A Probabilistic Word Class Tagging Module Based On Surface Pattern Matching

Robert Eklund
Stockholm

Abstract

This paper[1] treats automatic, probabilistic tagging. First, residual, untagged, output from the lexical analyser SWETWOL[2] is described and discussed. A method of tagging residual output is proposed and implemented: the *left-stripping method.* This algorithm, employed by the module ENDTAG, recursively strips a word of its leftmost letter, and looks up the remaining 'ending' in a dictionary. If the ending is found, ENDTAG tags it according to the information found in the dictionary. If the ending is not found in the dictionary, a match is searched in ending lexica containing statistical information about word classes associated with the ending and the relative frequency of each word class. If a match is found in the ending lexica, the word is given graded tagging according to the statistical information in the ending lexica. If no match is found, the ending is stripped of what is now its left-most letter and is recursively searched in dictionary and ending lexica (in that order). The ending lexica – containing the statistical informaiton – employed in this paper are obtained from a reversed version of *Nusvensk Frekvensordbok* (Allén 1970), and contain endings of one to seven letters. Success rates for ENDTAG as a stand-alone module are presented.

1 Introduction

One problem with automatic tagging and lexical analysis is that they are never (as yet) 100 % accurate. Varying tagging algorithms, using different methods, arrive at success rates in the area of 94–99 %.[3] After machine analysis there remains an untagged residue, and the complete output may – somewhat roughly – be divided into three subgroups:

1 A group of unambiguously tagged words.

2 A group of homographs given alternative tags.

3 A residual group lacking tags.[4]

---

[1]This paper is an abbreviated version of my diploma paper in computational linguistics with the same title, presented in April 1993 at the department of linguistics, computational linguistics, Stockholm University.

[2]Karlsson 1990; Koskenniemi 1983a,b; Pitkänen 1992.

[3]See e.g. Church (1988), Garside (1987), DeRose (1988).

[4]There is a bulk of words which is never found in this group, preponderatingly those belonging to the closed words classes, since these normally are found in the lexicon.

Proceedings of NODALIDA 1993, pages 83-95

Whereas the second of these groups is treated in Eriksson (1992), the task undertaken in this paper is to develop an algorithm for tagging material which has come through lexical analysis untagged.

The paper falls into the following areas:

First, untagged, residual output from the lexical analyser SWETWOL (Karlsson 1990; Koskenniemi 1983a,b; Pitkänen 1992) is described and analysed. This is done in order to pin down what input is, in one way or another, problematic to an automatic tagger. This is covered in section 3.

This paper presents a probabilistic tagger – henceforth ENDTAG – which tags according to statistics on the relations between final-letter combinations and word classes. The statistical information was obtained from the listings in NFO (Allén 1970) and collected in special ending lexica. This is described in section 4.

The ENDTAG module is presented in section 5. ENDTAG is based on what is here called the *left-stripping algorithm*, which recursively strips a word from its leftmost letter and compares the remaining ending[1] with the statistical information in ending lexica described in section 4.

The results of ENDTAG are evaluated in section 6.

2 Method

The untagged material used in this paper consists of residual files from the lexical analyser SWETWOL in Helsinki. SWETWOL was run on 831.289 words, whereof 10.988 came out untagged. Since SWETWOL yields output files of words on a word-for-word basis – thus ignoring (more or less) things like lexicalised phrases, particle verbs (ubiquitous in Swedish) and the like, words were only analysed one-by-one. A conjectural supposition is that a higher rate of accuracy is to be expected if context is also considered, as attempts with purely heuristic parsers show (cf. Källgren 1991b;c, Brodda 1983). On the other hand, it can be argued that there is palpable explanatory value in trying to find out how much information can be extracted from the words alone, neglecting their immediately adjacent 'text-mates'.

