UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Assigning Finnish words with inflection codes

Hurskainen, Arvi

2023-04-27

http://hdl.handle.net/10138/358090

Hurskainen, A 2023 'Assigning Finnish words with inflection codes' Technical Reports on Language Technology, no. 98, SALAMA - Swahili Language Manager, Helsinki.

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Technical Reports on Language Technology
Report No 98 2023
http://www.njas.helsinki.fi/salama

Assigning Finnish words with inflection codes[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Rule-based description of the language such as Finnish requires a careful encoding of all inflecting words. This includes verbs, nouns, adjectives, pronouns, and numerals. The Institute for the Languages of Finland (KOTUS) has worked for several years for classifying Finnish words into individual inflection classes. The classes are marked with combinations of letters and numbers. The classification of nominals starts from 1 and continues until 51. Many of these classes have also sub-classes, and they are marked with an additional code, using capital letters, such as A, B, C etc. The classification of verbs starts from 52 and continues until 73. In addition, KOTUS marks word classes of the words. Here I mark nominals with numbers preceded by the letter N, and the verbs with the letter V. Also verbs have sub-classes, marked with capital letters. The KOTUS list does not mark the inflection classes in compound words. The user should find the inflection code from the last word of the compound word. The wordlist has currently more than 104,000 words, and the number is increasing over the years. It is not exhaustive, but it covers a substantial number of words occurring in speech and writing. More than half of the words are compound words, and they are without inflection codes.

The list is defective in that it does not include information on the front/back inflection patterns. The vowels of the word stem define whether the word should have a front or back vowel inflection. Particularly problematic are compound words, because the last member of the compound defines the inflection pattern, and it is not easy to define computationally where the boundary is.

In this report I describe how the word list can be enhanced, so that it also has information on front/back inflection. Because back vowel inflection is more common, I have defined it as default and left without marking. Only front vowel inflection is marked.

I also show how front/back inflection codes can be added to such compound words that are not in the KOTUS list.

Key Words: *morphological analysis, inflection codes, front/back vowel inflection.*

---

[1] The report is issued under licence CC BY-NC

1 Introduction

The list of Finnish words, compiled by the Institute for the Language of Finland (KOTUS)[2], is a huge resource for anybody wishing to construct a rule-based language description of Finnish. Each word is assigned a code, except for compound words, which helps to decide how the word should inflect.

However, the list does not have information on the front/back vowel inflection, which would be essential in handling the words computationally. Also compound words without inflection codes should be provided with inflection codes.

The vowels of the word stem decide the inflection type. In case the stem vowels include one of *y*, *ä* or *ö*, the word has front vowel inflection. More difficult are cases, where the stem vowels are *e* or *i*. These are neutral vowels, and they can occur in both inflection types. However, if the word stem does not have hard back vowels, such as *a*, *o*, or *u*, the word inflects according to front pattern.

So far everything is clear. The problem arises, however, in cases where the word is a compound. The rule is that the compound word inflects according to the last member of the compound. If the boundaries of the members would be marked, the inflection pattern could be easily resolved. But this is not the case. Compounds are written as single words regardless the number of members. The compilers of the KOTUS list did not see the marking of boundaries important. Neither did they see the marking of the inflection type important in relation to front/back inflection.

Now when we consider the construction of the marking algorithm, we must take this problem into account. We cannot rely on the distance from the word end to the left in considering whether the word should be marked or not. There are compound words, which have a very short last member with front inflection and with neutral vowels, and in front of it is a member with back inflection. There are also such single-member words, that have in the end a long sequence of characters with only front vowels and neutral vowels, but then suddenly further to the left a back vowel, which prompts back vowel inflection. The conclusion is that we do not have a secure basis for correctly marking the words. We proceed in steps and try first by inspecting the last part of the words.

2 Positive rules

We write a set of rules that check, starting from the end of the word, whether there is one of the front vowels *y*, *ä*, and *ö* (1).

(1)
```
perl -pe 's/([yäö] [NV]\S+)\;/$1 f;/gm' | \
perl -pe 's/([yäö]. [NV]\S+)\;/$1 f;/gm' | \
perl -pe 's/([yäö].. [NV]\S+)\;/$1 f;/gm' | \
perl -pe 's/([yäö]... [NV]\S+)\;/$1 f;/gm'
```

[2] https://www.kotus.fi/aineistot/sana-aineistot/nykysuomen_sanalista

The first rule checks the last character of the word, and the fourth rule checks the fourth character from the end. These rules catch short words with front inflection, also such compounds that have a short last member. These rules are safe, because they check only the few final characters of the word. However, they do not catch such words that have only vowels e or i, and have front vowel inflection. Such vowels cannot be included into the rules, because they may also be part of back vowel words.
Words such as in (2) can be safely marked.

