A semantic tagger for the Finnish language

Laura Löfberg[1], *Scott Piao*[2], *Paul Rayson*[2],
Jukka-Pekka Juntunen[3], *Asko Nykänen*[3], *and Krista Varantola*[1],
[1] University of Tampere, Finland
laura.lofberg@uta.fi krista.varantola@uta.fi
[2] Lancaster University, UK
s.piao@lancaster.ac.uk paul@comp.lancs.ac.uk
[3] Kielikone Ltd., Finland
jpj@kielikone.fi asko@kielikone.fi

Abstract

This paper reports on the current status and evaluation of a Finnish semantic tagger (hereafter FST), which was developed in the EU-funded Benedict Project. In this project, we have ported the Lancaster English semantic tagger (USAS) to the Finnish language. We have re-used the existing software architecture of USAS, and applied the same semantic field taxonomy developed for English to Finnish. The Finnish lexical resources have been compiled using various corpus-based techniques, and the resulting lexicons have then been manually tagged and used for the FST prototype. At present, the lexicons contain 33,627 single lexical items and 8,912 multi-word expression templates.

In the evaluation, we used two sets of test data. The first test data is from the domain of Finnish cooking, which is both sufficiently compact and sufficiently versatile. The second data is from Helsingin Sanomat, the biggest Finnish daily newspaper. As a result, the FST produced a lexical coverage of 94.1% and a precision of 83.03% on the cooking test data and a lexical coverage of 90.7% on the newspaper data. While there is much room for improvement, this is an encouraging result for a prototype tool. The FST will be continually improved by expanding the semantic lexical resources and improving the disambiguation algorithms.

1. Introduction

Corpus annotation is a vital part of corpus-based language study and NLP (natural language processing). Over the last several decades, a wide range of annotation schemes and tools have been developed, such as POS tagging, syntactic parsing and named entity identification and classication, etc. The research in this area has mainly focussed on English, but in the last five years more tools are becoming available for other languages as well. Recently, semantic annotation has received increasing attention in this research area, and various tools have been developed for this purpose. For example, in Lancaster an English semantic tagger has been developed for annotating corpora with semantic field information (Rayson et al, 2004). Such a semantic annotation scheme and tool have various applications, such as discourse analysis, text domain analysis, information extraction and software requirements engineering (Sawyer et al, 2002). In this paper, we present our work on the development of software and lexical resources for the Finnish language based on the Lancaster semantic tagger and taxonomy and report on our evaluation of this tool.

Very little previous work has been reported in the area of semantic tagging of the Finnish language. In addition, there were no Finnish lexical sample tasks held at the first three *Senseval* word sense disambiguation evaluation exercises[1]. A sub-task of semantic tagging is that of semantic labelling of named entities (terms such as people, organisations, places and temporal expressions). In Finnish this is the approach taken by Connexor's Machinese Metadata tool[2], Aunimo et al (2004), and Makkonen et al (2002). To carry out more extensive Finnish semantic tagging, Cheadle and Gambäck (2003) extended Connexor's Machinese Syntax tool with sense annotation for an adaptive speech interface. Lagus et al (2002) applied clustering algorithms to Finnish verbs in a corpus of 13 million words of magazines and newspapers.

Our FST was developed in the EU-funded Benedict Project[3]. The aim of this project was to discover an optimal way of catering for the needs of dictionary users in modern electronic dictionaries by using state-of-the-art language technology. Studies of dictionary use and the potential of modern electronic dictionaries were used to define the needs of users. A major feature of the Benedict intelligent dictionary, the end-result of the project, is a context-sensitive dictionary search tool. It not only helps the user to find the correct main entry but also highlights the relevant sense of the looked-up item. This search tool is based on the English and Finnish semantic taggers, which provide semantic field information for the words under consideration. In contrast to "shallow intelligence" applications such as spelling correction and morphological analysis, assisted by the semantic taggers, this tool is capable of doing "deeper intelligence" searches in order to capture the correct word sense. Varantola (forthcoming) defines the shallow and deep intelligence as follows: ”Shallow intelligence could be used to describe what spell-checking systems and cross-referencing links in dictionary entries do. These systems will help in determining the correct spelling and finding synonyms, near-synonyms, antonyms and more 'mechanical' information in general. Deeper intelligence, on the other hand, would entail access to user-definable user profiles, user-specified filters and display modes, such as browser modes and look-up modes, full and reduced displays of data categories, user alerts, etc." Essentially, the correct sense of search word is specified by using the context of the search word for disambiguation.