The success rate of any automatic tagger or analyser, *per se* and in comparison with other automatic taggers, is of course dependent on what tagset is being employed. The more general it is, i.e., the fewer the tags,

[1]The word 'ending' will throughout this paper denote any word final letter cluster, be this a grammatical suffix or not.

the more 'accurate' the output will be, due to the lack of more subtle subcategories. Since it was judged important that the tagset easily harmonise with already existing tagsets employed in other systems, the ending statistics were obtained from *Nusvensk Frekvensordbok*, NFO hereinafter (Allén 1970). I opted to adhere to the tagset employed therein, thus, the tags employed in this paper constitute a proper subset of the NFO tags.. It should be pointed out that NFO also contains tags for subcategories. The tagset employed by ENDTAG is shown in table 1.

TABLE 1 : The tagset employed by the ENDTAG module.

ABBREVIATION	WORD CLASS
ab	adverb
al	article
an	abbreviation
av	adjective
ie	infinitival marker
in	interjection
kn	conjunction
nl	numeral
nn	noun
pm	proper noun (proprium)
pn	pronoun
pp	preposition
vb	verb
**	non-Swedish unit
NT[1]	*Not tagged in NFO*

The ENDTAG module was implemented in COMMON LISP.

3 A Curt Description of the Untagged Output

In order to pin down what needs to be accounted for in tagging algorithms for arriving at better figures, one naturally has to scrutinise, with as great a punctilio as possible, the contents of that residual group of untagged words. I will here briefly list just a few observations made.[2]

In the untagged material, **proper** and **place nouns** abound! This is not really surprising, since they do not to any greater extent exhibit consistent morphological patterns.[3] It is also hard to list them all in the lexicon. Liberman and Church (1992) mention that a list from the Donnelly marketing organisation 1987 contains 1.5 million proper nouns (covering

[1]Since it was found that not all words in the computer readable version of NFO were tagged, an additional tag was created to render the format consistent. Hence, the tag 'NT' was added.

[2]For a more detailed account, the reader is referred to Eklund 1993.

[3]Of course *some* consistent patterns can be found. Thus the suffix *-(s)son* in Swedish typically denotes a surname, as in *Eriksson*, *Svensson* etc.

72 million American households). Since these have any number of origins, it is not feasible to cover them either with morphological rules or with a lexicon.[1]

Abbreviations were also common, which is more surprising, since all these should, it is assumed, have been expanded/normalised in the pre-processing. A related – and harder – problem concerns lexical abbreviations and **acronyms**.

Compounds constitute a notorious problem in all automatic processing of Swedish. Because they are legion, compounds constitute a very dire problem for any tagging module working on Swedish text. It might even be hard to decide where the compound border is located.

A related problem is encountered in what I call **complex compounds**. By that I mean compound words created in ways diverging from the 'normal' compounding of two ordinary words. One example of this is when more than two words are compounded. Instances of such compounds are:

Djursholms-Bromma-Lidingö-gängen
'The Djursholm-Bromma-Lidingö gangs'
karpatisk-balkansk-bysantiska
'Carpatian-Balcanian-Bysanthinian'
du-och-jag-ensamma-i-världen
'you-and-I-alone-in-the-world'

These clearly exhibit a *word-hyphen-word* pattern which could be formalized thus:

X-(Y-)*Z

These, I assume, would normally obtain the correct tag if one just looked at **Z** alone, and tagged accordingly. Compounds like these, I have found, were rather common in psychological terminology, where it also typically was used rather freely as to word class. A 'word' such as *du-och-jag-ensamma-i-världen* may be used as an adjective or a noun, for example.

---

[1] Something that could be considered here is majuscule heuristics, but this is not done without problems since upper case letters appearing in texts might indicate a wide variety of different phenomena. For example, the first letter of each sentence in a typical text, Roman figures, initials, titles and headings etc. Because of these problems, I chose to let the algorithm exempt majuscules altogether. For further discussion on majuscules, cf. e.g. Libermann & Church (1992), Eeg-Olofsson (1991:IV *et passim*), Källgren (1991b) and Sampson (1991).