(2)
```
aluejärjestö N2 f
alttarikehä N10
alppitähti N7-F f
alppiyhdistetty N1-C f
```

3 Using negation in rules

We continue checking more vowels to the left. Instead of trying to figure out the positive triggers in marking the front vowel words, we can define what there should not be. We write a negative rule using hard back vowels, that is *a*, *o* and *u*. The rule is in (3).

(3)
```
perl -pe 's/([^aou][^aou][^aou][^aou][^aou][^aou][^aou][^aou]
[NV]\S+)\;/$1 f;/gm'
```

The rule reads: Add *f* to the end of the reading, if the word, followed by an inflection code, does not have *a*, *o*, or *u* in any of its eight last letter spaces.

The rule concerns only hard back vowels but not the neutral vowels *e* and *i*. The rule is safe in that it blocks out all such cases, where a hard back vowel appears in any of the eight last letter spaces. But if the word is longer, the rule is not safe (4).

(4)
```
ammatillinen N38 f
amatööriorkesteri N6 f
aluepoliittinen N38 f
aloitekykyinen N38 f
```

The problem occurs also when the last member of the compound is short, and the rule checks also characters that are not part of the last member (5).

(5)
```
aloitekyky N1-M
alkuväestö N2
alkuyö N19
```

These compounds should be marked with *f*, but they are not marked with the rule in (3), because a back vowel is found among the eight last characters of the word.

Although the rules above are rather safe, they leave many words unmarked. Also false marking cannot be avoided, because there is no mechanism for identifying the member boundaries in compound words. Therefore, we must use another approach.

4 Marking the word boundaries of compound words

If we can mark the compound words so that there is a mark showing the boundary of each member of the compound word, the task of adding inflection codes becomes very easy. Because the KOTUS list does not have such marking, we must find methods for doing this. In earlier reports I described the method of handling compound words in Finnish[3]. Using this version of the Finnish analyser as a starting point, and enhancing it according to the need, we are able to mark the boundary points in compound words.

4.1 Constructing the system for marking word boundaries in compound words

The SALAMA analyser for Finnish included already a method for marking word boundaries in compound words. The system was developed using a corpus of Finnish texts. Using this analyser, it was possible to mark the boundaries in part of the words in the KOTUS list. This list contains, however, also such words that occur only in speech and hardly ever in written text. Therefore, it was necessary to extract all such words in the KOTUS list that were not marked with the existing system and mark the boundaries manually. This laborious work made it possible to enhance the morphological analyser so that it also included the compound words in the KOTUS list.

As a result, we have an analyser that identifies word boundaries in compound words in the KOTUS list as well as in any running text with inflected word forms.

4.2 Test with the KOTUS list

In (6) are some extracts from the enhanced KOTUS wordlist[4]. The member boundaries of compound words were marked with enhanced SALAMA version.

(6)

```
amatöörin|oikeus N40                ambitio N3
amatööri|näyttelijä N12             ambitio|kysymys N39
amatööri|orkesteri N6               ambi|valenssi N5
amatööri|ottelu N2                  ambi|valentti N5-C
amatööri|urheilija N12              ambi|valenttinen N38
amatööri|urheilu N2                 ambra N9
amatööriys N40                      ambrosia N12
ambio|foni N12                      ambulanssi N5
ambio|foninen N38                   ambulanssi|lento|kone N48
```

