Porting an English semantic tagger to the Finnish language

Laura Löfberg[1], Dawn Archer[2], Scott Piao[2], Paul Rayson[2],
Tony McEnery[2], Krista Varantola[1], Jukka-Pekka Juntunen[3]

[1]University of Tampere, Finland
[2]University of Lancaster, UK
[3]Kielikone Oy, Finland

Abstract
Semantic annotation is an important and challenging issue in corpus linguistics and language engineering. While such a tool is available for English in Lancaster (Wilson and Rayson 1993), few such tools have been reported for other languages. In a joint Benedict project funded by the European Community under the 'Information Society Technologies Programme', we have been working towards developing a Finnish semantic tagger that will parallel the existing English semantic tagger. The intention is to avoid building a completely new system but to bootstrap using the existing software and the largely hand-constructed English lexical resources. In this paper, we report on our work to date, which includes (i) a comparative study of the grammar of English and Finnish, (ii) the tagging of an English-Finnish parallel corpus, and (iii) the building of a Finnish lexicon using existing lexicons and software such as a Finnish-English-Finnish machine-translation system, Finnish dependency parser and morphological analyser, etc. This paper also discusses some challenging issues that have arisen during the construction of the parallel semantic tagging system between English and Finnish, namely, the complications caused by the widely different grammatical systems of the two languages. We believe that our work will provide a valuable experience for the community working on cross-language annotation schemes.

1. Introduction

In the past decades, numerous tools have been developed for annotating linguistic information in corpora (i.e. part-of-speech taggers, syntactic parsers, etc.). Fewer tools have been developed to undertake semantic annotation, that is, the automatic assignment of semantic categories to words in a running text.[1] One such tool is the English Semantic Tagger (henceforth EST) developed at Lancaster University (Wilson and Rayson 1993).

The EST employs a set of semantic tags loosely based on Tom McArthur's (1981) *Longman Lexicon of Contemporary English*. That said, the tagset has been considerably revised in the light of practical tagging problems met in the course of previous research. The revised tagset is arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels. The following table shows the 21 labels at the top level of the hierarchy.[2]

Table 1: The top level of the USAS system

A general and abstract terms	B the body and the individual	C arts and crafts	E emotion
F Food and farming	G government and public	H architecture, housing and the home	I money and commerce in industry
K entertainment, sports and games	L life and living things	M movement, location, travel and transport	N numbers and measurement
O substances, materials, objects and equipment	P Education	Q language and communication	S social actions, states and processes
T Time	W World and environment	X psychological actions, states and processes	Y science and technology
Z Names and grammar			

Being a hybrid system, the EST combines lexicons and various disambiguation template rules. The entries of the lexicons provide possible semantic categories for words or multi-word units (MWU) – mostly idioms and fixed expressions. Given a key word or MWU, the disambiguation algorithm matches near contexts against template rules in

[1] Such tools are important for corpus linguistics and language engineering, and can potentially be applied to a wide range of corpus-based studies and practical NLP tasks.
[2] For the full tagset, see http://www.comp.lancs.ac.uk/ucrel/usas/.

order to determine correct semantic categories for the given key word/MWU. The EST is reported to have obtained 92% accuracy on general English texts (Rayson and Wilson 1996).

We have been working on improving the EST and constructing an equivalent Finnish Semantic Tagger (henceforth FST) as part of the Benedict project (the general aim of which is to promote the use of language tools in practical application).[3] In order to avoid re-inventing the wheel, the architecture of the FST is being designed so that it will parallel that of the EST. In other words, we envisage that a single program package (of parallel lexicons and rule sets) will be able to process both English and Finnish (by switching linguistic components).

We cannot construct a Finnish lexicon for the FST without first investigating the similarities and differences between the lexical structure of Finnish and the lexical structure of English, of course. A focus of this paper, then, is to highlight some of the grammatical features of Finnish in comparison to English. In addition, we discuss the proposed design of the Finnish lexicon, in particular, our intention of using existing lexicons and software to build the Finnish lexicon, before highlighting some of the practical issues that have arisen during the construction of the lexicon.

2. Distinct lexical features of Finnish language

Finnish is a Finno-Ugric language belonging to the Uralian language family, and is spoken mainly in Finland and by the people of Finnish origin living in Sweden and other countries. It uses the Latin alphabet set similar to the English alphabet with three exceptions: Å, Ä and Ö (the phonetic values of the letters are those of the International Phonetic Alphabet). [4] However, unlike English spelling, Finnish spelling is mostly phonetic, i.e. in principle, Finnish words are pronounced as they are written.