A similar problem concerns what I call **slash compounds** like:

Dannemora/Österby
Hornstein/Voristan

... where the slash (/) separates two words according to the formalised pattern:

X/Y

Other phenomena occuring amongst the untagged residue were **professional/special terms, diacritica, archaisms** and **numbers** in various forms.

Another rather amusing, problem is posed by a word like

aaaaahh!

This word is of a recursive disposition which could be formalised thus:

a+h+!+

... where the plus sign denotes any number, equal to or greater than one, and not necessarily the same number in all three instances.

A large part of the untagged output was made up of **foreign words**, expressions and quotations et cetera. Interestingly enough, some of the suffixes used in certain languages are sufficiently unambiguous to permit a graded tagging in Swedish. Thus, some endings of Latin origin, *-ium*, *-ukt* or *-tion*, and some endings of Greek origin, *-graf*, *-lit*, *-ark*, *-skop* or *-logi* are highly unambiguous as to word class.

A problem harder to solve is that of **new words** being continuously created, old words given new interpretations, and then being used as members of other word classes. Thus, even a word like the conjunction *but* can not be considered a sure-fire case. In a phrase like

'*But me no buts!!*'

... 'but' first occurs as a verb in its imperative form, and then as a noun in the plural.[1] One must also point out that *all* words, irrespective of word class, might be used as nouns in a meta-linguistic way, for instance:

---

[1] A Swedish, idiomatic, counterpart would perhaps be *Menna mig hit och menna mig dit!*, the story being a speaker annoyed with a listener who interrupts by saying *but* all the time!

A 'green' would suit this phrase better!
Thou employest too many a 'lest' in thy prolegomenon, young esquire!

Lexicalized phrases typically receive the wrong parses, especially if they allow other constituents to be included 'inside' them. Since, as mentioned before, the module works with but a one-word window, lexicalized phrases cannot be properly accounted for by the module.

4 Obtaining the Ending Lexica

If we are to tag on a probabilistic basis, we need statistical information on the ending/word class relation. Hence, the first task was to create a number of ending lexica containing information as to word classes associated with particular endings. As mentioned earlier, the ending lexica were obtained from the lists in NFO (Allén 1970). NFO is a listing based on one million running words obtained from the material PRESS-65 and exists in computer-readable format. It might be pointed out that NFO is based exclusively on newspaper texts, and that other types of texts would perchance result in different ending lists. (Then again, results always depend on the input material used.)

Ending lexica were created with endings of 1–7 letters[1] – one lexicon per ending length – and word classes and their relative frequencies were calculated. Thus, the final format is as follows:

("ENDING" ((WORD-CLASS$_1$ PERCENTAGE$_1$) (WORD-CLASS$_2$ PERCENTAGE$_2$) (WORD-CLASS$_n$ PERCENTAGE$_n$)))

Word class frequencies are given with four decimals, and the word classes appear in falling order according to frequency. Thus, an authentic typical lexicon entry (from the three-letter ending lexicon):

*("ari" (("nn" 0.7802) ("ab" 0.1209) ("pm" 0.0934) ("**" 0.0055)))*

In other words, if the three final letters of a Swedish words are *-ari*, then there is a 78 % probability that the word is a noun, a 12 % probability that it is an adverb, a 9 % probability that is a proper noun, and finally, a 0.5 % probability that it is a foreign word.

The output files of the ENDTAG module look exactly the same apart from the first member of the list which will be the entire word, instead of as above, a final letter cluster.

[1]The number seven was chosen without any reason in particular.

The number of entries for each of the ending lexica is shown in table 2.

TABLE 2 : Numbers of entries in the ending lexica obtained from NFO.

	Number of letters in each ending lexicon						
	one	two	three	four	five	six	seven
Number of entries in lexica	43	669	3 936	13 176	26 494	38 464	46 179

One thing which cannot be bypassed is the extent to which the number of word classes associated with an ending decreases with the number of letters in the ending, i.e., the longer the final letter cluster, the fewer word classes associated with that ending. Statistics showing these relationships are illustrated in table 3.

