, emphatic communication terms, etc.

Prototypical examples:

ahaa, ai_niin, ciao, huom, hyi, jne., joka_tapauksessa, meni_syteen_tai_saveen, ok, siis, skål

aika on rahaa, auta armias, hauskaa iltaa, Jumalan siunausta

Z5 GRAMMATICAL BIN

Prepositions/adverbs/conjunctions etc.

Prototypical examples:

heti_kun, kannalta, kuluttua, millainen, miten_tahansa, moinen, mones, mukaan, sillä, tällöin

muun muassa, niin kuin, sitten kun

Z6 NEGATIVE

Negative particles

Prototypical examples:

epä, lainkaan, non, yhtään_mikään

ei ensinkään, ei mistään hinnasta, vain kuolleen ruumiini yli

Z7 IF

Conditional terms

Prototypical examples:

jos, mikäli_mahdollista, sikäli_kun

Z8 PRONOUNS ETC.

Pronouns (standard and colloquial) etc.

Prototypical examples:

eräs, hän, joku, kuka_hyvänsä, me, muuan, mä, nämä, siitä, tuo, tämä_kaikki

se jokin

Z9 TRASH CAN

(Not in use in the Finnish semantic lexical resources)

Z99 UNMATCHED

Misspellings or words that have not been included in the lexicon as yet. (This tag is assigned when the program does not recognize a word in the input text.)

References

Aland, K., Black, M., Martini, C.M., Metzger, B.M., & Wikgren, A. (Eds.). (1975). *The Greek New Testament*. New York: United Bible Societies.

Alexander, M., Dallachy, F., Piao, S., Baron, A., & Rayson, P. (2015). Metaphor, popular science, and semantic tagging: Distant reading with the Historical Thesaurus of English. *Digital Scholarship in the Humanities*, *30*(suppl 1), i16–i27.

Al-Hejin, B. (2015). Covering Muslim women: Semantic macrostructures in BBC News. *Discourse & Communication*, *9*(1), 19–46.

Antoniou, G. (2012). *A semantic web primer*. Cambridge, MA: MIT press.

Archer, D., Wilson, A., & Rayson, P. (2002). Introduction to the USAS category system. Retrieved from http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf

Archer, D., McEnery, T., Rayson, P., & Hardie. A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A., & McEnery, T. (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 22–31). Lancaster: Centre for Computer Corpus Research on Language Technical Papers, University of Lancaster.

Archer, D., Rayson, P., Piao, S., & McEnery, T. (2004). Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. In Williams G., & Vessier S. (Eds.), *Proceedings of the 11th EURALEX Congress* (pp. 817–827). Université de Bretagne Sud.

Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In Lüdeling, A., & Kytö M. (Eds.), *Corpus Linguistics: An international handbook* (pp. 613–641). Berlin: Mouton de Gruyter.

Baker, C. F. (2012). FrameNet, current collaborations and future goals. *Language Resources and Evaluation*, *46*(2), 269–286.

Baker, C. F., & Fellbaum, C. (2009). WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 125–129). Association for Computational Linguistics.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Baron, A., & Rayson, P. (2009). Automatic standardization of texts containing spelling variation: How much training data do you need? In Mahlberg, M., González-Díaz, V., & Smith, C. (Eds.), *Proceedings of Corpus Linguistics 2009*. University of Liverpool.

Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, *20*(1), 41–67.

Baron, A., Tagg, C., Rayson, P., Greenwood, P., Walkerdine, J., & Rashid, A. (2011). Using verifiable author data: Gender and spelling differences in Twitter and SMS. Paper presented at ICAME 31, Oslo.

Botchway, B. O. (1989). *The impact of image and perception on foreign policy: An inquiry into American Soviet policy during presidents Carter and Reagan administrations, 1977–1988* (Vol. 37). Tuduv-Verlagsgesellschaft.

Brooke, J. (2001). *A semantic approach to automated text sentiment analysis* (Doctoral thesis). Stanford University.

Calvo Maturana, Maria del Coral. (2012). *Maternidad y voces poéticas en "The Adoption Papers" de Jackie Kay: Un estudio de estilistica de corpus* (Doctoral thesis). Universidad de Granada.

Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning* (pp. 152–164). Association for Computational Linguistics.

Casares, J. (1942). *Diccionario ideológico*. Barcelona: Gustavo Gili.

