

# COMPUTATIONAL MORPHOSYNTAX

Report on Research 1981—84

Edited by Fred KARLSSON

University of Helsinki  
Department of General Linguistics  
Hallituskatu 11—13  
SF-00100 HELSINKI 10  
FINLAND

PUBLICATIONS  
No. 13  
1985

Fred Karlsson

## MORPHOLOGICAL TAGGING OF FINNISH

### 1. Introduction

Readily available facilities for testing hypotheses and descriptive alternatives on sufficiently large natural corpora are mandatory when one tries to design a parser for running text. Of course, tagged corpora have many other uses as well, especially as tools for "ordinary working grammarians".

Tagging a corpus normally means supplying its word tokens with appropriate labels of part of speech membership by using computerized methods for the bulk of the work. However, a time-consuming final phase of manual checking and correction is always needed. In practice corpus tagging is therefore semi-automatic.

Most of the tagging projects so far completed have indeed been predominantly morphological even if syntactic prerequisites have often been taken in the selection of what tags to use, cf. the structure of the Brown Corpus of written American English (Francis 1964, 1980), the London-Oslo-Bergen Corpus of written British English (Garside and Leech 1982), and the London-Lund Corpus of spoken English (Svartvik, Egg-Olofsson, Forsheden, Öreström and Thavenius 1982). Important steps towards semiautomatic syntactic tagging of the two last-mentioned corpora have already been taken (Leech, Garside and Atwell 1983, Atwell, Leech and Garside 1984, Svartvik & al. 1982, Egg-Olofsson and Svartvik 1984). Of course, syntactic tagging presents problems of a quite different magnitude than those of morphological tagging. A good syntactic tagging program is in fact more or less equivalent to a parser (Brodda 1982a). Both enterprises have as their ultimate goal to assign constituent, functional or other structure to clauses and sentences.

As pointed out in the introduction to this volume, mor-

phology, and consequently also morphological tagging, is of special importance in Finnish. A full surface morphological description of the word tokens in a text provides a wealth of clues relevant to grammatical and semantic interpretation. Much of this information corresponds to what in more analytic languages would be supplied by "syntax" in the classical sense. E.g., given knowledge of the case marking of the nominal constituents, and the likewise overtly marked finiteness of the main verb in a Finnish clause, it is possible to unambiguously assign grammatical functions (subject, object, etc.) to the nominals.

Mainly for serving the purpose of constructing parsing models we have designed a morphological tagging program, FIN-TAG, running in batch mode that consists of fifteen BETW rule modules geared to apply in sequence to any input text. The input needs no pre-editing. The output is the original text so analyzed that (a) all word tokens are supplied with a part of speech label, and (b) all inflectional endings and clitics are segmented, i.e. separated from the preceding stem and eventual following endings by a morphological juncture. FINTAG thus is an extended tagging program in that it also performs full-fledged morphological analysis of the input words. The tagging format is intagging in the sense of Brodda (1982b). Part of speech labels are prefixed (separated by a colon) to the word forms, and the morphological junctures are infixes as characters of the word token. Analyzed sentences thus look as follows.

<u>Tagged word</u>	<u>Interpretation</u>
Ni:vuode=n	noun, ending =n (gen.sg.)
Naiku	noun, no endings (nom.sg.)
Vi:ron	finite verb, no endings (3.p.sg.pres.ind.)
A:kylmä=a	adjective, ending =ä (part.sg.)
Naika=a	noun, ending =a (part.sg.)

"=" is the basic juncture marker. Two more junctures are in use: "g" for the illative case, and " for possessive clitics.

Their motivation is simply to facilitate accessing those endings which would otherwise be laborious due to the wealth of forms. All other ending morphemes may be accessed just by listing their allomorphs prefixed by "-".

Multiple tags are assigned if the word form is ambiguous, e.g. VE/N:tu=n. Ambiguous or uncertain morphological segmentations are started, e.g. V/VE/N:vo\*1 (there is a juncture only if the proper interpretation is VE:vo=1, i.e. if the word is a finite past tense form). The final program module disambiguates multiple tags and unclear segmentations using contextual information provided by the current clause, in most instances only the preceding or following word.

The notable thing about this simple tagging format is that the joint information of (i) part of speech and (ii) segmented endings suffices to provide unique morphological identification of almost all morphs. E.g. the ending -t in isolation is ambiguous between nom. pl. and 2nd pers. pres. indic. act., but knowledge of part of speech membership picks the right interpretation. As shown by Brodda and Karlsson (1981), there are very few instances where an ending morph is homographic within the same part of speech.

## 2. The Part of speech tags

FINTAG uses the following restricted tag set. The defining properties of the various parts of speech are mainly morphological.