In this paper we will present our evaluation of the current FST, focusing on the following issues:

a) Problems caused by the widely different grammatical systems of English and Finnish during the construction of the Finnish semantic lexicon;
b) Technical questions to be solved when dealing with the morpho-syntactic features of Finnish;
c) Evaluation of the lexical coverage of the FST;
d) Evaluation of the precision of the FST and error analysis;
e) Problems to be solved in the future development of the FST.

This paper is a follow-up of our previous report on the on the early stage of the FST development (Löfberg et al. 2003). In the following sections, we discuss the evaluation of the current stage of the FST (June 2005). Our evaluation demonstrates that the tagger has already achieved a high lexical coverage and an encouraging level of precision. On the other hand, the problems encountered in this latest evaluation present tough and intriguing challenges to software development and to research on the construction of semantic lexicons.

[1] See http://www.senseval.org

[2] See http://www.connexor.com/software/metadata/

[3] This collaborative project (IST-2001-34237) was carried out from March 2002 to February 2005 involving the University of Tampere, Gummerus, Nokia and Kielikone from Finland plus Lancaster University and Collins Dictionaries from the UK. For further information see http://mot.kielikone.fi/benedict/

2. Development of the FST software

Aiming to provide semantic tools for bridging across the English and Finnish languages, during the development of the FST, we put much effort to achieve a close compatibility between the English and Finnish semantic taggers. We adopted the approach of porting the English semantic tagger to the Finnish language, both in terms of software architecture and semantic taxonomy. Although some adjustments and modifications were inevitable to cope with some unique features of Finnish language, our approach has been proven very successful.

In the Benedict project we have worked on both improving the existing EST (English Semantic Tagger) and developing a parallel tool for Finnish. With regard to the development of the FST software, we wished to evaluate the applicability of the existing English software framework for Finnish. Therefore the FST is largely based on the architecture of the existing EST (Java version), which has been designed in an Object-Oriented model[4]. The semantic categories developed for the EST were compatible with the semantic categorizations of objects and phenomena in Finnish. We must, however, keep in mind that Finnish is a non-Indo-European language employing morphological rules which are very different from those of English. In order to cope with the unique features of Finnish some modifications and changes were thus inevitable.

Unlike English, Finnish is a highly inflected language: generally, what is expressed in English through phrases or syntactic structures is expressed in Finnish via morphological affixation. For example, case endings are used to express relations between words (instead of prepositions) and morphemes are used to express plural and possessive relations as well as to denote morpho-syntactic concepts pertaining to verbs:

kukissani (in/among my flowers)
kuk/i/ssa/ni (base nominative form/plural marker/inessive case/possessive affix)

Tulisitko? (Would you come?)
tul/isi/t/ko (base verb form/conditional mood/second person singular/clitic affix)

Clearly, due to such flexible inflectional/derivational morphological changes as well as the numerous morphemes that can be attached to the base forms of Finnish nouns, verbs and adjectives can carry a very high information load. Other differences compared to English include:

- Word order is relatively free but not random.
- There are no articles.
- The Finnish language does not differentiate between genders.
- When the predicate verb is negated, the negation *ei* ('not') takes on the conjugation form that indicates person.

First of all, we needed a tool for analysing and decomposing the complex morpho-syntactic structures of Finnish words. For this purpose, we adopted a Finnish morpho-syntactic analyser and parser, named TextMorfo. TextMorfo provides an efficient and accurate tool for the analysis and decomposition of Finnish lexical items. Given a Finnish text, it extracts stems, lemmas, POS information etc. for each word. TextMorfo is used as the equivalent of the CLAWS POS tagger in the English semantic tagger framework. Figure 1 illustrates the parallel architecture of the USAS system consisting of EST and FST.

[4] Object-Oriented model is a software engineering technology for building flexible and adaptable software systems, which is being widely applied today (for further details, see Sommerville 2001).

Furthermore, although most of the letters of the Finnish alphabet are the same as in English, there are three additional letters in Finnish, Å, Ä, Ö, whose values fall outside the basic ASCII code set. To cope with this problem, we adopted the Unicode UTF-8 encoding scheme for the whole project. This freed us from a complex conversion problem in encoding. Adopting Unicode would also allow us to easily extend our framework to many other languages.