Finish is also very distinct from English in terms of its grammatical features. Indeed, English is by large an analytic language whereas Finnish is an agglutinative, synthetic language. Since we aim to build the FST lexicon to parallel that of the EST, the following sections compare some of the lexical features of the two languages in more detail.

2.1. Heavy morphological affixation

Due to its agglutinative nature, the Finnish language relies heavily on morphology. Generally, what is expressed in English through syntactic structure is expressed via morphological affixation in Finnish. For example, prepositions are used to express relations between words in English. In Finnish, case endings are used to express the same relations (the case endings are attached to the base forms of the words). Finnish also uses morphemes to express plural and possessive relations (the morphemes are added to the stems of nouns as suffixes), and to denote morpho-syntactic concepts pertaining to verbs, including tense, subject-conformant verb inflection, verb nominalization, interrogative form and some pragmatic verb variants. Such inflectional/derivational changes can also convey information about mood (indicative, conditional, potential, imperative) and voice (active, passive).

The flexible inflectional/derivational morphological changes and the high number of morphemes enable Finnish nouns, verbs and adjectives to carry a very high information load. Indeed, as the transliteration of the following example illustrates, a single noun can carry a meaning equivalent to an English phrase (in this case, *also in my plants*).

kasve/i/ssa/ni/kin [= base nominative form + plural marker + inessive case ending + possessive affix + clitic affix]
plant- s-in-my-also[= *also in my plants*]

To a non-Finnish speaker it means that Finnish words often look very long and unanalysable. This impression is further enhanced by the Finnish tradition of spelling N+N+ (Nn) compounds as single words. However, Finnish is not as incomprehensible as it seems, not least because it is extremely systematic and regular. Indeed, the morphology and sound changes of the Finnish language are governed by clear rules which can be operationalized. Automatic tools for analysing Finnish morphological units are therefore highly accurate and reliable.

2.2. Finnish in a nutshell

(1) *Articles*

While articles are one of the most frequent word classes in English, they are not used in Finnish. This is because determinedness can be expressed via case variation. It is worth noting that pronouns such as *se* ('it'), *joku* ('some'), *eräs* ('one') and the numeral *yksi* ('one') occasionally behave in an article-like manner. Unlike its English equivalent, the third person pronoun *hän* does not distinguish between genders (cf. 'he' and 'she').

[3] The partners of the project include language technology provider Kielikone, the Universities of Lancaster and Tampere, the publishing houses HarperCollins and Gummerus, and Nokia. For more information, see http://mot.kielikone.fi/benedict/.

[4] The first letter Å corresponds to the Finnish O sound and does not appear in Finnish words, but it occurs in proper nouns of Swedish origin which are common in Finland.

(2) *Negating the verb*
When the predicate verb is negated, the negation *ei* ('not') takes on the conjugation form that indicates person. The verb form is identical with the indicative form, e.g. *minä ajan -minä en aja* ('I drive' - 'I don't drive'), *me ajamme - me emme aja* ('We drive' - 'we don't drive').

(3) *Prepositions and postpositions*
As mentioned previously, Finnish uses case endings where English uses prepositions. Unlike English, Finnish also uses postpositions. Typically, they express relational aspects, i.e. *jälkeen* ('after') and *päällä* ('on').

(4) *Case*
In theory, Finnish nouns, adjectives and numerals may appear in 15 different cases. However, not all of them are likely to be found in real-life use, even though they are grammatically valid. Table 1 shows examples of uses of different cases in Finnish and their approximate equivalents in English (Karlsson 1982).