<u>For nominals</u>	
NE	proper name
N	other noun
A	adjective (including ordinal numerals)
PR	pronoun
NUM	cardinal numeral

For verbs

VF	finite verb form (including "passives")
V	the negative stem
V11	1st infinitive
V12	2nd infinitive
V13	3rd infinitive
VP1	1st active participle
VP2	2nd active participle
VPA1	1st passive participle
VPA2	2nd passive participle
VPP1	1st passive participle
VPP2	2nd passive participle
VPG	agent participle
VPN	negative participle

For (mostly) uninflected words

ADV	adverb (in the morphological sense)
pp	preposition, postposition
C	conjunction
I	interjection
ABBR	abbreviation

Of course, there are many borderline cases between two or more of the above parts of speech. Such intricate language-specific matters will not be further pursued here.

Only 21 tags and 3 junctures are in use. The number seems small in comparison to the 81 and 134 tags of the Brown and LOB Corpora, respectively. The main difference is that the present tagging system does not aim at a detailed subclassification of the major word classes. The tag system could fairly easily be amended since the lexicons (cf. below) systematically contain the most frequent representatives of all parts of speech. On the other hand, a wealth of information is extractable from the explicitly indicated morphological structure.

3. The tagging strategies

The fifteen BEVA rule modules contain a total of some 16,500 lines of substitution rules (a closer presentation of Brodda's BETA formalism is available in e.g. Brodda and Karlsson 1981). Some of the modules check for endings, others are condensed frequency-based lexicons looking for roots or stems belonging to specific parts of speech. There is no strict order between lexical identification of stems and morphological segmentation of endings. The rule modules have been sequenced both on linguistic and heuristic grounds. These considerations determining the optimal order of tag assignment are here called tagging strategies. The most important tagging strategies are the following. Their order of invocation is, generally speaking, the one presented. The n in an expression M<sub>n</sub> refers to the sequence number of the respective module.

- whenever a word form is conclusively tagged and segmented, prefix a plus-sign to it indicating that later rule modules need not enter this word (concerns all modules); M13 removes all plus-signs;
- conclusively tag and segment all monosyllabic word-forms, inflected or not (M1); this is a finite list with some 250 members;
- mark (with a special diacritic) such first syllables that could potentially occur as isolated monosyllables (for the purpose of facilitating ending identification) (M1);
- segment eventual clitics of all parts of speech (M2); M2 also contains an optimized stop list (N=85) preventing errors such as \*et<sub>u</sub>s<sub>k</sub>i<sub>n</sub> p<sub>r</sub>o e<sub>t</sub>u<sub>s</sub>k<sub>i</sub>=n;
- tag and, if proper, conclusively segment the most common (N=some 600) adverbs, prepositions, postpositions, and conjunctions that would otherwise be wrongly analyzed by

- the productive inflectional rules for nouns (M3): e.g. the adverb YÄRSIN should not be analyzed as a putative instructive plural \*YÄRS=I=N;
- the top 200 word-forms of a frequency count (The Oulu corpus, cf. Saukkonen & al. 1979) are tagged and segmented as wholes (M4a);
- the possessive suffixes are identified by M5 which also contains a stop list (N=90) of potential structural homonyms such as aspitiini;
- (the stems of) closed lexical classes are identified by early lexical lists (partly mentioned above);
- the central inflectional (and a few derivational) endings of nouns, adjectives, pronouns, numerals and verbs are segmented in a strict order dispersed over M3-M13; the most frequent structural homonyms are always exempted from segmentation by stop lists; most verb endings reside in M6, most nominal endings in M9;
- when possible, predict the part of speech tags using previously segmented inflectional or derivational endings of the word token (e.g. EMÄINE and TON occur in adjectives only, YÄI in verbs /either finites or participles/ only, SSÄ in nominals and participles only, etc.);
- use (inconclusive) frequency-based stem-lexicons for tagging adjective and verb stems;
- by default, M13 predicts that all remaining untagged word tokens are nouns;
- M15 disambiguates multiple tags and uncertain ending segmentations.

A comparison with the English corpora in regard to tagging strategies is instructive. The Brown corpus tagging program employed look-up in a dictionary of 2,860 words. If the word was not found there, its ending was checked against a suffix list of 416 entries. 61% of the lexicon words and 51% of the listed suffixes provided unique tags (Francis 1980:201).

In comparison to Brown, the LOB project has enlarged both the lexicon (over 7,000 words) and the suffix-list (660 word-endings) (Leech, Garside and Atwell 1983:13).

The survey of Spoken English project employed a dynamic high-frequency lexicon containing some 1,000 items. For each entry the program maintained its cumulative total frequency as well as a list of all possible tags, each with its partial frequency. If a word token was encountered in the lexicon, it was assigned the most frequent tag. Additionally, a few heuristic suffix and context (frame) rules were used (Svartvik & al. 1982:66-68).