TABLE 3 : Number of word classes associated with number of letters in endings (percentages). Zero percent area is marked with bold line.

	Number of letters in ending						
Number of word classes	one letter	two letters	three letters	four letters	five letters	six letters	seven letters
one	30.2	39.5	52.9	71.4	85.4	92.1	94.9
two	23.0	13.8	20.9	19.4	12.3	7.2	4.7
three	–	12.1	12.9	6.2	1.8	0.6	0.3
four	23.0	9.6	6.2	2.0	0.3	0.1	0.1
five	4.7	7.3	3.3	0.7	0.1	–	–
six	7.0	5.2	2.1	0.2	–	–	–
seven	–	4.0	1.0	0.1	–	–	–
eight	9.3	3.7	0.5	–	–	–	–
nine	2.3	1.6	0.2	–	–	–	–
ten	4.7	1.5	0.1	–	–	–	–
eleven	11.6	0.7	–	–	–	–	–
twelve	9.3	0.1	–	–	–	–	–
thirteen	9.3	0.7	–	–	–	–	–
fourteen	4.7	–	–	–	–	–	–
fifteen	2.3	–	–	–	–	–	–

A detailed description of the contents of the ending lexica will not be given here, but one example will perhaps serve as an indicator as to how the module works. Table 4 shows that probability rises as a function of increasing length for three noun declensions in Swedish.

TABLE 4 : Noun percentages (plural/definite/genitive) for Swedish noun declensions one, two and three.

	Paradigm according to the pattern *o / a / e* + suffix (i.e. the three first noun declensions in Swedish).		
Declensions	*-r* (plural)	*-r n a* (plural+definite)	*-r n a s* (plural+definite +genitive)
First declension	74.8	96.9	100.0
Second declension	26.5	97.8	98.4
Third declension	41.5	94.9	97.6

5 A Description of the Left-Stripping Algorithm

The tagging problem has been approached by many a linguist in many a way. Morphological models of Swedish have been provided by Hammarberg (1966), Kiefer (1970), Linell (1972, 1976), Cedwall (1977), Hellberg (1978)[1], Brodda (1979), Blåberg (1984), Eeg-Olofsson (1991:III), Ejerhed and Bromley (1986) and others. These works however, predominantly treat either very specific areas of Swedish morphology with varying degrees of minutae, or are generative models for Swedish word formation.

Probabilistic parsing as such, has been described by e.g. Sampson (1991) and Church (1987). As for tagging, probabilistic/statistical methods in general have been used by e.g. Johansson and Jahr (1982), Marshall (1987), Garside and Leech (1982), Church (1987) and Garside (1987) in the tagging of the LOB Corpus. Eeg-Olofsson (1991:I;IV) describes a statistical model for word-class tagging, and DeRose (1988) treats grammatical disambiguation by means of statistical methods. Johansson and Jahr's project aimed at improving the suffix lists developed for the Brown Corpus by Greene and Rubin (1971). They basically worked by means of a prediction of word classes in relation to grammatical suffixes, and to a certain extent also prefixes. Ejerhed (1988), Karlsson (1990), Källgren (1991a;b), Magnberg (1991) and Eriksson (1992) employ probabilistic methods for lexical analysis. Recent methods have been proposed by Samuelsson (1994) and Cutting (1994).

[1]Implemented by Ivan Rankin (1986).

The algorithm presented in this paper – the left-stripping algorithm – works by simple surface structure pattern matching. The concept is to strip a word of its leftmost letter, look for the resulting 'word' – i.e., the previous word sans its first letter – in a dictionary (e.g. SWETWOL for Swedish). If it is found, the word is tagged according to the dictionary, and the procedure is repeated with the next word. If it is not found, and the number of letters in the word is small enough to have a corresponding ending lexicon, i.e., the same number of letters, the word is looked for in that ending lexicon. If it is found in the ending lexicon, it is tagged, and the whole procedure is repeated with the next word. If it is not found, the word is stripped of what is now its leftmost letter, searched for in the dictionary et cetera. If no match is found even at the final (one) letter stage, the word is tagged thus:

("ENDING" ((NONE 0.0)))

The rationale behind this somewhat pleonastic design of the word class list is a desire to keep the format consistent. The flow chart in FIGURE 1 describes the module.