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC): Pronounced "Luke",… and other useful facts. In McCarthy, P., & Boonthum, C. (Eds.),

Applied natural language processing and content analysis: Identification, investigation, and resolution (pp. 206–209). Hershey, Pennsylvania: IGI Global.

CL Research. (n.d.). Welcome to CL Research. Retrieved from http://www.clres.com

Collins English Dictionary (2000) (electronic resource). Glasgow: Harper-Collins Publishers.

Curran, J. R., & Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised Lexical Acquisition* (pp. 59–66). Association for Computational Linguistics.

Davidson, G. W. (Ed.). (2002). *Roget's Thesaurus of English words & phrases*. London: Penguin Books.

Davis, B., & Mason, P. (2013). Computer-aided identification of stance shifts and semantic themes in electronic discourse analysis. In H. Lim, & F. Sudweeks (Eds.), *Innovative methods and technologies for electronic discourse analysis*. Hershey: ICI.

Day, A.C. (1992). *Roget's Thesaurus of the Bible*. HarperSanFransisco.

Demmen, J., Semino, E., Demjen, Z., Koller, V., Hardie, A., Rayson, P., & Payne, S. (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*, *20*(2), 205–231.

Dornseiff, F. (1970). *Der deutsche Wortschatz nach Sachgruppen*. Berlin: Walter de Gruyter & Co.

El-Haj, M., Rayson, P., Piao, S., & Wattam, S. (Forthcoming). Creating and validating multilingual semantic representations for six languages: Expert versus non-expert crowds. EACL 2017 Workshop on Sense, Concept and Entity Representations and Their Applications. Valencia.

Ellit. (n.d.). Retrieved from http://ellit.fi/. Last accessed November 26, 2012.

ESRC Centre for Corpus Approaches to Social Science (2016). NewsHack 2016 retrospective. Retrieved from http://cass.lancs.ac.uk/?p=1978

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, Mass: MIT Press.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. (2003). Background to Framenet. *International Journal of Lexicography, 16*(3), 235–250.

Fischer, A. (2004). The notional structure of thesauruses. In Kay, C., & Smith, J. J. (Eds.), *Categorization in the history of English* (pp. 41–58). John Benjamins Publishing.

Fondazione Bruno Kessler. (2009a). WordNet Domains hierarchy. Retrieved from http://wndomains.fbk.eu/hierarchy.html

Fondazione Bruno Kessler. (2009b). WordNet Domains labels. Retrieved from http://wndomains.fbk.eu/labels.html

Fortier, P. A. (1989). Some statistics of themes in the French novel. *Computers and the Humanities*, *23*(4), 293–299.

FrameNet. (n.d.-a). About FrameNet. Retrieved from https://framenet.icsi.berkeley.edu/fndrupal/about

FrameNet. (n.d.-b). FrameNets in other languages. Retrieved from https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages

Gacitua, R., Sawyer, P., & Rayson, P. (2008). A flexible framework to experiment with ontology learning techniques. *Knowledge-Based Systems*, *21*(3), 192–199.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language* (pp. 233–237). Association for Computational Linguistics.

Gangemi, A., Guarino, N., & Oltramari, A. (2001). Conceptual analysis of lexical taxonomies: The case of WordNet top-level. In *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 285–296).

Garside, R., & Rayson, P. (1997). Higher-level annotation tools. In Garside, R., Leech, G., & McEnery, T. (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 179–193). New York: Longman.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In Garside, R., Leech, G., & McEnery, T. (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). New York: Longman.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, *28*(3), 245–288.

Granger, S., Paquot, M., & Rayson, P. (2006). Extraction of multi-word units from EFL and native English corpora: The phraseology of the verb "make". *Phraseology in Motion I: Methoden und Kritik*, 57–68.

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T-R., & Alho, I. (2004). *Iso suomen kielioppi* ("Comprehensive grammar of Finnish"). Helsinki: Finnish Literature Society.

Hallig, R., & von Wartburg, W. (1963). *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas*. Berlin: Akademie-Verlag.

Hancock, J. T., Woodworth, M. T., & Porter, S. (2013). Hungry like the wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology*, *18*(1), 102–114.

Hardie, A. (2004). *The computational analysis of morphosyntactic categories in Urdu* (Doctoral thesis). Lancaster University.