Table 2: Examples of cases of Finnish language

Finnish singular numeral, adjective and noun	Case	Approximate English equivalents
yksi punainen omena	nominative	'one red apple'
yhde/n punaise/n omena/n	genitive	'one red apple's
punainen omena/ *yhd/en punais/en omena/n*	accusative (the object form which is identical with the nominative or the genitive form)	(I can see) 'a red apple' (I ate) 'the red apple'
yhte/nä punaise/na omena/na	essive	'as one red apple'
punais/ta omena/a	partitive (object form)	(I was eating) 'a red apple'
punaiseksi omenaksi	translative	(It will turn) 'into a red apple'
yhde/ssä punaise/ssa omena/ssa	inessive	'in one red apple'
yhde/stä punaise/sta omena/sta	elative	'from one red apple'
yhte/en punaise/en omena/an	illative	'in/into one red apple'
yhde/llä punaise/lla omena/lla	adessive	'on one red apple'
yhde/ltä punaise/lta omena/lta	ablative	'from one red apple'
yhde/lle punaise/lle omena/lle	allative	'onto one red apple'

(5) *Compound words*
Compounds are stems consisting of more than one root, such as 'car-wash', 'black market', etc (Jackson and Amvela 2000: 79). Some English compounds have spaces between element roots. However, in this particular case, we use the term compound to refer to those formed by concatenating two or more words without a blank space between them. In Finnish, compounding is a very productive means of word-formation. Compounds are formed mostly of nouns, but words of other parts of speech can also appear in compounds. In theory, it is possible to combine any number of N+N+ (Nn) to form a compound. In reality, nevertheless, most compounds consist of an N+N combination in which the modifying N is either in the nominative form or in the genitive form.

Lexicalised compounds may have meanings which differ from sum of the meanings of the element words. Such compounds are usually included in dictionaries with their own definitions. On the other hand, the compounds of ad hoc formation have meanings that can be deduced from that of the element words. In this sense it is very similar to the English the N of N structure with the difference that in ad hoc N+N compounds the semantic relation between the modifier and the head is not explicit but has to be deduced with extra-linguistic knowledge, e.g. *saunakahvit* (sauna coffees) - coffee and cakes and/or savouries offered after bathing in the sauna has ended (Karlsson 1982).

Table 3: Examples of Finnish compounds

Finnish compound	Constituents		English equivalents
osa/päivä/työ	*osa* *päivä* *työ*	'part' 'day' 'work'	'part-time job'
sähkö/paimen	*sähkö* *paimen*	'electricity' 'shepherd'	'electric cattle fence'
auringon/keltainen	*aurinko+ the genitive-forming morpheme* *keltainen*	'sun' 'yellow'	'yellow like the sun'

(6) *Flexible word order*
As in any agglutinative language, the word order in Finnish is relatively flexible, but not random (Vilkuna 2000: 32). This is because the information about part-of-speech and syntactic function of a word is usually embedded in its inflectional/derivational pattern in text context. It is therefore possible to change word order without changing the core meaning of the sentence. Nonetheless, changing word order inevitably affects the thematic structure of the clause, resulting into new emphases and nuances.

In Finnish new information is usually given at the end of the sentence, but it can be topicalized by moving it to the beginning of the sentence. For instance, the words in the sentence: *Nina rakastaa oliiveja* (‘Nina loves olives') can be put in six different orders:

Nina rakastaa oliiveja ('Nina loves olives');
Oliiveja Nina rakastaa ('Olives Nina loves');
Rakastaa oliiveja Nina ('loves olives Nina');
Nina oliiveja rakastaa ('Nina olives loves');
Rakastaa Nina oliiveja ('loves Nina olives');
Oliiveja rakastaa Nina ('olives loves Nina').

All of the above sentence variants are grammatically correct, but only the first one is in the unmarked, natural word order.

As shown in our previous contrastive study, the Finnish language has a number of distinct grammatical features widely different from English. Such distinction presents a challenge to us in porting the EST framework into the Finnish language. In the following section, we report our current progress in this work and discuss related challenging issues.

3. Building Finnish lexicons for the FST
Note that the purpose of our work is to build a Finnish semantic tagger that makes use of the existing architecture of the EST. It would be ideal if we could simply switch the English linguistic database of the EST with Finnish counterparts. Unfortunately, things are not that straightforward. Indeed, our research to date suggests that changes to the EST framework are necessary if we are to deal with the distinct grammatical features of Finnish. The changes needed range from simple tasks such as dealing with three unique letters of Finnish (Å, Ä and Ö) to complex tasks such as adjusting tagsets to reflect dictinct syntactic features of the Finnish language. In the following sections, we will discuss the problems that need be solved for developing the FST.

3.1. Tackling the rich Finnish morphology
In the semantic tagger, part-of-speech information provides a basis for determining the semantic category of a word. In the Lancaster EST, the C7 tagset of the CLAWS tagger is used for this purpose. In order to develop the FST, we need a Finnish counterpart POS tagset and tagger. The C7 tagset or other English POS tagsets are not appropriate for the Finnish language without modification.