Since BETA rules correspond fairly directly both to entries in a lexicon and suffix stripping rules, we may conclude that PINTAG is roughly five times "larger" than the Brown tagging program, and even two times larger than LOB in terms of lexical and morphological coverage. Of course, this is a very crude measure which only gives a hint of how the programs relate to each other. The main difference between PINTAG and LOB in regard to tagging strategies concerns the dispersion of lexical and morphological recognition in PINTAG. This is only to be expected, given the different surface appearance of the languages. Lexical and morphological recognition feed each other in intricate ways.

Below we return to an evaluation of success rate (precision) and contextual disambiguation rules.

#### 4. Processing cycle

The following is an example of how a short text is processed through the rule modules. The outputs of all modules activated in tagging this particular text are included. Note that most inflected words may be processed by several modules. Boldface shows where a word received its final analysis. Underlining indicates that some aspect of a word was processed by the respective module but a conclusive analysis was not reached. M1 and M13 trivially affect all word tokens. Incorrect results are starred. In this particular example, no multiple tags happened to be assigned and therefore the main disambiguator module, M15, was not activated.

##### Output of M1

TÄZMÄN KO'KOELMAN KI'RJOITUKSE<sup>T</sup> O'VAT PA'RIA KO'LMEA LU'KUUN O'TTAMATTA SY'NTYNEET VII'DEN VII'ME VUOIDEN AI'KANA. E'RÄÄT NIIZSTÄ +VF:ON JU'LKISTETTU RA'DIOSSA, E'RÄÄT E'SITELMINÄ. AI'HEPIIRI +VF:ON VE'RRATEN KI'RJAVA: MU'KANA +VF:ON EONSINNÄ-KIN TIEDEPOLITIIKKAAN +C:JA TIEOTEENFILOSOFIAAN LIITTYVÄ KI'RJOITUKSIA.

##### Output of M2

TÄZMÄN KO'KOELMAN KI'RJOITUKSE<sup>T</sup> O'VAT PA'RIA KO'LMEA LU'KUUN O'TTAMATTA SY'NTYNEET VII'DEN VII'ME VUOIDEN AI'KANA. E'RÄÄT NIIZSTÄ +VF:ON JU'LKISTETTU LE'HDISTÖSSÄ, E'RÄÄT RA'DIOSSA, E'RÄÄT E'SITELMINÄ. AI'HEPIIRI +VF:ON VE'RRATEN KI'RJAVA. MU'KANA +VF:ON EONSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA TIEOTEENFILOSOFIAAN LIITTYVÄ KI'RJOITUKSIA.

##### Output of M3

TÄZMÄN KO'KOELMAN KI'RJOITUKSE<sup>T</sup> O'VAT PA'RIA KO'LMEA LU'KUUN O'TTAMATTA SY'NTYNEET VII'DEN VII'ME VUOIDEN AI'KANA. E'RÄÄT NIIZSTÄ +VF:ON JU'LKISTETTU LE'HDISTÖSSÄ, E'RÄÄT RA'DIOSSA, E'RÄÄT E'SITELMINÄ. AI'HEPIIRI +VF:ON +ADV:VERRATEN KI'RJAVA. +PP:MUKANA +VF:ON +ADV:ENSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA TIEOTEENFILOSOFIAAN LIITTYVÄ KI'RJOITUKSIA.

##### Output of M4

+PR:TÄMÄ=N KO'KOELMAN KI'RJOITUKSE<sup>T</sup> +VF:O=VAT PA'RIA KO'LMEA LU'KUUN O'TTAMATTA SY'NTYNEET VII'DEN +A:VIIME +N:VUODE=N +N:AIKA=NA. E'RÄÄT +PR:NI=I=STÄ +VF:ON JU'LKISTETTU LE'HDISTÖSSÄ, E'RÄÄT RA'DIOSSA, E'RÄÄT E'SITELMINÄ. AI'HEPIIRI +VF:ON +ADV:VERRATEN KI'RJAVA. +PP:MUKANA +VF:ON +ADV:ENSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA TIEOTEENFILOSOFIAAN LIITTYVÄ KI'RJOITUKSIA.

##### Output of M6

+PR:TÄMÄ=N KO'KOELMAN KI'RJOITUKSE<sup>T</sup> +VF:O=VAT PA'RIA KO'LMEA LU'KUUN +VI3:OTTA=MA=TTA +VPA2:SYNTY=NEE=T VII'DEN +A:VIIME +N:VUODE=N +N:AIKA=NA. E'RÄÄT +PR:NI=I=STÄ +VF:ON +VPP2:JULKISTE=TTU LE'HDISTÖSSÄ, E'RÄÄT RA'DIOSSA, E'RÄÄT E'SITELMINÄ. AI'HEPIIRI +VF:ON +ADV:VERRATEN +A:KIRJAVA: +PP:MUKANA +VF:ON +ADV:ENSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA \*TIEOTEENFILOSOFI=KIN LIITTY=V=I=Ä KI'RJOITUKSIA.