Harvard University. (n.d.-a). Welcome to the General Inquirer home page. Retrieved from http://www.wjh.harvard.edu/~inquirer

Harvard University. (n.d.-b). Descriptions of Inquirer categories & use of Inquirer dictionaries. Retrieved from http://www.wjh.harvard.edu/~inquirer/homecat.htm

Harvard University. (n.d.-c). Lasswell Dictionary. Retrieved from http://www.wjh.harvard.edu/~inquirer/lasswell.htm

Harvard University. (n.d.-d). Marker categories. Retrieved from http://www.wjh.harvard.edu/~inquirer/kellystone.htm

Harvard University. (n.d.-e). How the General Inquirer is used and a comparison of General Inquirer with other text-analysis procedures. Retrieved from http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html

Haverinen, K., Laippala, V., Kohonen, S., Missilä, A., Nyblom, J., Ojala, S., ...Ginter, F. (2013). Towards a dependency-based PropBank of general Finnish. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013)* (pp. 41–57). Linköping University Electronic Press.

Heikkinen, V., Lehtinen, O., & Lounela, M. (2001). Kuvia kirjoitetusta suomesta ("Images of written Finnish"). *Kielikello, 3*, pp. 12–15.

Helsingin Sanomat. (n.d.). Helsingin Sanomat. Retrieved from http://www.hs.fi

Hendrickx, I., & Marquilhas, R. (2011). From old texts to modern spellings: An experiment in automatic normalisation. *Journal for Language Technology and Computational Linguistics*, *26*(2), 65–76.

Historical Thesaurus of English. (n.d.). HT Semantic hierarchy. Retrieved from http://historicalthesaurus.arts.gla.ac.uk/downloads/HT_Semantic_Hierarchy.pdf

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Huntsman, J. F. (1975). (Review of the book *A Concept Dictionary of English* by J. Laffal). *Computers and the Humanities, 9*(1), 46–50.

Hüllen, W. (1990). Rudolf Hallig and Walther von Wartburg's Begriffssystem and its non-/acceptance in German linguistics. In Schmitter, P. (Ed.), *Essays towards a history of semantics* (pp. 126–168). Münster: Nodus Publikationen.

Hüllen, W. (2004). *A history of Roget's Thesaurus: Origins, development, and design*. New York: Oxford University Press.

Hüllen, W. (2006). *English dictionaries, 800–1700: The topical tradition*. New York: Oxford University Press.

Hüllen, W. (2009). *Networks and knowledge in Roget's Thesaurus.* New York: Oxford University Press.

Hyvönen, E., Viljanen, K., Tuominen, J., & Seppälä, K. (2008). Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In *Proceedings of the European Semantic Web Conference* (pp. 95–109). Berlin Heidelberg: Springer.

Häkkinen, K. (1994). *Agricolasta nykykieleen* ("From Agricola to modern language"). Juva: Werner Söderström Oy.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, *24*(1), 2–40.

Iltalehti. (n.d.). Iltalehti. Retrieved from http://portti.iltalehti.fi/keskustelu//showthread.php?t=660179. Last accessed January 25, 2012.

Institute for Computational Linguistics "A. Zampolli". (n.d.). Introduction to the EAGLES initiative. Retrieved from http://www.ilc.cnr.it/EAGLES96/edintro/node6.html#SECTION00040000000000000000

Jackson, H., & Zé Amvela, E. (2000). *Words, meaning and vocabulary: An introduction to modern English lexicology*. London: Cassell.

Jäppinen, H. (1989). *Synonyymisanakirja* ("Synonym Dictionary"). Porvoo: WSOY.

Jäppinen, H., & Ylilammi, M. (1986). Associative model of morphological analysis: An empirical inquiry. *Computational Linguistics*, *12*(4), 257-272.

Jönsson-Korhola, H., & White, L. (2002). *Tarkista tästä. Suomen kielen rektioita suomea vieraana kielenä opiskeleville* ("Check here. Finnish rections for learners of the Finnish language"). Helsinki: Finn Lectura.

Kannisto, P., & Kannisto, S. (2005). *La Habanera—Matkalla oravanpyörästä onnenpyörään* ("La Habanera—On the road from the rat race to the wheel of fortune"). Retrieved from http://kirjatohtori.blogspot.com/2005/01/oravanpyora-onnenpyora-habanera.html

Kari, E. (1993). *Naulan kantaan*. *Nykysuomen idiomisanakirja* ("To Hit the Nail on the Head. Idiom Dictionary of Modern Finnish"). Helsinki: Otava.