Moreover, considering the complex morphological structure of Finnish words, a specially designed morphological analyser is needed to identify stems and affixes embedded in Finnish words. Furthermore, in order to provide the syntactic tags, the Finnish texts need to be syntactically processed with a Finnish parser. For example, *Ajoimmeko?* (='Did we drive?') comprises a full sentence by itself. This word is derived from the base form *ajaa* ('to drive'). In order to analyse its morphological structure, we need several tags to annotate attributes. By way of illustration:

the category of the word is *verb*,

it is a *finite* form,
it is in the *active* voice (as opposed to the passive voice),
it is in the *indicative* mood (as opposed to the conditional, potential mood, etc.),
it is in the *past tense*,
it is in the *1st person plural* form (we),
it ends with the clitic *-ko* which implies a question.

As previously highlighted, a single Finnish word form carries the same information as an English clause. It is worth noting that a single Finnish verb may have thousands of different forms, many of which are semantically quite different from each other. We therefore need some kind of a method to separate these words syntactically, i.e. a different 'tag' for all. Fortunately, software modules designed for this purpose are available.

3.2. Morfo, TextMorfo and Dependency Parser

A software package designed by the Finnish language technology company, Kielikone Ltd, perfectly serves the purpose of analysing POS and morphological structure of Finnish words. This package includes three tools capable of analysing Finnish text in various aspects (Juntunen 2002).

The first tool, called Morfo, analyses the morphological structure of Finnish words. Given a word or compound, Morfo extracts all morpho-syntactic information from it and returns the candidate base forms with syntactic categories. It also splits the compounds into component elements.

The second tool, called DCParser, is a full dependency parser of the Finnish language. Given a raw Finnish sentence, it returns a dependency tree representing the dependency structure of the input sentence. The parser is deterministic and linear in time behaviour (Arnola 1998).

The last tool, called TextMorfo, works on the output of the DCParser to disambiguate word forms. That is, based on the candidate interpretations of the input word produced by the Morfo tool, it selects the correct interpretation in the given context. TextMorfo also extracts nominal and other constituent phrases. For example, given the word *Ajoimmeko*, the TextMorfo returns:

Ajaa (Ajoimmeko), Verb, Imp Act Ind P 1P ko

When necessary, this entry can be converted into an input format of the EST-core engines.

The TextMorfo is an efficient tool that can potentially be used as a Finnish equivalent to the English POS tagger and lemmatiser. This kind of software makes it possible and feasible to build a Finnish semantic tagger.

3.3. Tackling the compounds

There are a large number of compounds in the Finnish language, as is the case for most languages. As mentioned earlier, lexicalised compounds carry meanings which are different from the sum of the meanings of their elements, and they function syntactically as single words. Although there is an argument for including such compounds in the lexicon, we have found that it may not be feasible to compile an extensive compound lexicon. The main obstacle to such a lexicon is the unlimited number of potential compounds. For example, *hätäpuhelin* ('emergency phone'), *yleisöpuhelin* ('public phone'), *käsipuhelin* ('portable phone') etc. are all types of phones and should automatically get the semantic tag of a phone. Such new words are being coined frequently; therefore it is doubtful that anyone can ever collect these compounds exhaustively.

We suggest that a practical and reasonable solution to this problem is to break the Finnish word into element words. Each word will be assigned its own tag. At the final stage, multiple membership tags will be formed automatically by combining the tags of the compound parts. For example, N+N+ (Nn) compounds may receive two, three or even more tags, as the sample output shows below:

*suodatusaika*_NNT1_T1/A1.1.1 ('filtering_A1.1.1 time_T1')
*kasvukorkeus*_N3.7/N3.2+/A2.1 ('growing_N3.2/A2.1 height_N3.7')

In case of a lexicalised compound which has its own entry in a dictionary, the definition provided by the dictionary will be used instead of going through the disintegrating/integrating procedure.

Besides the types of compounds discussed so far, there is another type of compound whose element words cannot be separated. The reason is that when the element words are separated, they have different meanings from their combined form. A typical example of such a compound is the word *jälki-* which is often translated as the prefix 'post-' in English. However, when used on its own, the word jälki often means 'imprint'. 'track', 'trace' or 'mark'. In Finnish, this is a more general phenomenon than in English, though.