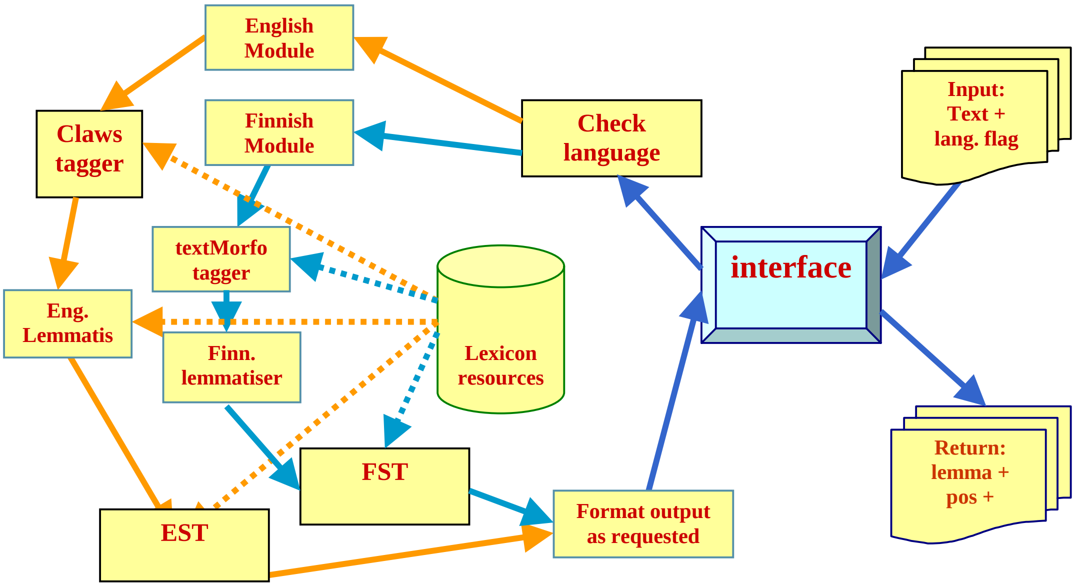

Figure 1: Outline of USAS package including EST and FST

Another distinct feature of Finnish language is its widespread use of compounds; English multi-word expressions are often conveyed by compounds in Finnish language. Finnish compounds are typically formed by attaching two or more words together without space in between. Consequently these compounds function syntactically as single words. Most often compounds are formed of nouns, but words of other parts of speech can also appear in compounds. In another study, we focussed on phrasal verbs in the English multi-word-expression (hereafter MWE) list and showed that the Finnish equivalents are single word items resulting in a shift in balance between lexical resources across language (Mudraya et al, forthcoming).

We differentiate between two types of compounds: (a) lexically petrified compounds, and (b) secondly transparent compounds and ad hoc compounds. The meaning of a petrified compound (e.g. *tietokone*, *sähköpaimen*; see Table 1 below) is not equal to the sum of the meanings of its constituent parts. Such compounds normally occur as headword entries in dictionaries, and hence we included them in the lexicon of single lexical items as individual entries. The second group of transparent or ad hoc compounds (e.g. *kalakeitto*, *nilkkavamma*, see below) have meanings that can be deduced from that of the element words. The meaning of a transparent lexicalized compound is not necessarily the sum of the meanings of its parts, but they are close enough to be deduced from the parts. In terms of the semantic granularity, we consider them sufficiently different for our purposes. Table 1 shows some sample Finnish compounds.

In the case of ad hoc compounds, the meaning is clearly the sum of the parts. For the purposes of our analysis, it is possible to group these two basic types into the same category of transparent compounds. For example,

keittokirja (’cookery book’) – lexicalized and semi-transparent
keittokirjavalikoima – *keittokirja* (’cookery book’) + *valikoima* (’selection’) = ad hoc compound

There are practically no limits to the number of possible compounds. Therefore it would be simply impossible to try to include all the possible combinations in the lexicon. To solve this problem, we have added a new component to the FST framework called the ‘compound engine’ to identify and flag Finnish compounds that are not included in the lexicon. For example, the compound engine would produce the following entry for the compound *nilkkavamma:*

nilkkavamma : <w pos="Noun/Noun" mwe="com" sem="B2-/B1" lem="vamma/nilkka">nilkkavamma</w>.

As shown above, the second part of the compound is placed first. This is because it is more significant in terms of meaning (in this case *vamma* ‘injury’). Consequently, the first part that modifies the second part of the compound is placed second (in this case *nilkka* ‘ankle’) after a slash.