Karlsson, F. (1999). *Finnish: An essential grammar*. London: Routledge.

Kay, C., & Alexander, M. (2010). Life after the Historical Thesaurus of the OED. *Dictionaries*: *Journal of the Dictionary Society of North America, 31*(1), 107–112.

Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (2009). Unlocking the *OED*: The story of the Historical Thesaurus of the OED. In Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (Eds.), *Historical Thesaurus of the Oxford English Dictionary* (pp. xiii–xx). Oxford: Oxford University Press.

Kettunen, K., & Löfberg, L. (forthcoming). Tagging named entities in 19th century and modern Finnish newspaper material with a Finnish semantic tagger. NoDaLiDa 2017. Gothenburg.

Kettunen, K., Mäkelä, E., Kuokkala, J., Ruokolainen, T., & Niemi, J. (2016). Modern tools for old content—In search of named entities in a Finnish OCRed historical newspaper collection 1771–1910. Krestel, R., Mottin, D., & Müller, E. (Eds.), *Proceedings of conference "Lernen, Wissen, Daten, Analysen" (LWDA 2016*).

Kettunen, K., & Pääkkönen, T. (2016). Measuring lexical quality of a historical Finnish newspaper collection—Analysis of garbled OCR data with basic language technology

tools and means. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Kielipankki (n.d.). Taajuussanasto9996 ("Frequency lexicon9996"). Retrieved from https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotTaajuussanasto9996

Kielitoimiston sanakirja. 2016. Helsinki: Kotimaisten kielten keskus. URN:NBN:fi:kotus-201433, ISSN 2343-1466. Verkkojulkaisu HTML. This publication is updated regularly. Last update February 29, 2016.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, *31*(2), 91–113.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Klebanov, B. B., Diermeier, D., & Beigman, E. (2008). Automatic annotation of semantic fields for political science research. *Journal of Information Technology & Politics*, *5*(1), 95–120.

Knowles, G. (1996). Corpora, databases and the organization of linguistic data. In Thomas, J., & Short, M. (Eds.), *Using corpora for language research* (pp. 36–56). London: Longman.

Kornai, A. (2007). *Mathematical linguistics*. Springer Science & Business Media.

Koskenniemi, K. (2013). *Johdatus kieliteknologiaan, sen merkitykseen ja sovelluksiin* ("An introduction to language technology, its significance and applications"). Helsinki: Helsingin yliopisto. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/38503/kt-johd.pdf?sequence=1

Koskenniemi, K., Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M., ..., Piehl, A. (2012). *The Finnish language in the digital age*. META-NET White Paper Series. Springer. Retrieved from http://www.meta-net.eu/whitepapers/e-book/finnish.pdf

Kotimaisten kielten keskus (2007). Presidentti Halosen uudenvuodenpuheet ("President Halonen's New Year's speeches"). Retrieved from http://kaino.kotus.fi/korpus/teko/meta/presidentti/halonen/halonen_coll_rdf.xml

Kotimaisten kielten keskus (2006). Suomalaisen kirjallisuuden klassikoita ("Classics of Finnish literature"). Retrieved from http://kaino.kotus.fi/korpus/klassikot/meta/klassikot_coll_rdf.xml

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Kytzler B. (2005). (Review of the book *Statistischer Schlüssel zum Vokabular in Vergils Eklogen* by D. Najock). *Bryn Mawr Classical Review*. Retrieved from http://bmcr.brynmawr.edu/2005/2005-01-14.html

Laffal, J. (1967). Characteristics of the three-person conversation. *Journal of Verbal Learning and Verbal Behavior*, *6*(4), 555–559.

Laffal, J. (1973). *A concept dictionary of English*. Halsted Press.

Laffal, J. (1970). Toward a conceptual grammar and lexicon. *Computers and the Humanities*, *4*(3), pp.173–186.

Laffal, J. (1995). A concept analysis of Jonathan Swift's A Tale of a Tub and Gulliver's Travels. *Computers and the Humanities*, *29*(5), 339–361.

Lancaster University. (n.d.). Wmatrix corpus analysis and comparison tool. Retrieved from http://ucrel.lancs.ac.uk/wmatrix

Lancaster University. (2016). About VARD 2. Retrieved from http://ucrel.lancs.ac.uk/vard/about

Lee, D., Jeong, O. R., & Lee, S. G. (2008). Opinion mining of customer feedback data on the web. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication* (pp. 230-235). ACM.