Considering the complicated structure of Finnish words, we suggest that a specially designed algorithm is needed to ensure that the program is able to pick the correct semantic tag for each part of a compound. For example, a

compound could be split into its elements, then each of the element could be given a special syntactic attribute like 'CompPart' before processing the lexicon. In this way, the rule

jälki_CompPart XXXX

would fire only when the word *jälki* occurred in a compound. In the future, a set of generic rules in the MWU lexicon will be needed for analysing parts of compounds.

3.4. Tackling the flexible word order

Like English, Finnish has plenty of fixed phrases and idioms, such as

mennä kalaan ('to go fishing' or literally 'go into a fish')
olla marjassa ('to be berry picking', literally 'be in a berry')

However, the word order in Finnish is very flexible. For example, one could say:

Marjaan mentyäni huomasin, että olin jättänyt korin kotiin.
('After having gone berry-picking I realized that I had left the basket at home.')

As in English, there might be several words between the actual idiom components. For example,

Olisi kiva mennä joskus sinunkin kanssasi marjaan.
('It would be nice to go some day with you berry-picking')

In order to identify such non-consecutive components of single idioms, we need to design a special algorithm. A brute force solution would be to list every idiom in every possible formation, but it would require an excessive amount of time and human-labour. One of the ultimate solutions for this problem would be to include processing dependency trees instead of linear surface form text streams. In the dependency trees approach, the tree topology and attributes of the nodes specify the grammatical relationships between words.

In addition to the main lexicon, we will also consider attributes-based rules such as:

*_Base:ajaa_* *_Base:auto_Case:{ Nom/Gen/Acc/Part/Iness/Adess} M3

where the items separated by '/' indicate possible cases. This rule would be able to tag all different ways to drive a car with semantic tag M3. However, it would exclude, for example, the act of driving someone out.

3.5. Construction of Finnish lexicon

The construction of a Finnish lexicon is probably the most laborious part of the porting project of the semantic tagger. To alleviate the problem, automatic methods/existing software will be used whenever possible. For instance, the Kielikone lexicons will be used for a rough translation of the existing EST lexicon. This technique has proved fruitful in the past where bilingual English-Polish dictionaries were used to assign semantic tags to Polish texts (Lewandowska-Tomaszczyk et al, forthcoming). Nevertheless, a time-consuming and meticulous manual post-editing phase is still needed.

3.5.1. Test corpus

Initial pilot tests include the construction of a parallel and comparable English and Finnish corpus. The Finnish corpus was compiled from texts found at http://www.kahvilasi.net/ - a Finnish web page for aficionados of good coffee. The collected Finnish texts were slightly amended (i.e. some grammatical and typographical errors were corrected), and then machine-translated into English and post-edited. Both the Finnish and English corpus were then compared and, where necessary, amended so that their content was as close as possible (in semantic terms). The Finnish corpus contains 2,063 words and the English corpus contains 3,473 words.

3.5.2. Initial lexicon experiments

In our initial experiment, we aimed to test the possibility of using a translated EST lexicon to tag Finnish text, as well as to construct a hand-tagged parallel corpus. The Finnish corpus was parsed with Kielikone's DC Parser to find all the base forms of the words. Altogether 1,489 base forms were extracted. They were arranged in alphabetical order, and English-Finnish translation lexicons were used to find the English equivalents for them. Finally, the resulting list of the English equivalents was tagged with the EST lexicon.

As can be expected, this method produced some noise to the lexicon: some Finnish words received more semantic tags than they should have. Extra tags were manually removed while missing tags were added to obtain a clean test lexicon for the subfield of *coffee*.

This lexicon was then applied back to the sample corpus. Afterwards, the corpus was manually checked to make sure that each word had only one correct semantic tag. This sample corpus will be used further for developing a statistical algorithm for the FST.

3.5.3. Future development of the lexicon

After having conducted initial experiments with different approaches, we conclude that existing lexicons and resources can help our work in porting the EST into the Finnish language. In terms of semantic categories, we found it is wisest to start with the most concrete, unambiguous semantic categories such as food, drink, animals, plants, anatomy and physiology, numbers and measurements etc. More abstract categories can be added later on. With abstract categories, automatic methods seem more a burden than help.