Finnish compound	**Constituents**		**English equivalent**
tietokone	tieto / kone	'knowlegde' 'machine'	'computer'
sähköpaimen	sähkö / paimen	'electricity' 'shepherd'	'electric cattle fence'
kalakeitto	kala / keitto	'fish' 'soup'	'fish soup'
nilkkavamma	nilkka / vamma	’ankle' ’injury'	’ankle injury'

Table 1: Examples of Finnish compounds

Despite the modifications and changes described above, in general the architecture of the FST software mirrors that of the EST components. This makes it easier to maintain and improve the tools as a single package. We are currently applying and evaluating the same framework for Russian in the Assist project[5].

3. Creating lexical resources for FST

The main lexical resources of the FST include lexicons for tagging single words and multi-word expressions. We built the Finnish lexical resources using a variety of corpus-based techniques, and the resulting wordlists were then manually semantically classified. In the beginning we tagged the 6,000 most frequent Finnish words based on a large corpus and some other word lists from different domains. We have exploited readily available resources, including word lists of different domains from Kielikone’s machine translation lexicon and the Web; however, a meticulous post-editing phase has still been essential. Afterwards, the lexicon has been further expanded by feeding texts from various sources into the FST and classifying words that remain unmatched. Overall, the lexicon development has for the most part been manual work which is both laborious and time-consuming. Nonetheless, as will be shown, this has assured a high quality and reliability of lexical resource.

At present, the FST lexicons contain 33,627 single lexical items and 8,912 multi-word expression templates. This compares to 52,785 single lexical items and 18,809 MWEs in the EST. During the

[5] The Automated Semantic Assistance for Translators project, see the ASSIST website for more details: http://www.comp.lancs.ac.uk/ucrel/projects.html#assist

compilation of the Finnish lexicons, we have used the identical tagset and followed the same guidelines as those applied to the USAS English lexicons. Theoretically speaking, therefore, the English and Finnish lexicons are comparable[6]. Nevertheless, the structures of lexical entries are slightly different.

The main difference lies in the fact that the English lexicon contains inflectional variants whereas the Finnish counterpart consists of only lemmas, or base forms. Because we had no reliable automatic English lemmatiser at the start of the EST construction, and there are limited number of English inflectional forms in English, it was decided to include inflectional forms in the English lexicon[7]. However, our observation on the Finnish morpho-syntactic structure soon revealed it is not a practical approach to the FST lexicon construction. Due to the highly inflectional and agglutinative nature of Finnish, if we included inflectional variants in the FST lexicon, that would have resulted into an uncontrollable size of lexicon. Provided with the highly accurate Finnish morpho-syntactic analyser TextMorfo, we decided to compile the Finnish lexicon only with lemmas/basic forms. When FST is applied to running text, TextMorfo is used to reduce the Finnish words into lemmas and basic forms, which are matched against the lexicon entries.

In addition, a couple of extra POS tags are used for Finnish lexicon. For example, the tag *CompPart* is used to mark a special group of Finnish words which solely appear as the first part of compounds and which are never used independently. Another such tag is PL Marker, which is used to mark those nouns that usually appear in plural form. Also we treat fixed expressions such as *viime_aikoina* 'lately') as single lexical items. Figure 2 and Figure 3 show sample entries of single-word and MWE lexicons respectively.

vieras	Adjective	A6.2- X2.2- Z2 A9-
vihkiäis	CompPart	S4
Wihuri	Proper	Z3/I2.1 Z1
Viiala	Proper	Z2
viidestoista	Numeral	N4
viikonloppuisin	Adverb	T1.3
viilentyä	Verb	O4.6-/A2.1 S1.2.1-/A2.1
viileäkaappi	Noun	O4.6-/O3
viime_aikoina	Adverb	T3---
villahousut	Noun PL	B5

Figure 2: Examples from the lexicon of single lexical items

onneksi olkoon	Z4
onnesta sekaisin	E4.1+
onni suosii rohkeaa	Z4
optinen lukija	Y2
Oracle Finland	Z3/I2.1
osoittaa elonmerkkejä	L1+
ostaa sika säkissä	I2.2/X2.4-

Figure 3: Examples from the lexicon of multi-word expression templates

[6] The full tagset and guidelines are available on-line at http://www.comp.lancs.ac.uk/ucrel/usas/. The tags used in the examples of this paper are explained in the appendix.