[3] http://www.njas.helsinki.fi/salama/enhanced-method-for-describing-compound-words.pdf
http://www.njas.helsinki.fi/salama/translating-compound-words-from-fin-to-eng.pdf
http://www.njas.helsinki.fi/salama/translating-finnish-compound-words-in-context.pdf
[4] Inflection codes were added also to the more than 50,000 compound words.

```
ameba N9
amenorrea N12
amerikan|englanti N5-J
amerikan|rauta N9-F
amerikan|salaatti N5-C
amerikan|suomalainen N38
amerikan|suomi N7
amerikkalainen N38
amerikkalaistua V52
amerikkalaisuus N40
ametisti N5
amfetamiini N5
amfi|bio|ajo|neuvo N1
amfi|bio|lento|kone N48
amfibio|panssari|vaunu N1
amfi|teatteri N6
amfora N10
amino|happo N1-B
amiraali N5
ammateittain N99
ammateittainen N38
ammatillinen N38
ammatillinen N99
ammatillistaa V53
ammatillistua V52
ammatillisuus N40
ammatin|harjoitus N39
ammatin|kuva N10
ammatin|opettaja N10
ammatin|valinnan|ohjaaja N10
ammatin|valinnan|ohjaus N39
ammatin|valin N9-J
ammatti N5-C
ammatti-ihminen N38
ammatti|aine N48
ammatti|ajo|kortti N5-C
ammatti|ala N9
ammatti|apu N1-E
ammatti|autoilija N12
ammatti|auttaja N10
ammatti|etiikka N14-A
ammatti|haureus N40
ammatti|julkaisu N2
ammatti|järjestö N2
ammatti|kalastaja N10
ammatti|kasvatus N39
ammatti|kateus N40
ammatti|kieli N26
ammatti|kirjailija N12
ammatti|kirjallisuus N40
ammatti|kokemus N39
ammatti|korkea|koulu N50
ammatti|koulu N1
ammatti|koululainen N38
ammatti|koulun|opettaja N10
ammatti|koulutus N39
ammatti|kunnia N12
ammatti|kunta N10-J
ammatti|kunta|laitos N39
ammatti|kurssi N5
ammatti|kurssi|keskus N39
ammatti|kuva N10
ammatti|lainen N38
ammattilais|joukkue N48
ammattilais|ottelu N2
ammattilais|sopimus N39
ammattilaistaa V53
ammattilaistua V52
ammattilais|urheilija N12
ammattilais|urheilu N2
ammatti|laisuus N40
ammatti|lehti N7-F
ammatti|liikenne N48-J
ammatti|liitto N1-C
ammatti|mainen N38
ammatti|mainen N99
ammatti|maisuus N40
ammatti|mies N42
ammatti|ministeri N6
ammatti|muusikko N4-A
ammatti|nimike N48-A
ammatti|nyrkkeilijä N12
ammatti|nyrkkeily N2
ammatti|opetus N39
ammatti|opinto N1-J
ammatti|oppi|laitos N39
ammatti|oppilas N41
ammatti|oppilas|lauta|kunta
N10-J
ammatti|osasto N2
ammatti|poliitikko N1-A
ammatti|pätevyys N40
ammatti|rikollinen N38
ammatti|salaisuus N40
ammatti|sana N9
ammatti|sanasto N2
ammatti|slangi N5
ammatti|sotilas N41
ammatti|suoja N10
ammatti|taidoton N34-C
ammatti|taito N1-F
ammatti|taitoinen N38
```

```
ammatti|tappaja N10
ammatti|tauti N5-F
ammatti|termi N5
ammatti|toveri N6
ammatti|tulo N1
ammatti|tutkinto N1-J
ammatti|työ N19
ammatti|työntekijä N12
ammatti|työ|väki N7-D
ammatti|urheilija N12
ammatti|urheilu N2
ammatti|yhdistys N39
ammatti|yhdistys|liike N48-A
ammatti|yhdistys|toiminta N9-J
ammatti|ylpeys N40
amme N48
öljy N1
öljy-yhtiö N3
öljy|alue N48
öljy|esiintymä N10
öljy|happo N1-B
öljy|hiekka N9-A
öljy|hoito N1-F
öljyinen N38
öljy|johto N1-F
öljy|kakku N1-A
öljy|kalvo N1
öljy|kamiina N13
öljy|kangas N41-G
öljy|kangas|takki N5-A
öljy|kanisteri N6
öljy|kankainen N38
öljy|kannu N1
öljy|kastike N48-A
öljy|kasvi N5
öljy|katastrofi N5
öljy|kenttä N10-C
öljy|kerrostuma N10
öljy|lamppu N1-B
öljy|lautta N9-C
öljy|liuske N48
öljy|lähde N48-F
öljy|läikkä N10-A
öljy|läiskä N10
öljy|lämmitteinen N38
öljy|lämmitys N39
öljy|löytö N1-F
öljy|maa N18
öljy|maalaus N39
öljy|maali N5
öljymäinen N38
öljyn|etsin N9-J
öljyn|jalostamo N2
öljyn|jalostus N39
öljyn|jalostus|teollisuus N40
öljyn|keräys|alus N39
öljyn|kulutus N39
öljyn|paine N48
öljyn|paine|mittari N6
öljyn|poraaja N10
öljyn|poraus N39
öljyn|poraus|lautta N9-C
öljyn|suodatin N33-C
öljyn|torjunta N9-J
öljyn|tuonti N5-J
öljyn|tuotanto N1-J
öljyn|tuottaja N10
öljyn|tuottaja|maa N18
öljyn|vaihto N1-F
öljyn|viejä N10
öljyn|viejä|maa N18
öljyn|vienti N5-J
öljy|palmu N1
öljy|pastelli N5
öljy|pisara N12
öljy|pitoinen N38
öljy|pohatta N14-C
öljy|pohja N10
öljy|pohjainen N38
öljy|politiikka N9-A
öljy|poltin N33-C
öljy|pumppu N1-B
öljy|puomi N5
öljy|puu N18
öljy|sampoo N20
öljy|satama N10
öljy|sora N10
öljy|sora|päällyste N48
öljy|sora|päällysteinen N38
öljy|sorastaa V53
öljy|sorastus N39
öljy|sora|tie N19
öljy|suo N26
öljy|säiliö N3
öljy|tahra N9
öljy|takki N5-A
öljy|tankkeri N6
öljy|tankki N5-A
öljy|teollisuus N40
öljy|tila N9
öljytä V52-C
öljy|tuho N1
```

```
öljy|tuikku N1-A
öljy|tuote N48-C
öljy|tynnyri N6
öljytä V74
öljy|vahinko N1-G
öljy|valaistus N39
öljy|valo N1
öljy|varanto N1-J
öljy|varasto N2
öljy|vara N9
öljy|väki|rehu N1
öljy|väri N5
öljy|väri|liitu N1-F
öljy|väri|maalaus N39
öljyyntyä V52-J
öljy|sheikki N5-A
örinä N12
```