Leech, G. (1997). Introducing corpus annotation. In Garside, R., Leech, G., & McEnery, T. (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1–18). New York: Longman.

Leino, A., & Leino, P. (1990). *Synonyymisanasto* ("Synonym lexicon"). Helsinki: Otava.

L'Hôte, E., & Lemmens, M. (2009). Reframing treason: Metaphors of change and progress in new Labour discourse. *CogniTextes. Revue de l'Association française de linguistique cognitive*, *3*.

Lindén, K., & Carlson, L. (2010). FinnWordNet—WordNet på finska via översättning ("FinnWordNet—WordNet of Finnish via Translation"). *LexicoNordica—Nordic Journal of Lexicography*, *17*, 119–140.

Lindén, K., & Niemi, J. (2014). Is it possible to create a very large wordnet in 100 days? An evaluation. *Language Resources and Evaluation*, *48*(2), 191–201.

Lindén, K., Niemi, J., & Hyvärinen. M. (2012). Extending and updating the Finnish WordNet. In Santos, D., Lindén, K., & Ng'ang'a, W. (Eds.), *Shall we play the festschrift game? Essays on the occasion of Lauri Carlson's 60th birthday* (pp. 67–98). Berlin: Springer.

Litkowski, K. C. (1997). Category development based on semantic principles. *Social Science Computing Review*, *15*, 394–409.

Louw, J. P. (1973). Discourse analysis and the Greek New Testament. *The Bible Translator*, *24*(1), 101–118.

Louw, J. P. (1982). *Semantics of New Testament Greek*. Augsburg Fortress Publishing.

Louw, J. P., & Nida, E. A. (1988). *Greek-English lexicon of the New Testament, based on semantic domains*. New York: United Bible Societies.

Löfberg, L., Juntunen, J-P., Nykänen, A., Varantola, K., Rayson, P., & Archer, D. (2004). Using a semantic tagger as dictionary search tool. In Williams, G., & Vessier, S. (Eds.), *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)* (pp. 127–134). Lorient: Université de Bretagne Sud.

Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P., Nykänen, A., & Varantola, K. (2005). A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics 2005 Conference*. Proceedings from the Corpus Linguistics Conference Series on-line e-journal.

Macmillan. (n.d.). *Macmillan Dictionary*. Retrieved from http://www.macmillandictionary.com

Magnini, B., & Cavaglia, G. (2000). Integrating subject field codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S., & Stainhaouer G. (Eds.), *Proceedings of LREC 2000, Second International Conference on Language Resources and Evaluation* (pp. 1413–1418). Athens.

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS One*, *9*(8), e105937.

McArthur, T. (1981). *Longman lexicon of contemporary English*. Harlow: Longman.

McArthur, T. (1986). *Worlds of reference: Lexicography, learning and language from the clay tablet to the computer*. Cambridge: Cambridge University Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McTavish, D. G., & Pirro, E. B. (1990). Contextual content analysis. *Quality and Quantity*, *24*(3), 245–265.

Miller, G. A. (1998). Foreword. In Fellbaum, C. (Ed.), *WordNet: An electronic lexical database* (pp. xv–xxii). Cambridge, Mass: MIT Press.

MOT Collins English Dictionary. HarperCollins Publishers. Available via the commercial MOT language service https://www.sanakirja.fi

Mudraya, O., Babych, B., Piao, S., Rayson, P., & Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. In *Proceedings of Corpus Linguistics 2006* (pp. 290–297). St. Petersburg.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Najock, D. (2004). *Statistischer Schlüssel zum Vokabular in Vergils Eklogen* (Vol. 177). Georg Olms Verlag.

Narayanan, S., Fillmore, C. J., Baker, C. F., & Petruck, M. R. L. (2002). FrameNet meets the Semantic Web: A DAML+OIL frame representation. In *Proceedings of the 18th National Conference on Artificial Intelligence*. Edmonton, Alberta: AAAI.

Nenonen, M. (2002). *Idiomit ja leksikko*. *Suomen kielen lausekeidiomien syntaktisia, semanttisia ja morfologisia piirteitä* ("Idioms and the lexicon. Syntactic, semantic and morphological features of Finnish phrasal idioms") (Doctoral thesis). University of Joensuu.

Nida, E. A. (1949). *Morphology: The descriptive analysis of words*. Ann Arbor: University of Michigan Press.