[7] In fact, subsequently we did include an English lemmatiser in the system based on earlier research using POS tags (Beale, 1987).

The Finnish lexicons are used by the FST software for identifying semantic categories of the words in running texts. Again for the sake of compatibility, the FST produces output in similar format as that of EST, except using an attribute tag ***mwe="com"*** to mark-up Finnish compounds. Fig. 4 shows a sample output of FST. As shown in this sample, the word “keittokirjan” (cookery book) is identified as a compound, which consists of constituents *keitto* and *kirja*. Actually the compound was missing from the lexicon, but the compound engine was able to process it successfully.

```
<w pos="Pronoun" mwe="0" sem="Z8" lem="eräs">Erään</w>
<w pos="Noun/Noun" mwe="com" sem="Q4.1/F1 Q1.2/F1 N5/Q1.2/F1" lem="kirja/keitto">keittokirjan</w>
<w pos="Preposition" mwe="0" sem="Z5" lem="mukaan">mukaan</w>
<w pos="Adverb" mwe="0" sem="Z5" lem="tällöin">tällöin</w>
<w pos="Verb" mwe="0" sem="A3+ Z5" lem="olla">on</w>
<w pos="Adverb" mwe="0" sem="A14 Z4" lem="vain">vain</w>
<w pos="Verb" mwe="0" sem="S8+ E6- A1.1.1" lem="huolehtia">huolehdittava</w>
<w pos="Pronoun" mwe="0" sem="Z8" lem="se">siitä</w>
<w pos="_Delimiter" mwe="0" sem="PUNC" lem=",">,</w>
<w pos="Conjunction" mwe="0" sem="Z5" lem="että">että</w>
<w pos="Noun" mwe="0" sem="F1" lem="mauste">mausteita</w>
<w pos="Verb" mwe="0" sem="Z6" lem="ei">ei</w>
<w pos="Verb" mwe="0" sem="X2.2-" lem="unohtaa">unohdeta</w>
<w pos="_Delimiter" mwe="0" sem="PUNC" lem=".">.</w>
<w pos="_EndOfSentence" mwe="0" sem="Z99" lem="NULL">NULL</w>
```

Figure 4: FST sample output

4. Evaluation

During the Benedict Project, the Finnish semantic tagger has been evaluated on various test data. In our most recent evaluation, we examined the lexical coverage and precision of the tool on a pair of test corpora. The first test corpus was a random collection of texts with the topic of the past and present of Finnish cooking (4,264 words) while the second one was a random collection of articles from *Helsingin Sanomat*, the biggest Finnish daily newspaper (24,452 words).

First, we tested the lexical coverage of FST. As a result, the tool produced a lexical coverage of 94.1% on the Finnish cooking test data and 90.7% on the *Helsingin Sanomat* data, as shown in Table 2. Such a result is very promising, considering the fact that we have developed the resources from scratch over three years. This lexical coverage is comparable to that of the English tagger (Piao et al, 2004), which has been built and improved over more than a decade. Obviously, the different lexical coverages on different domains can be expected, which predicts varying degrees of performance of the tool across domains and genres. Although our evaluation is far from being conclusive, it demonstrates that the current FST, as it stands, already provides a practically useful tool in terms of lexical coverage.

Next, we examined precision of the tool on a subset of the Finnish cooking test data (3,044 words). In details, after tagging the test data using FST, a Finnish linguist (the first author of this paper) manually checked the output. The precision was calculated as the percentage of the tags manually checked to be correct out of the total number of tags in the automatically tagged text. As we explained earlier, a Finnish compound is split into two constituent parts by FST, each receiving a separate tag. If any one of the constituent tags is incorrect, the compound as a whole is considered to be incorrectly tagged. But because a compound is a single orthographical unit, it is counted as one error. On the other hand, if a MWE is incorrect tagged, each of its constituent words is counted as one error. As a result a total of 515 errors were found, resulting into a precision of

((3044–515)÷3044)×100%=83.03%. While this is an encouraging result for a prototype tool, the moderate precision shows that it is a tough challenge to build an accurate semantic tagger for Finnish.