##### Output of M9

+PR:TÄMÄ=N KOKOELMA=N KIRJOITUKSE<sup>T</sup> +VF:O=VAT PARI=A KOLMEA LU'KUUN +VI3:OTTA=MA=TTA +VPA2:SYNTY=NEE=T VII'DE=N +A:VIIME +N:VUODE=N +N:AIKA=NA. ERÄÄ<sup>T</sup> +PR:NI=I=STÄ +VF:ON +VPP2:JULKISTE=TTU LEHDISTÖ=SSÄ, ERÄÄ<sup>T</sup> RADIO=SSÄ, ERÄÄ<sup>T</sup> ESITELMINÄ. AIHEPIIRI +VF:ON +ADV:VERRATEN +A:KIRJAVA: +PP:MUKANA +VF:ON +ADV:ENSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA \*TIEOTEENFILOSOFI=A\*AN +VPAL:LIITTY=V=I=Ä +N:KIRJOITUKS=I=A.

##### Output of M12

+PR:TÄMÄ=N KOKOELMA=N KIRJOITUKSE<sup>T</sup> +VF:O=VAT PARI=A +NUM:KOLME=A LUKUKUN +VI3:OTTA=MA=TTA +VPA2:SYNTY=NEE=T +NUM:VIIDE=N +A:VIIME +N:VUODE=N +N:AIKA=NA. +PR:ERÄÄ<sup>T</sup> +PR:NI=I=STÄ +VF:ON +VPP2:JULKISTE=TTU LEHDISTÖ=SSÄ, +PR:ERÄÄ<sup>T</sup> RADIO=SSA, +PR:ERÄÄ<sup>T</sup> ESITELMINÄ. AIHEPIIRI +VF:ON +ADV:VERRATEN +A:KIRJAVA: +PP:MUKANA +VF:ON +ADV:ENSINNÄ=KIN TIEDEPOLITIIKKAAN +C:JA \*TIEOTEENFILOSOFI=A\*AN +VPAL:LIITTY=V=I=Ä +N:KIRJOITUKS=I=A.

##### Output of M13 = Final output

PR:TÄMÄ=N N:KOKOELMA=N N:KIRJOITUKSE=T VF:O=VAT \*N:PARI=A

NUM:KOLME=A N:LUKU&UN V13:OTTA=MA=TTA VPA2:SYNTY=NEE=T  
 NUM:VIIDE=N A:VTIME N:VUODE=N N:AIKA=NA. PR:ERÄ=T PRNI=1-STÄ  
 VE:ON VPP2:JULKISTE=TU N:LEHDISTÖ=SSÄ, PR:ERÄ=T N:RADIO=SSA,  
 PR:ERÄ=T \*N:ESITELMINÄ. N:AIHEPIIRI +VF:ON +ADV:VERÄTÄN  
 +A:KIRJAVÄ +PP:NUKANA +VE:ON +ADV:ENSINÄ=KIN N:TIEDEPOLITIIK-  
 KA&AN +C:JA \*N:TIEHENPILOSOFI=A\*AN +VPAL:LIITTY=V=I=Ä +N:KIR-  
 JOTUKS=I=A.

Eight rule modules have been activated. It is clearly seen how most of the tagging load resides with M1 (monosyllables), M3 (adverbs, etc.), M4a (the frequent word forms), M6 (most verb endings), M9 (nominal case endings), M12 (certain pronominal, numeral, and adjectival stems), and M13 (by default, nouns).

The example cycle produced three errors: N:pari=a (wrong part of speech, N pro NUM), N:esitelminä (undersegmented pro N:esitelmi=nd), and N:tieteentilosoifi=a\*an (wrong segmentation and ending identification pro N:tieteentilosoifi\*an). The errors are due to surface homonyms where wrong analyses were picked due to the lack of sufficient lexical and/or contextual criteria.

### 5. Disambiguation of multiple tags

The theoretically most interesting parts of tagging programs tend to be those disambiguating multiple tags. Tag disambiguation is a particularly vexing problem in languages with little surface morphology and extensive zero declination: English is a paradigmatic example. In the Brown project, some 61 % of the words in the lexicon and only 51 % of the listed ending strings provided unique tags (Francis 1980). Despite the use of context frame rules covering several words on both sides of the target and sometimes applying in several left-right and/or right-left cycles in strings of words with multiple tags, roughly one fourth of the multiply tagged words had to be disambiguated manually.

These experiences led the LOB project to develop a program

which computes transitional probabilities between one tag and the next for any combinations of tags. All possible tag sequences were treated as first-order Markov chains. This simple probabilistic algorithm achieved a success rate of 94 % in disambiguating multiple tags when tried on a 15,000 word sample.