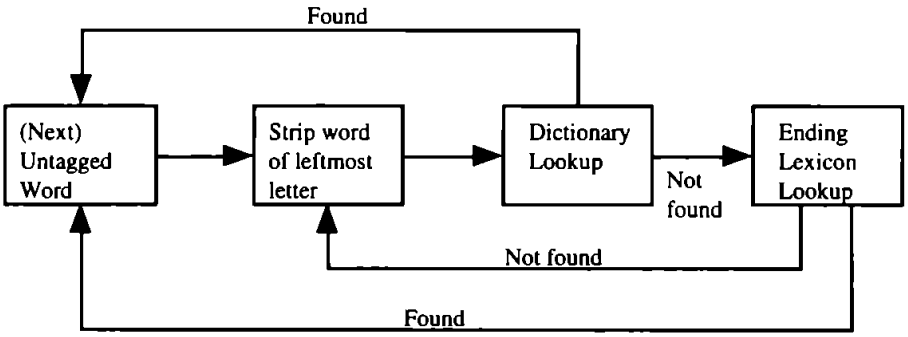

FIGURE 1 – Flow chart of the ENDTAG module.

As mentioned earlier, 'ending' here denotes the *n* final letters of a word, irrespective of whether these be grammatical suffixes, common combinations of any kind or unique word-final clusters. The dictionary lookup is likely to succeed before the ending lexicon, since the length of a complete word (normally) perforce exceeds the length of its ending.[1]

The module iterates over the untagged output list and strips the words recursively until a match is found in either the dictionary or the ending lexica. In the test run carried out here, no dictionary was employed, and the sub-routine intended to perform the dictionary lookup was foregone.

[1]In some instances, however, it might be hard to tell the difference between a word and its ending. Thus, in quoting John Lennon's *Give Peace A Chance*: *...Ragism, Tagism, / This-ism, that-ism, ism, ism.* ...it might be hard to tell the difference between the word and the ending in *ism*.

6 Results and Discussion

Since the output files of the module provide *graded* tagging, it is somewhat hard to discuss the results in terms of 'hits' or 'misses'. What could be discussed is how often the word class with the *highest* percentage is also the 'correct', word class. Although the module was not conceived as being used as a stand-alone module, it is of a certain interest to check its capabilities a such. Thus, a test run was carried out on 316.599 already tagged – and manually checked – words in the *Stockholm-Umeå Corpus* (Källgren 1991a). The leftmost member in the resulting output lists of ENDTAG were compared to the tags in SUC. The percentages are given in table 5.

TABLE 5 : Figures indicating the percentages of right tagging of words for different word classes.

WORD CLASS	PERCENTAGE
Infinitival marker	100
Nouns	93
Verbs	93
Prepositions	82
Adjectives	78
Adverbs	78
Conjunctions	69
Proper nouns	66
Pronouns	63
Interjections	37
Numerals	26
Abbreviations	16

One interesting feature of ending-list based tagging is the method's inherent capabilitites regarding the tagging of *new* words (cf. Greene & Rubin 1971). Since word formation obeys morphological rules, one may predict that neologisms and inflected loan words should be given rather accurate tags by the module.

One could also point out that one of the contributions of this work is the actual ending lexica *per se*. These have not been scrutinised in detail, but could presumably provide interesting information if studied.