Test Corpus	Total Tokens	Unmatched Tokens	Lexical Coverage
Finnish cooking corpus	4,264	231	94.1%
Helsingin Sanomat newspaper corpus	24,452	2,278	90.7%

Table 2: Evaluation of lexical coverage

5. Error analysis and discussion

As our evaluation reveals, the lexical coverage of FST has already reached a high level, and it will be further improved by expanding the lexicon into wider domains. The real challenge, however, lies in developing disambiguation methods for the FST. In order to gain a deeper insight into the existing problems, we classified and analysed the errors occurred in our evaluation. Table 3 lists ten error types identified in the evaluation in descending order in terms of the number of errors.

No.	Error type	Number of errors	Percentage(%)
1	missing words	200	39
2	wrong order of senses	147	29
3	errors caused by the TextMorfo	45	9
4	auxiliary verbs	42	8
5	errors caused by the compound engine	30	6
6	missing multi-word expression templates	21	4
7	missing senses	12	2
8	errors caused by multi-word-expression templates	11	2
9	spelling errors in the test text	5	1
10	wrong tag for an ordinal number	2	0
11	**Total**	**515**	**100**

Table 3: Ten FST error types

With respect to the nature of errors, these error types can be divided into three major categories: a) lexicon related errors, b) disambiguation related errors, and c) errors caused by factors outside FST. Of them, the first error category can be resolved by expanding and improving FST lexicons while the second category can be solved by developing and improving disambiguation algorithms. As to the third category (see error type 9), it is related to the quality of the test data, and thus beyond the scope of our discussion in this paper.

In table 3, the error types of (1) missing words, (6) missing multi-word expression templates and (7) missing senses are caused due to the incomplete coverage of the lexical resource. Put together, they account for 46.21% of the total errors. To reduce such errors, the FST lexicon needs to be improved. For example, to reduce the errors of type (1), the lexicon needs to be expanded to cover a wider vocabulary; to solve problems of type (6), the MWE templates need to be expanded and improved to capture those expressions.

On the other hand, the error types (2), (3), (4), (5), (8) and (10) belong to the second category, making up 53.79% of the entire errors. Such errors are mainly related to the mis-performance or lack of the disambiguation components. For instance, the error type (2*)* is caused due to the lack of proper semantic disambiguation algorithms while the error type (3) is caused by the mis-performance of the TextMorfo tool. To reduce errors in this category, we need to develop and improve disambiguation algorithms and components for identifying the correct semantic category for a given word from a list of candidates based on context information.

Regarding the error distribution, the two major error types are (1) *missing words* and (2) *wrong order of senses*, containing 200 and 147 errors respectively. Put together, these two types account for 68% of the total errors. They point to the priority tasks for FST improvement: continual expansion of the lexicon and developing disambiguation algorithms for finding true sense from candidates. To address the main error types identified in our evaluation, we propose the following solutions.

1) Continue to expand and improve the lexicons, including the MWE templates.
2) Develop efficient disambiguation components, including compiling context rules, improving MWE templates and applying context-based sense disambiguation algorithms.
3) Improve the accuracy of the TextMorfo tool, whose error is the third largest source of errors in our evaluation (see Table 3).
4) Add a component for recognising auxiliaries. In our evaluation, 42 errors were related to the auxiliary verb *olla* ('to be'). Currently the *olla* is always tagged with {A3+ Z5}; the correct tag is Z5.
5) Improve the compound component. Our evaluation proves it works effectively, but it also makes some unexpected mistakes.

First of all, by expanding and improving the FST lexicon resources, we will address about 47% of the errors (refer to error types 1, 6. 7, 8 and 10 in table 3). This involves enhancement of several aspects of the FST lexicon, including lexical coverage, adding missing senses to the lexicon entries, and improving MWE templates. In particular, the expanded lexicon should cover the majority of the core Finnish vocabulary in order to be practically useful.

An important task is to expand and improve the MWE templates. In our experiment, 11 our of the total 14 MWE's in the test data were missed. Such a result is not surprising, because so far we have devoted only a limited amount of time and efforts into the building of MWE templates. The current version of the FST MWE templates contain noun and verb phrases, proper names and true idioms. Due to the highly inflectional feature of Finnish language, the FST MWE component works differently from its English counterpart. The FST first analyses the words syntactically using TextMorfo to reduce them into their base forms. It then matches the basic forms of the MWE constituent words against the MWE templates. In our evaluation, three malfunctional MWE templates caused the 11 errors. Perhaps we need to reconsider the conception of MWE templates from the Finnish point of view and design an system that can reliably recognise Finnish MWE's. In future, it should also be possible for FST to recognise the MWE's with optional embedded elements.