In comparison to English, it is a priori clear that the morphological richness of Finnish should entail a substantial reduction in word-level homography. This materializes in the fact that FINTAG, prior to disambiguation, introduces multiple tags for only 7.3 % of the word tokens in running text (estimated on the basis of a corpus containing 8,100 words). This is in harmony with the findings of Brodda and Karlsson (1981) concerning ending homography. Roughly, every tenth word is ambiguous in regard to either part of speech membership or morphological segmentation. The longest tag sequence assigned contains four tags and two ambiguous junctures:

V/VF/VII/N:pää\*s\*tä (a) negative verb stem V:päästä  
 (b) 2nd pers. sg. imperative VF:päästä  
 (c) 1st infinitive nominative VII:pääs\*tä  
 (d) relative sg. N:päästä

Also cf. the classical example ajusta which is seven ways ambiguous in terms of grammatical interpretation:

V/VF/N:alu\*s\*ta (a) negative verb stem V:ajusta  
 (b) 2nd person sg. imperative VF:ajusta  
 (c) nominative sg. ajusta  
 (d) relative sg. N:ajusta (of alku)  
 (e) relative sg. (of Alu 'trade mark')  
 (f) partitive sg. alus=ta (of alus)  
 (g) - " - (of alunen)

Roughly half of the ambiguous tag sequences in the test corpus belong to two types, either A/N (adjective vs. noun, e.g. Suomalainen 'Finnish; Finn'), or PP/ADV (preposition-postposition vs. adverb, e.g. keskeillä 'in the middle of; in the mid-

die'). Some 45 % belong to the rest of the types in the following list (the items are not in strict descending frequency order).

MULTIPLE TAGS	EXPLANATION	EXAMPLE
A/N	adjective vs. noun	suomalainen
PP/ADV	preposition vs. adverb,	keskellä
	postposition vs. adverb	vieressä
PR/ADV	pronoun vs. adverb	tässä
VCG/VI3	agent participle vs. 3rd infinitive	hankkimaan
V/VII	negative pass. stem	
	vs. 1st infinitive	tulla
VF/VII	finite 3rd person sg. pres. indic.	antaa
	vs. 1st infinitive	
VPAl/VP	1st act. participle nom. pl. vs. finite verb 3rd person pl. pres. indic.	antavat
V/VF	negative stem vs. finite 2nd person sg. imperative	anna
V/VF/N	negative stem vs. finite 3rd person sg. pres. indic. vs. noun	voi
VF/N	finite verb vs. noun	voim
PP/N	preposition vs. noun	ilman

The vast majority of these ambiguities is decidable given either a simple memory of a crucial word or constituent that occurred earlier in the same clause, or a one-word look-ahead. Even though the current Helsinki implementation of BETA is somewhat awkward in dealing with syntactic generalizations, those in the list above present no serious difficulties. The following basic look-ahead strategy thus solves several ambiguities:

Look-ahead strategy<sub>1</sub>

Check whether the current ambiguous nominal has the same case and number as the following word.

If the outcome is positive, several ambiguities dissolve:

MULTIPLE TAGS	IF NEXT WORD HAS SAME CASE AND NUMBER, REDUCE TO	OTHERWISE REDUCE TO
A/N	A	N
PR/ADV	PR	ADV
VPG/VI3	VPG	VI3
VPAl/VP	VPAl	VP

The tags in the middle column mark pronominal modifiers: adjectives, pronouns, agent participles (e.g. VPG:hankki-maan), and 1st (ie. present) participles. If the simple case/number test fails, the interpretation is as in the third column (noun, adverb, 3rd infinitive, finite verb).

The prepositions also need a straightforward look-ahead test for finding out whether their government requirement is satisfied:

Look-ahead strategy<sub>2</sub>

Check whether the case of the following word is as required by the current preposition.

If so, PP/ADV and PP/N reduce to PP, otherwise to ADV and N. The precision of the look-ahead strategies is close to perfection. This should be no surprise in view of the mutually reinforcing interplay between probability of occurrence and language structure. Since the modifier interpretation presupposes one particular case-number combination out of a paradigm of 28 (2 numbers, 14 cases), the probability is small for the current word to be a head subsequently followed by a head occurring in the same case and number. Occasionally, this situation arises especially in strings of nouns in the genitive if one of them is a topicalized object.

The remaining ambiguities may be solved by a combination of (1) a set of syntactic demons keeping track of certain basic

properties, especially polarity and finiteness, so far identified during the analysis of the current clause, and (ii) applying heuristic strategies given the information supplied by the demons. Thus, to mention the most salient ones (in a slightly simplified form in regard to coordinate structures):

MULTIPLE TAGS	REDUCE TO	IF
V/VII	V	the negative verb has been found
V/VII	VII	a finite verb has been found
VF/VII	VF	in initial position (= imperative)
VF/VII	VF	no finite verb has been found
VF/VII	VII	a finite verb has been found
V/VF	V	the negative verb has been found
V/VF	VF	no finite verb has been found
VF/N	VF	"-"
VF/N	N	a finite verb has been found

These strategies, and a few minor ones not treated here, together conclusively and correctly disambiguate some 93 % of all multiple tags. Of course the strategies, due to their heuristic nature, give rise to errors but the impact of these is minor (cf. section 6). The theoretical conclusion to be drawn is that morphological and syntactic disambiguation do not, normally, presuppose excessive computation but are resolved by local means, i.e. in terms of the immediate context. If this were not the case, language understanding would be slowed down to an impracticable level.