Nida, E. A. (1975). *Componential analysis of meaning: An introduction to semantic structures*. The Hague: Mouton.

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation*. Brill.

Nurmi, T. (1998). *Uusi suomen kielen sanakirja* ("New Dictionary of Finnish"). Helsinki: Gummerus Kustannus.

Ogilvie, D.M., Stone, P.J., & Shneidman, E.S (1966). Some characteristics of genuine versus simulated suicide notes. In Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M., *The General Inquirer: A computer approach to content analysis* (pp. 527–535). Cambridge: The MIT Press.

O'Halloran, K. (2011). Limitations of the logico-rhetorical module: Inconsistency in argument, online discussion forums and electronic deconstruction. *Discourse Studies*, *13*(6), 797–806.

Ollikainen, L. (1994). *Messages from the point of no return: A conceptual and empirical analysis of suicide notes left by suicide victims* (Doctoral thesis). Turku: University of Turku.

Ooi, V. B. Y, Tan, P. K. W., & Chiang, A. K. L. (2007). Analyzing personal weblogs in Singapore English: The WMatrix approach. *eVariEng (Journal of the Research Unit for Variation, Contacts, and Change in English), 2*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy*, *31*(6), 539–548.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Piao, S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, T., & Wilson, A. (2005). A large semantic lexicon for corpus annotation. In *Proceedings of the Corpus Linguistics 2005 Conference*. Proceedings from the Corpus Linguistics Conference Series on-line e-journal.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., & Rayson, P. (2015). Development of the multilingual semantic annotation system. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL HLT 2015)* (pp. 1268–1274).

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Phoey, L.T., & Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In Calzolari et al. (Eds.), *10th edition of the Language Resources and Evaluation Conference (LREC2016)* (pp. 2614–2619). European Language Resources Association (ELRA).

Piao, S., Rayson, P., Archer, D., & McEnery, T. (2004). Evaluating lexical resources for a semantic tagger. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 499–502). Lisbon.

Piao, S., Rayson, P., Archer, D., & McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language 19*(4), 378–397.

Piao, S., Rayson, P., Archer, D., Wilson, A., & McEnery, T. (2003). Extracting multi-word expressions with a semantic tagger. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, ACL 2003* (pp. 49–56). Sapporo.

Preiss, J., & Stevenson, M. (2004). Introduction to the Special Issue on Word Sense Disambiguation. *Computer Speech and Language*, *18*(3), 201–207.

Prentice, S., Taylor, P. J., Rayson, P., Hoskins, A., & O'Loughlin, B. (2011). Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict. *Information Systems Frontiers*, *13*(1), 61–73.

Pulkkinen, P. (1972). *Nykysuomen kehitys* ("The development of modern Finnish"). Suomalaisen Kirjallisuuden Seura.

Qian, Y., & Piao, S. (2009). The development of a semantic annotation scheme for Chinese kinship. *Corpora*, *4*(2), 189–208.

Rada Mihalcea. (n.d.). Senseval. Retrieved from http://www.senseval.org

Rashid, A., Greenwood, P., Walkerdine, J., Baron, A., & Rayson, P. (2012). Technological solutions to offending. In Quayle, E., & Ribisl, K. (Eds.), *Understanding and preventing online sexual exploitation of children* (pp. 228–243). London: Willan.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison* (Doctoral thesis). Lancaster University.

Rayson, P. (2005). Right from the word go: Identifying multi-word-expressions for semantic tagging. Invited talk at BAAL Corpus Linguistics SIG/OTA Workshop Identifying and Researching Multi-word Units.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP tasks* (LREC 2004) (pp. 7–12). Lisbon.

Rayson, P., & Baron, A. (2011). Automatic error tagging of spelling mistakes in learner corpora. In Meunier F., De Cock S., Gilquin G., & Paquot M. (Eds.), *A taste for corpora. In honour of Sylviane Granger* (pp. 109–126). Amsterdam: John Benjamins.

Rayson, P., Emmet, L., Garside, R., & Sawyer, P. (2001). The REVERE project: Experiments with the application of probabilistic NLP to systems engineering. In *Natural Language Processing and Information Systems* (pp. 288–300). Berlin Heidelberg: Springer.

Rayson, P., & Smith, N. (2006). The key domain method for the study of language varieties. In *The Third Inter-Varietal Applied Corpus Studies (IVACS) group International Conference on" Language at the Interface"*. University of Nottingham.