For those words which have a one-to-one translation entry in the Finnish-English lexicon, a fully automatic approach can be used, i.e. for a given Finnish word its semantic tag can be found in the EST lexicon via its English translation equivalent. Also, automatic methods can be used in subdomains where there are semantically tagged lexicons available, such as scientific names of animals and plants and other things. Finally, in order to guarantee a wide coverage of the lexicon, the translated and post-edited lexicon is compared against a Finnish-English dictionary or Finnish monolingual dictionary to find missing Finnish words.

Building idiom templates seems to be the most challenging part of our work. Little help from automatic tool can be expected. We have found it is unfeasible to translate all of the rules of the existing EST idiom list due to the regular expressions used within it. Possible help may come from the lexicons of the Kielikone's existing Finnish-English machine translation system TranSmart. This system has a quite extensive set of context-specific rules for translating Finnish to English and, to some extent, it does some semantic analysis, too. Unfortunately, the rules of TranSmart are in a format that is not compatible with the EST. However, efforts are being made to make use of this tool wherever possible.

4. Conclusion

In this paper, we described our research work on building a Finnish lexicon in the efforts of building a Finnish semantic tagger via porting the English semantic tagger architecture to the Finnish language. We examined distinct grammatical features of Finnish in comparison to English, discussed a list of challenging issues that have arisen in the course of our work, and proposed solutions to cope with these problems.

Our experiments have shown that different grammatical features between different languages may present tough challenges in building a system that is compatible with multiple languages. In our particular case, the distinct agglutinative and synthetic nature of Finnish and the analytic nature of English require different algorithms and tools for annotating the same type of linguistic information. In particular, the complex and flexible morphological structure of Finnish words entails a drastically different approach to extracting POS information. Our experiments to date have also shown that the semantic categories developed for the EST are mostly compatible with the semantic categorizations of objects and phenomena in Finnish and can thus be applied cross-linguistically in prototypical cases. Certainly, some minor deviations, such as the lack of gender markers in Finnish, will cause discrepancy between related categories of these two languages.

The unresolved issues mentioned above are both interesting and challenging to us. As our work progresses, we will seek solutions to these issues, which will be beneficiary to the corpus research community in general.

Acknowledgements

The work described in this paper was carried out within the Benedict Project funded by the European Community under the 'Information Society Technologies' Programme (project reference IST-2001-34237). The partners of the project are language technology provider Kielikone, the Universities of Lancaster and Tampere, the publishing houses HarperCollins and Gummerus, and Nokia. For more information, see the web site http://mot.kielikone.fi/benedict/.

References

Arnola, Harri 1998 On Parsing Binary Dependency Structures Deterministically in Linear Time. COLING 98. Montreal

Hakulinen, Auli; Karlsson, Fred 1979 *Nyky-suomen lauseoppia* ('Syntax of Contemporary Finnish'). Suomalaisen kirjallisuuden seura.

Jackson, Howard and Etienne Zé Amvela 2000 *Words, Meaning and Vocabulary - An introduction to modern English lexicology*, London: Casell.

Juntunen, Jukka-Pekka, Kielikone Ltd 2002 Features and API of Kielikone Language Processes for Automatic Processing of Finnish

Karlsson, Fred 1982. *Suomen peruskielioppi* ('Finnish Grammar'). Suomalaisen kirjallisuuden seura.

Lewandowska-Tomaszczyk B, Oakes M and Rayson P (forthcoming). Annotated Corpora for Assistance with English-Polish Translation. In Wilson A, Rayson P and McEnery T (eds.) *Corpora by the Lune: a festschrift for Geoffrey Leech*. Peter Lang, Frankfurt.
McArthur Tom 1981 *Longman Lexicon of Contemporary English*. Longman London.
Rayson Paul and Wilson Andrew 1996 The ACAMRIT semantic tagging system: progress report. In Evett L J and Rose T G (eds.) *Language Engineering for Document Analysis and Recognition, LEDAR, AISB96 Workshop proceedings*, pp. 13-20. Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK.
Vilkuna, Maria 2000. *Suomen lauseopin perusteet* ('The basics of Finnish Syntax'). Kotimaisten kielten tutkimuskeskus. Edita.
Wilson Andrew and Rayson Paul 1993 Automatic Content Analysis of Spoken Discourse. In: Souter C and Atwell E. (eds.), *Corpus Based Computational Linguistics*. Amsterdam: Rodopi. pp. 215-226