A more challenging task is to develop efficient disambiguation algorithms and components. Currently the only disambiguation method implemented in the FST if the MWE templates. If a word is not a part of a MWE, the FST assigns it all the candidate tag(s) found in the lexicon. Where the word has only one tag (sense) in the lexicon, it is highly probable that the tag is the correct one.

However, there are many words that have more than one candidate tags in the lexicon. Quite often, the first tag in the candidate list is the correct one, since they are arranged in a usage frequency rank order. Nonetheless, in our evaluation such a practice produced 147 errors. Obviously, we need to develop more sophisticated algorithms of disambiguation. For example, for those words which form MWEs, we need to write templates to capture them. Ideally wildcard symbols should be used to allow a single template to match multiple MWE patterns, as is the case for EST (e.g. –*kg NNU N3.5). In addition, we need to develop context rules to mirror the EST equivalents (e.g. VB*[Z5] (R*n) (XX) (R*n) V*G*).

As we mentioned earlier, Finnish is a highly inflectional language and thus Finnish words can carry a very high information load with their rich affixation potential. Therefore, we could make use of information about e.g. verbal rection and valence patterns in disambiguation. For example, the verb *lainata* can possibly mean either 'to borrow' (A9+) or 'to lend' (A9-). At present the FST cannot disambiguate between these two senses, instead it assigns the tag (A9+) to this word indiscriminately. In fact, the verb *lainata* appears with different cases which provides clues for the correct sense in the given context. In the next development stage of the FST, we should pay a closer attention to the possible application of such information to the disambiguation.

Finally, the FST compound component needs improvement. Firstly, when the program processes a compound, it does not check the lexicon before passes it to the compound engine. It should be noted that a compound may have already been included in the single word lexicon, in which case it is more likely a lexicalised compound. Although in some cases the result produced in this way can still be considered correct (with transparent and ad hoc compounds), but for the lexically petrified compounds, it can cause errors. The correct algorithm should be that the program always checks an input word in the lexicon first – compound or not – before it feeds the word into the compound engine. If the word is found in the lexicon, it should be treated as a non-compound word. Secondly, the compound engine does not recognize inflected first parts of compounds (usually in the genitive case), tagging them with Z99 (unmatched). Although this still can produce a correct output for those ad hoc compounds, of which the latter parts convey the core meanings of the items, this is by no means an intelligent algorithm. The compound component will be improved to identify the inflectional variants of compound constituent parts. Thirdly, the Finnish compounds which consist of more than two constituent parts. Because the TextMorfo can only split a compound into two parts, currently FST can only identify the last constituent part of such a Finnish compound while often assigning Z99 (unrecognised item) to the remaining parts of the compound. To solve this problem, the TextMorfo needs to be improved.

We envisage that, with the improvements proposed above, FST will provide an efficient tool for semantic annotation. Such semantic taggers can have various applications for both corpus linguistics and practical NLP tasks. For example, during the Benedict project, a supplementary disambiguation system, called Domain Detection System (DDS), was developed for dictionary look-up purposes. It effectively tries to disambiguate a word by looking at its context. It assumes that domain specific words co-occur and thus assigns the most likely semantic tag for the word (Löfberg et al. 2004). Lessons learnt in the DDS development could also be valuable for improving the semantic taggers in future.

6. Conclusion

In this paper we have reported on the development of the Finnish semantic tagger to date. We have briefly discussed the challenges presented to us by the widely different grammatical systems between English and Finnish during the construction of the Finnish semantic lexicon and the development of the FST software. Furthermore, we have presented our evaluation of both the lexical coverage and the precision of the current FST, discussed the problems found in our evaluation, and proposed solutions to address these problems in the future development of the tool.

Overall, the results are promising and the current software functions reliably to a large extent. As we further expand the semantic lexical resources and improve disambiguation algorithms, the performance of the FST will be improved. Although FST was developed in this project mainly for the context-sensitive dictionary look-up for the Finnish language, the package of multilingual semantic taggers employing compatible semantic field frameworks and taxonomies across different languages, such as the EST/FST package, can have various potential applications in corpus linguistics, lexicography and language engineering, including discourse analysis, development of searching tools for translations in bilingual/multilingual dictionaries, and cross-language information retrieval.

References

Aunimo, L., Makkonen, J. and Kuuskoski, R. (2004) Cross-Language Question Answering for Finnish. *Proceedings of the Web Intelligence Symposium, Finnish Artificial Intelligence Conference*, September 2004, Vantaa, Finland.