A few examples illustrate how multiple tags are disambiguated. The examples come pairwise with input above and disambiguated output below.

V/VF:ole ADV:nyt ADV:hiljaa.  
VF:ole ADV:nyt ADV:hiljaa.

VF/VII:saatta=a N:kalle V/VII:tul=ja.  
VF:saatta=a N:kalle VII:tul=ja.

N:kalle VF:ei V/VF/N:toim\*i ADV:reilu=sti.  
N:kalle VF:ei V:toimi ADV:reilu=sti.

VF:en V/VF:osta V/VF/PR:tuo\*ta N:auto=a.  
VF:en V:osta PR:tuo=ta N:auto=a.

N:kalle VF:halua=a VF/VII:anta=a N:lahja=n.  
N:kalle VF:halua=a VII:anta=a N:lahja=n.

VPAL/VF:tule\*vat N:ostaja=t VPAL/VF:saatta\*vat VII:rikastu=a.  
VPAL:tule=va=t N:ostaja=t VF:saatta=vat VII:rikastu=a.

N:kalle V/VF/N:tapaa VPAL/VF:tule\*vat N:ostaja=t.  
N:kalle VF:tapaa VPAL:tule=va=t N:ostaja=t.

A/N:japanilaise=t N:auto=t VF:o=vat VF/PR:tuo=n PP/ADV:takana.  
A:japanilaise=t N:auto=t VF:o=vat PR:tuo=n PP:takana.

PP/ADV:takana VF:ol=i A/N:japanilainen.  
ADV:takana VF:ol=i N:japanilainen.

PR:he VF:ajo=i=vat N:tuula=n VPG/VII:osta=ma=lla N:auto=lla.  
PR:he VF:ajo=i=vat N:tuula=n VPG:osta=ma=lla N:auto=lla.

N:kalle VF:rikastu=i VPG/VII:osta=ma=lla C:ja VPG/VII:myy=mä=lla.  
N:kalle VF:rikastu=i VII:osta=ma=lla C:ja VII:myy=mä=lla.

N:lapse=n VII3/N:synty\*mään VF:on ADV:vain N:kuukausi.  
N:lapse=n N:syntymään VF:on ADV:vain N:kuukausi.

## 6. Evaluation

So far, FINTAG has been applied on a larger scale only to a corpus containing 66,000 word-forms. This is the HKV corpus containing basically neutral newspaper and magazine prose (cf. Hakulinen, Karlsson and Vilkuna 1980 for detailed documenta-

tion). During the testing phase, twelve smaller corpora of 1,500 words each, from diverging styles, were used for avoiding stylistic bias and oversimplification.

FINTAG has an average precision (correctness rate) of some 96-97 %, i.e. only 3-4 % of the tagged and segmented words have to be manually corrected. Strict criteria were applied in evaluating correctness. Thus, a tagged word token is regarded as correct only if all of the following three criteria are satisfied:

- (a) the part of speech label is proper and unique, i.e. disambiguated in context if multiple,
- (b) all endings are properly segmented, and no hesitantly located junctures (indicated by "\*\*") remain,
- (c) the root has not been "invaded" by overgenerating suffix segmentation rules.

Improper or multiple part of speech tags remaining after disambiguation (MIS) are counted as errors, as are all improper or lacking morphological junctures.

The precision of FINTAG is at the same level as that reported by Leech, Garside and Atwell (1983:22) for the final version of the LOB tagging program. These figures are not strictly comparable, however, since LOB uses many more tags than FINTAG (cf. above), whereas FINTAG provides more detailed morphological analyses than LOB.

The manual correction phase is a time-consuming and laborious one. Four cycles of proofreading seem to be needed if (almost) all remaining errors are to be eliminated.

The level 96-97 % is not an absolute limit. The program could still be amended and modified, and the precision level pushed towards 98-99 %. Trivially, this can be done by extending the lexical lists used in various parts of the program suite. More interestingly, tag disambiguation could also be made somewhat more efficient. However, upon approaching 100 %, tagging problems tend to be identical to those that have to be

solved by a high-quality parsing algorithm.

For the sake of comparison, we note that 23 % of the words had to be manually corrected in the course of the Brown project (Francis 1980:202). Svartvik & al. (1980:69) report a similar 80 % precision level for the Survey of Spoken English tagging phase. These projects used considerably smaller lexicons and more restricted suffix analysis methods than LOB and FINTAG.

#### 7. Corpus uses

FINTAG was designed primarily for generating corpora needed by the parsing project. Robust parsers and "performance grammars" cannot be constructed without rapid access to well organized, representative corpora. We conclude by a concrete example demonstrating how further processing of a tagged corpus may yield efficient tools.