Rayson, P., & Stevenson, M. (2008). Sense and semantic tagging. In Lüdeling, A., & Kytö, M. (Eds.), *Corpus linguistics*. *An international handbook* (pp. 564–579). Berlin: Mouton de Gruyter.

Roberts, J., Kay, C., & Grundy, L. (1995). *A Thesaurus of Old English*. London: King's College.

Robertson, T. (1859). *Dictionnaire idéologique: Recueil des mots, des phrases, des idiotismes et des proverbes de la langue française classés selon l'ordre des idées*. Derache.

Roget, P. (1852). Introduction to the first edition. In Davidson, G. W. (Ed.) (2002), *Roget's Thesaurus of English Words & Phrases*. London: Penguin Books.

Roget's Thesaurus of the Bible. (n.d.). Colin Day & Roget's Thesaurus of the Bible. Retrieved from http://www.colinday.co.uk

Rundell, M. (Ed.). (2002). *Macmillan English Dictionary*. London: Macmillan.

Ruokatieto. (n.d.). Suomalaisen ruokakulttuurin ulottuvuuksia ("Dimensions of Finnish culinary culture"). Retrieved from http://www.ruokatieto.fi/Suomeksi/Ruokakulttuuri

Ruppel, K. (Forthcoming). The lexicography of Finnish. In P. Hanks, & G-M. de Schryver (Eds.), *International handbook of modern lexis and lexicography*. Berlin: Springer-Verlag.

Ruppel, K., & Sandström, C. (2014). Stora finska ordböcker i ett historiskt perspektiv ("Large Finnish dictionaries in a historical perspective"). *LexicoNordica*, *21*, 141–160.

Sanders, D. (1873). *Deutscher Sprachschatz geordnet nach Begriffen zur leichten Auffindung und Auswahl des passenden Ausdrucks: Ein stilistisches Hülfsbuch für jeden Deutsch Schreibenden* (Vol. 1). Hoffmann & Campe.

Saukkonen, P., Haipus, M., Niemikorpi, A., & Sulkala, H. (1979). *Suomen kielen taajuussanasto* ("Frequency lexicon of Finnish"). WSOY: Porvoo.

Schmidt, K.M. (1986). Concept versus meaning: The contribution of computer-assisted content analysis and conceptual analysis to this disputed area. In *En hommage à Charles Muller: Méthodes quantitatives et informatiques dans l'étude des textes* (pp. 779–993). Geneva: Slatkine.

Schütze, H., & Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, *33*(3), 307–318.

Semantic Computing Research Group. (n.d.-a). National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012. Retrieved from http://www.seco.tkk.fi/projects/finnonto

Semantic Computing Research Group. (n.d.-b). Ontologies by SeCo. Retrieved from http://www.seco.tkk.fi/ontologies

Semino, E., Hardie, A., Koller, V., & Rayson, P. (2005). In Barnden, J., Lee, M., Littlemore, J., Moon, R., Philip, G., & Wallington, A. (Eds.) *Corpus-based approaches to figurative language: A Corpus Linguistics 2005 colloquium* (pp. 145–154). Birmingham: University of Birmingham Cognitive Science Research Papers.

Seppälä, K., & Hyvönen, E. (2004). Asiasanaston muuttaminen ontologiaksi. Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista ("Changing a keyword thesaurus into an ontology. General Finnish Ontology as an example of the FinnONTO Model"). Helsinki: National Library of Finland.

Shapero, Jess Jann (2011). *The language of suicide notes* (Doctoral thesis.). University of Birmingham.

Sharoff, S., Babych, B., Rayson, P., Mudraya, O., & Piao, S. (2006). ASSIST: Automated Semantic Assistance for Translators. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations* (pp. 139–142). Association for Computational Linguistics.

Shawar, B. A., & Atwell, E. (2003). Using dialogue corpora to train a chatbot. In *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 681–690). Lancaster: Centre for Computer Corpus Research on Language Technical Papers, University of Lancaster.

Silfverberg, M. (2015). Reverse engineering a rule-based Finnish named entity recognizer. Retrieved from https://kitwiki.csc.fi/twiki/pub/FinCLARIN/KielipankkiEventNERWorkshop2015/Silfverberg_presentation.pdf.

Simm, W., Ferrario, M. A., Piao, S., Whittle, J., & Rayson, P. (2010). Classification of short text comments by sentiment and actionability for VoiceYourView. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on* (pp. 552–557). IEEE.