Another point worth making is the module's limitations. Primo, it works on a *brute force* basis, rather than with linguistic *finesse*. The fact that it is not based on grammatical or morphological descriptions or models of Swedish, precludes generation, whence it follows that the module is not bi-directional, a lack we will have to make do with if we want to be able to handle foreign entries. Secundo, as already pointed out, the ending information in the ending lexica is perforce dependent upon the material on which they are based (in this case NFO). Tertio, tagging is graded. If an unambiguous tagging is desired, the module must succeed at lengths greater than (in most cases) three to four letters.

As a final remark, it could be said that no *one* tagging strategy, hitherto, has been able to solve this task fully. A combination of several different methods might increase success rates. A combination of a lexically based method (SWETWOL) with a statistically based method (ENDTAG), disambiguated by a module like the one described by Eriksson (1992) could enhance success rates in automatic word class recognition.

References

Allén, Sture. 1970. *Nusvensk frekvensordbok baserad på tidningstext (Frequency dictionary of present-day Swedish), 1. Graphic Words, Homograph Components, 2. Lemmas, 3. Collocations, 4. Morphemes, Meanings*. Almqvist & Wiksell International, Stockholm.

Blåberg, Olli. 1984. *Svensk Böjningsmorfologi. En tvånivåbeskrivning*. Unpublished Master's Thesis, Department of General Linguistics, University of Helsinki.

Brodda, Benny. 1979. *Något om de svenska ordens fonotax och morfotax: iakttagelser med utgångspunkt från experiment med automatisk morfologisk analys*. PILUS 38, December 1979, Stockholm University.

Brodda, Benny. 1983. *An Experiment with Heuristic Parsing of Swedish*. Stockholm University, Dept. of Linguistics.

Cedwall, M. 1977. *Semantisk analys av processbeskrivningar i naturligt språk*. Linköping Studies in SCIENCE AND TECHNOLOGY, Dissertations No. 18.

Church, Kenneth Ward. 1988. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text*. In *Proceedings of the Second Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pp. 136–143.

Cutting, Douglas. 1994. *Porting a Stochastic Part-of-Speech Tagger to Swedish*. In Eklund (ed), *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*. Stockholm.

DeRose, Steven. 1988. *Grammatical Category Disambiguation by Statistical Optimization*. In COMPUTATIONAL LINGUISTICS 14:1.

Eklund, Robert. 1994. (ed). *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*. Stockholm.

Eklund, Robert. 1993. *A Probabilistic Word Class Tagging Module Based On Surface Pattern Matching*. BA Thesis, Dept. of Linguistics, Stockholm University.

Eeg-Olofsson, Mats. 1991. *Word-Class Tagging, Some computational tools.* Ph. D. Thesis including the following papers: *i. A probability model for computer-aided word-class determination, ii. Software Systems for Computational Morhpology – An Overview, iii. A morphological Prolog system for Swedish based on analogies, iv. Probabilistic word-class tagging of a corpus of spoken English.* Gothenburg University, institutionen för språkvetenskaplig databehandling.

Ejerhed, Eva and Hank Bromley. 1986. *A Self-extending Lexicon: Description of a World Learning Program.* in Karlson 1986, pp. 59–70.

Ejerhed, Eva. 1988. *Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods.* In *Proceedings of the Second Conference in Applied Natural Language Processing*. Austin, Texas.

Eriksson, Gunnar. 1992. *Att två ord för att få en tolkning*. Dept. of Linguistics, Computational Linguistics, Stockholm University.

Garside, Roger. 1987. *The CLAWS word-tagging system.* in Garside, Leech and Sampson (1987), chapter 3.

Garside, J. L. and Geoffrey Leech. 1982. *Grammatical Tagging of the LOB Corpus: General Survey.* in Johansson 1982.

Garside, Roger, Geoffrey Leech and Geoffrey Sampson. 1987. *The Computational Analysis of English: a Corpus-Based Approach.* Longmans, London.

Greene, Barbara B. and Gerald Rubin. 1971. *Automatic Grammatical Tagging of English.* Providence, R. I., Department of English, Brown University.