The tagged and corrected HKV corpus, containing 10,150 simplex clauses, was subjected to a second manual phase of preparation during which all punctuation marks were disambiguated for the purpose of indicating clause structure. Punctuation marks not indicating sentence or clause boundaries were changed to characters not occurring elsewhere in the corpus. This procedure provided a corpus where sentences, complex or simplex, are unambiguously defined as strings occurring between elements drawn from set (a). Simplex clauses are delineated by elements from set (b):

- (a) . : ; ! ?
- (b) . . : ; ! ? ,

One version exists as a file where each clause is represented only by its tags, junctures, and endings. The initial number sequence is an identifier.

Simplex clauses in terms of tags and inflectional allomorphs

10002 N=N A=N N N- C N VP V.  
10018 N VP A=N A6 A C N\$SEEN VPA1=VA.  
10034 N=LLA VF ADV V PR=TA=AN N=TTA A=STA N=STA VII=A,  
10052 C NE=N N VP VPA2=NYT PR=N,  
10064 PR VP VPA2=LGT N=KIN VII=DA.  
10074 C N C N VP=ISI=VAT N=LLE ADV N=J=A,  
10090 VP=ISI PR=I=LLE ADV=STI ADV N=X.  
10120 PR=N VP=A PR=EN A=I=EN N=EN N=N NE NE=N N:  
10138 A N VP VPA2=NYT VII=MA&AN N=T,  
10150 C/VF=KA N=TA V=TA VII=MA&AN N=I=DEN PPA&AN,  
10162 C A N=I=NA VP ADV=STI VPP2=TY.

Here, the shapes of all inflectional morphs have been preserved, including vowel harmony. This file provides morphologically useful information on the distribution and use of allomorphs. For syntactic purposes, a further version of the same material has been prepared, just by using the mainframe operating system's command REPLACE, where all allomorphs have been reduced to morphosyntactic properties:

Simplex clauses in terms of tags and morphosyntactic features

10002 N=GEN=SG A=NOM=SG N=GEN=SG N- C N=NOM=SG VP=PRES=IND=SG3  
V=NEG.  
10018 N=NOM=SG VP=PRES=IND=SG3 A=GEN=SG A=NOM=SGs A=NOM=SG C  
N=ILL=SG VPA1=NOM=SG.  
10034 N=ADE=SG VP=PRES=IND=SG3 ADV V=NEG PR=PTV=SG N=PTV=SG  
A=ELA=SG N=ELA=SG VII=NOM,  
10052 C NE=GEN=SG N=NOM=SG VP=PRES=IND=SG3 VPA2=NOM=SG  
PR=GEN=SG,  
10064 PR=NOM=SG VP=PRES=IND=SG3 VPA2=NOM=SG N=NOM=SG=KIN  
VII=NOM.  
10074 C N=NOM=SG C N=NOM=SG VP=COND=PL3 N=ALL=SG ADV N=PTV=PL,  
10090 VP=COND=SG3 PR=ALL=PL ADV=STI ADV N=PTV=SG.  
10120 PR=GEN=SG VP=PRES=IND=SG3 PR=GEN=PL A=GEN=PL N=GEN=PL  
N=GEN=SG NE=NOM=SG NE=GEN=SG N=NOM=SG:  
10138 A=NOM=SG N=NOM=SG VP=PRES=IND=SG3 VPA2=NOM=SG VII=ILL  
N=NOM=PL,

10150 C/VF=PRES=IND=SG3=KA N=PTV=SG V=PSS=NEG VII=ILL N=GEN=PL  
PP=ILL,  
10162 C A=NOM=SG N=ESS=PL VP=PRES=IND=SG3 ADV=STI VPP2=NOM=SG.

This file provides fast and easily expressible access to surface morphosyntactic patterns even by mainframe operating systems. E.g., if one wants to know how often, and under what conditions, genetical singular nominal premodifiers are separated from a head inflected in the nominative singular by one or more adjectival and/or pronominal premodifiers, one asks for instances of tag sequences such as "N=GEN=SG A=NOM=SG N=NOM=SG", "N=GEN=SG PR=NOM=SG A=NOM=SG N=NOM=SG", "N=GEN=SG A=NOM=SG A=NOM=SG N=NOM=SG". Clause-initial finite verbs are found by picking instances of VP occurring in the first two columns, etc. The corpus is big enough to ensure that at least the prototypical patterns can be found.

If the fairly rigid commands of the operating system are not flexible enough, the BETA system may be used for defining search keys. This makes possible defining virtually any pattern, e.g. "instances of composite tenses with at least two NPs of adverbial phrases between the main verb and the participle", "nouns in the partitive that occur before the main verb", "initial sequences of more than two adverbial phrases", "temporal clauses occurring before the main clause", "participles as premodifiers in the beginning of the clause before the finite verb", etc.