Simpson, J. A., & Weiner, E. S. C. (Eds.). (1989). *The Oxford English Dictionary*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. Cambridge: The M.I.T. Press.

Stevenson, M., & Wilks, Y. (2003). Word sense disambiguation. In *Oxford handbook of computational linguistics* (pp. 249–265). Oxford University Press.

Suomi24. (n.d.). Kuolema ja suru ("Death and mourning"). Retrieved from http://keskustelu.suomi24.fi/debate/3838

Tagg, C., Baron, A., & Rayson, P. (2012). "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. *Lingvisticae Investigationes*, *35*(2), 367–388.

The Global WordNet Association. (n.d.). Wordnets in the world. Retrieved from http://globalwordnet.org/wordnets-in-the-world

Thomas, J., & Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas, & M. Short (Eds.) *Using corpora for language research* (pp. 92–109). London: Longman.

Tjong Kim Sang, E. F., & De Meulder, F. (2003, May). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (pp. 142–147). Association for Computational Linguistics.

Unger, L. (2007). *Kauniita valheita* ("Beautiful lies"). Retrieved from http://www.elle.fi/lifestyle/nettikirja/?sec=kokotarina&r=true. Last accessed November 12, 2012.

University Centre for Computer Corpus Research on Language. (n.d.-a). Benedict—The New Intelligent Dictionary. Retrieved from http://ucrel.lancs.ac.uk/projects.html#benedict

University Centre for Computer Corpus Research on Language. (n.d.-b). Hallig_Wilson model. Retrieved from http://ucrel.lancs.ac.uk/usas/Hallig_Wilson/Hallig_Wilson_Frameset.htm

University Centre for Computer Corpus Research on Language. (n.d.-c). Louw & Nida model. Retrieved from http://ucrel.lancs.ac.uk/usas/Louw&Nida/Louw&Nida_frameset.htm

University Centre for Computer Corpus Research on Language. (n.d-d). UCREL CLAWS7 tagset. Retrieved from http://ucrel.lancs.ac.uk/claws7tags.html

University of Glasgow. (n.d.-a). Conference papers, talks and reports. Retrieved from http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/metaphor/conferences

University of Glasgow. (n.d.-b). Semantic Annotation and Mark-Up for Enhancing Lexical Searches. Retrieved from http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels

Utriainen, T. & Honkasalo, M-L. (1996). Women writing their death and dying: Semiotic perspectives on women's suicide notes. *Semiotica, 109*(3–4), 197–220.

Valderrábanos, A. S., Dïaz, E. T., & Pérez, M. A. D. P. (1994). An automatic information extraction system for the Diccionario Ideológico de la Lengua Española by Julio Casares. *Literary and Linguistic Computing*, *9*(3), 203–208.

Véronis, J. (2001). Sense tagging: Does it make sense? In *Proceedings of Corpus Linguistics 2001*. Lancaster.

Wehrle, H., & Eggers, H. (1961). *Deutscher Wortschatz*. Stuttgart: Ernst Klett Verlag.

Whissell, C. M., & Dewson, M. R. (1986). A dictionary of affect in language: III. Analysis of two biblical and two secular passages. *Perceptual and Motor Skills*, *62*(1), 127–132.

Wilson, A., & Leech, G. N. (1993). Automatic content analysis and the stylistic analysis of prose literature. *Revue: Informatique et Statistique dans les Sciences Humaines 29*, 219–234.

Wilson, A., & Rayson, P. (1993). Automatic content analysis of spoken discourse. In Souter, C., & Atwell, E. (Eds.), *Corpus-based computational linguistics* (pp. 215–226). Amsterdam: Rodopi.

Wilson. A., & Thomas, J. (1997). Semantic annotation. In Garside, R., Leech, G., & McEnery, T. (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 53–65). New York: Longman.

Wilson, A. (2002). Developing conceptual glossaries for the Latin Vulgate Bible. *Literary and Linguistic Computing*, *17*(4), 413–426.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 347–354). Association for Computational Linguistics.

Yle. (n.d.). Yle. Retrieved from http://yle.fi/keskustelut/printthread.php?t=3511&pp=40. Last accessed February 15, 2012.

Zillig, W. (2014). Wer war Hugo Wehrle? Retrieved from http://epub.ub.uni-muenchen.de/20824/1/zillig_wer_war_hugo_wehrle_version_2.3.pdf