Beale, A.D. (1987). Towards a Distributional Lexicon. In: R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: A Corpus-based Approach.* London: Longman, pp. 149 - 162.

Cheadle, M. and Gambäck, B. (2003) Robust semantic analysis for adaptive speech interfaces. In C. Stephanidis, (ed.) *Universal Access in HCI: Inclusive Design in the Information Society, volume 4*, pp. 685-689, Mahwah, New Jersey, June. Lawrence Erlbaum Associates.

Lagus, K., Airola, A. and Creutz, M. (2002) Data analysis of conceptual similarities of Finnish verbs. In *proceedings of CogSci 2002, 24th annual meeting of the Cognitive Science Society, August 2002, Fairfax, Virginia, USA*, pp. 566 - 571.

Löfberg, L., Archer, D., Piao, S., Rayson, P., McEnery, T., Varantola, K., Juntunen, J-P. (2003). Porting an English semantic tagger to the Finnish language. In *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16.* UCREL, Lancaster University, pp. 457 - 464.

Löfberg L, Juntunen J-P, Nykanen A, Varantola K, Rayson P, Archer D. (2004). Using a semantic tagger as dictionary search tool. In Williams G. and Vessier S. (eds.) *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004),* Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Volume I, pp. 127-134.

Makkonen, J., Ahonen-Myka, H., Salmenkivi, M. (2002) Applying semantic classes in event detection and tracking. In *Proceedings of the International Conference on Natural Language Processing (ICON'02),* Mumbai, India.

Mudraya, O., Piao, S.L., Löfberg, L., Rayson, P., Archer, D. (forthcoming). English-Russian-Finnish Cross-Language Comparison of Phrasal Verb Translation Equivalents. Accepted for presentation at *Phraseology 2005*, Louvain-la-Neuve, Belgium.

Piao, S. L., 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)*, 26-28 May 2004, Lisbon, Portugal, Volume II, pp. 499-502.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, 25th May 2004, Lisbon, Portugal, pp. 7-12.

Sawyer, P., Rayson, P., and Garside, R. (2002) REVERE: support for requirements synthesis from documents. *Information Systems Frontiers Journal*. 4 (3), Kluwer, Netherlands, pp. 343 - 353.

Sommerville, I. (2001) *Software Engineering* (6th Edition). Addison-Wesley

Varantola, K. (forthcoming). The contextual turn in learning to translate. In Bowker, L. (ed.) *Text-based Studies: Lexicography, Terminology, Translation. In honour of Ingrid Meyer.*

Appendix

The EST/FST semantic tagset is arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels. A tag can be composed of:

1. an upper case letter indicating general discourse field.
2. a digit indicating a first subdivision of the field.
3. (optionally) a decimal point followed by a further digit to indicate a finer subdivision.
4. (optionally) one or more ‘pluses’ or ‘minuses’ to indicate a positive or negative position on a semantic scale.
5. (optionally) a slash followed by a second tag to indicate clear double membership of categories or a compound.

Antonymity of conceptual classifications is indicated by +/- markers on tags; comparatives and superlatives receive double and triple +/- markers respectively. (For further details, see http://www.comp.lancs.ac.uk/ucrel/usas/.)
The following tags are used in the examples of this paper:

A1.1.1	General actions, making etc.
A14	Exclusivizers/particularizers
A2.1	Affect:- Modify, change
A3	Being
A6.2	Comparing:- Usual/unusual
A9	Getting and giving; possession
B1	Anatomy and physiology
B2	Health and disease
B5	Clothes and personal belongings
E4.1	Happy/sad
E6	Worry, concern, confident
F1	Food
I2.1	Business: Generally
I2.2	Business: Selling
L1	Life and living things
N1	Numbers
N3.5	Measurement: Weight
N4	Linear order
N5	Quantities
O3	Electricity and electrical equipment
O4.6	Temperature
Q1.2	Paper documents and writing
Q4.1	The Media:- Books
S1.2.1	Approachability and Friendliness
S4	Kin
S8	Helping/hindering
T1.3	Time: Period
T3	Time: Old, new and young; age
X2.2	Knowledge
X2.4	Investigate, examine, test, search
Y2	Information technology and computing
Z1	Personal names
Z2	Geographical names
Z3	Other proper names
Z4	Discourse Bin
Z5	Grammatical bin
Z6	Negative
Z7	If
Z99	Unmatched