If one wishes to examine the original wording of the examples thus found, it is available on yet another file with parallel clause numbering. Thus the above structural skeletons materialize as the following clauses:

Full clauses

10002 VALTIO=N ENSI VUODE=N TULO- JA MENARVIO EI YLLA&VA.  
10018 BUDETTIESITYS ON KOHTUULLISE=N JAKK&VA, REALIS&TINEN JA  
PAKOTILANTEE&SEEN MUKAUTU=VA.  
10034 HALLITUKSE=LLA EI KUITENK&AN OLE MI=TA=AN AIHE=TT&VA VAL-  
LAISE=STA KIITOKSE=STA ILA&TU=A.

10052 SILLÄ SORSA=N KABINETTI ON TEH=NYT SE=N,  
 10064 MIKÄ ON OL=LUT PAKKO=KIN TEH=DÄ,  
 10074 JOS VEROMYLLY JA INFLAATIO PYÖRITTÄ=ISI=VÄT VALTIO=LLE  
   ENERMIÄN TULO=J=A,  
 10090 LÖYTY=ISI NI=I=LLE VARMA=STI MYÖS KÄYTTÖ=Ä,  
 10120 SE=N OSOITTA=A MONI=EN LIHAV=I=EN VUOSI=EN VALTIOVARAIN-  
   MINISTERI=N JOHANNES VIROLAISE=N VÄITE:  
 10138 POLIITINEN ELÄMÄ ON KYPSY=NYT TUNNUSTA=MAAN TOSIASIA=T,  
 10150 EI=KÄ TULEVAISUUT=TA PYRI=TYÄ RAKENTTA=MAAN TOIIVE=I=DEN  
   VARA=AN,  
 10162 KUVEN VIIME VUOS=I=NA ON VALITETTAVA=STI TEH=TY.

References

- Aarts, J. and Weijts, W. 1984, eds. Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research. Amsterdam: Rodopi.
- Atwell, E. 1983. "Constituent Likelihood Grammar". ICAME News 7, 34-66.
- Atwell, E., Leech, G. and Garside, R. 1984. "Analysis of the LOB Corpus: Progress and Prospects". In Aarts and Weijts (1984, eds., 41-52).
- Brodde, B. 1982a. "Vad slags objekt är en tagg och vad slags objekt är det man taggar med dessa taggar?". In Källgren (1982, ed., 14-26).
- 1982b. "Problems with Tagging - and a Solution". Nordic Journal of Linguistics 5:2, 93-116.
- Brodde, B. and Karissov, F. 1981. An Experiment with Automatic Morphological Analysis of Finnish. University of Helsinki, Department of General Linguistics, Publications No. 7.
- Beg-Olofsson, M. 1982. "En språkstatistisk modell för ordklassmärkning i löpande text". In Källgren (1982, ed., 27-30).
- Beg-Olofsson, M. and Svartvik, J. 1984. "Four-level tagging of spoken English". In Aarts and Weijts (1984, eds., 53-64).
- Francis, W. N. 1964. A Standard Sample of Present-Day English for Use with Digital Computers. Report to U.S. Office of Education on Cooperative Research Project No. E-007. Providence: Brown University.
- 1980. "A Tagged Corpus - Problems and Prospects". In Studies in English Linguistics to Randolph Quirk, eds. S. Greenbaum, G. Leech and J. Svartvik, London: Longman, 192-209.
- 1982. "Problems of Assembling and Computerizing Large Corpora". In Johansson (1982, ed., 7-24).
- Garside, R. and Leech, G. 1982. "Grammatical Tagging of the LOB Corpus: General Survey". In Johansson (1982, ed., 110-117).
- Hakulinen, A., Karlsson, F. and Wilkuna, M. 1980. Suomen tekstilauseiden pilteistä. Kvantitatiivinen tutkimus. University of Helsinki, Department of General Linguistics,

- Publications No. 7.
- Johansson, S. 1982, ed. Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities.
- Jurberg, J. 1982. "Syntax". In Källgren (1982, ed., 31-39).
- Källgren, G. 1982, ed. Tagging. PLOS 47. Institute of Linguistics, University of Stockholm.
- Leech, G., Garside, R. and Atwell, E. 1983. "The Automatic Grammatical Tagging of the LOB Corpus". ICAME News 7, 13-33.
- Saakkonen, P. & al. 1979. Suomen kielen laajuussanasto. Porvoo: WSOY.
- Svartvik, J. and Egg-Olofsson, M. 1982. "Tagging the London-Lund Corpus of Spoken English". In Johansson (1982, ed., 85-109).
- Svartvik, J., Egg-Olofsson, M., Forsheden, O., Orestrom, B. and Thavenius, C. 1982. Survey of Spoken English. Report on Research 1975-1981. Lund Studies in English 63. Lund: CWK Gleerup.