Hammarberg, Björn. 1966. *Maskinell generering av böjningsformer och identifikation av ordklass*, pp. 59–70 in *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning*. Sture Allén (ed.), Gothenburg University.

Hellberg, Staffan. 1978. *The Morphology of Present-Day Swedish.* DATA LINGUISTICA 13, Sture Allén (ed.), Department of Computational Linguistics, University of Gothenburg, Almqvist & Wiksell International, Stockholm.

Johansson. Stig. 1982. *Computer Corpora in English Language Research.* Norwegian Computing Centre for the Humanities, Bergen.

Johansson, Stig and Mette-Cathrine Jahr. 1982. *Grammatical Tagging of the LOB Corpus: Predicting Word Class from Word Endings.* In Johansson 1982.

Karlsson, Fred. 1986. Papers from the Fifth Scandinavian Conference on Computational Linguistics. Helsinki, December 11–12, 1985. Publications No. 15, Department of General Linguistics, Helsinki University.

Karlsson, Fred. 1990. *Constraint Grammar as a Framework for Parsing Running Text.* In Hans Karlgren (ed.), *Papers presented to the 13th International Conference on Computational Linguistics*, vol. 3, pp. 168–173, University of Helsinki, Department of General Linguistics.

Karlsson, Fred. 1992. *SWETWOL: A Comprehensive Morphological Analyser for Swedish.* pp. 1–45, NORDIC JOURNAL OF LINGUISTICS, Volume 15, Number 1, Scandinavian University Press.

Kiefer, Ferenc. 1970. *Swedish Morphology*. Skriptor, Stockholm.

Koskenniemi, Kimmo. 1983a. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Department of General Linguistics, University of Helsinki, Publication No. 11.

Koskenniemi, Kimmo. 1983b. *Two-Level Morphology for Word-Form Recognition and Production*. Department of General Linguistics, University of Helsinki, Publication No. 13.

Källgren, Gunnel. 1991a. *Storskaligt korpusarbete på dator. En presentation av SUC-korpusen*. In SVENSKANS BESKRIVNING 18, Lund University Press.

Källgren, Gunnel. 1991b. *Making maximal use of surface criteria in large-scale parsing: the MorP parser*. In Pilus 60, Dept. of Linguistics, Stockholm University.

Källgren, Gunnel. 1991c. *Parsing without lexicon: the MorP System*. European Chapter of the Association for Computational Linguistics, Berlin.

Lennon, John. *Give Peace A Chance*. Northern Songs, 1969.

Liberman, Mark Y. and Kenneth W. Church. 1992. *Text Analysis and Word Pronunciation in Text-to-Speech Synthesis*. In S. Furui and M. M. Sundhi (eds.), *Advances in Speech Technology*, Marcel Dekker, New York.

Linell, Per. 1972. *Remarks on Swedish Morphology*. Ruul 1, Department of Linguistics, Uppsala University.

Linell, Per. 1976. *On the Structure of Morphological Operations*. In LINGUISTISCHE BERICHTE 44/76, pp. 1–29.

Magnberg, Sune. 1991. *A Rule-Based System for Identifying Major Syntactic Constituents in Swedish*. Department of Linguistics, Stockholm University.

Marshall, Ian. 1987. *Tag selection using probabilistic methods*. In Garside, Leech and Sampson (1987), chapter 4.

Pitkänen, Kari. 1992. *SWETWOL / Major Changes in 1992*. Research Unit for Computational Linguistics, University of Helsinki.

Rankin, Ivan. 1986. *SMORF – an Implementation of Hellberg's Morphology System*. Department of Computer and Information Science, Linköping University.

Sampson, Geoffrey. 1991. *Probabilistic Parsing*. In Svartvik, Jan (ed.), *Directions in Corpus Linguistics 82*, Stockholm, 4–8 August 1991, Mouton de Gruyter, Berlin.

Samuelsson, Christer. 1994. *Morphological Tagging Based Entirely on Bayesian Inference*. In Eklund (ed), *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3–5 June 1993*. Stockholm.