Now we have a possibility to write a simple rule for marking words with front vowel inflection (7).

(7)
```
perl -pe 's/((^|\||\-)[^aou]+ [NV]\S+)/$1 f/gm'
```

Note that the rule has as its left boundary either the beginning of line `'^'`, a dash `'-'`, or a vertical bar `'|'`. They are left boundaries of the last member of compound words, and of single words. When we apply the rule to the list in (6), we get the result as in (8).

(8)
```
amatöörin|oikeus N40
amatööri|näyttelijä N12 f
amatööri|orkesteri N6
amatööri|ottelu N2
amatööri|urheilija N12
amatööri|urheilu N2
amatööriys N40
ambio|foni N12
ambio|foninen N38
ambitio N3
ambitio|kysymys N39 f
ambi|valenssi N5
ambi|valentti N5-C
ambi|valenttinen N38
ambra N9
ambrosia N12
ambulanssi N5
ambulanssi|lento|kone N48
ameba N9
amenorrea N12
amerikan|englanti N5-J
amerikan|rauta N9
amerikan|salaatti N5-C
amerikan|suomalainen N38
amerikan|suomi N7
amerikkalainen N38
amerikkalaistua V52
amerikkalaisuus N40
ametisti N5
amfetamiini N5
amfi|bio|ajo|neuvo N1
amfi|bio|lento|kone N48
amfibio|panssari|vaunu N1
amfi|teatteri N6
amfora N10
amino|happo N1-B
amiraali N5
ammateittain N99
ammateittainen N38
ammatillinen N38
ammatillinen N99
ammatillistaa V53
ammatillistua V52
ammatillisuus N40
ammatin|harjoitus N39
ammatin|kuva N10
ammatin|opettaja N10
ammatin|valinnan|ohjaaja N10
ammatin|valinnan|ohjaus N39
ammatin|valin N9-J
ammatti N5-C
ammatti-ihminen N38 f
ammatti|aine N48
ammatti|ajo|kortti N5-C
```

```
ammatti|ala N9
ammatti|apu N1-E
ammatti|autoilija N12
ammatti|auttaja N10
ammatti|etiikka N14-A
ammatti|haureus N40
ammatti|julkaisu N2
ammatti|järjestö N2 f
ammatti|kalastaja N10
ammatti|kasvatus N39
ammatti|kateus N40
ammatti|kieli N26 f
ammatti|kirjailija N12
ammatti|kirjallisuus N40
ammatti|kokemus N39
ammatti|korkea|koulu N50
ammatti|koulu N1
ammatti|koululainen N38
ammatti|koulun|opettaja N10
ammatti|koulutus N39
ammatti|kunnia N12
ammatti|kunta N10-J
ammatti|kunta|laitos N39
ammatti|kurssi N5
ammatti|kurssi|keskus N39
ammatti|kuva N10
ammatti|lainen N38
ammattilais|joukkue N48
ammattilais|ottelu N2
ammattilais|sopimus N39
ammattilaistaa V53
ammattilaistua V52
ammattilais|urheilija N12
ammattilais|urheilu N2
ammatti|laisuus N40
ammatti|lehti N7 f
ammatti|liikenne N48-J f
ammatti|liitto N1-C
ammatti|mainen N38
ammatti|mainen N99
ammatti|maisuus N40
ammatti|mies N42 f
ammatti|ministeri N6 f
ammatti|muusikko N4-A
ammatti|nimike N48-A f
ammatti|nyrkkeilijä N12 f
ammatti|nyrkkeily N2 f
ammatti|opetus N39
ammatti|opinto N1-J
ammatti|oppi|laitos N39
ammatti|oppilas N41
ammatti|oppilas|lauta|kunta
N10-J
ammatti|osasto N2
ammatti|poliitikko N1-A
ammatti|pätevyys N40 f
ammatti|rikollinen N38
ammatti|salaisuus N40
ammatti|sana N9
ammatti|sanasto N2
ammatti|slangi N5
ammatti|sotilas N41
ammatti|suoja N10
ammatti|taidoton N34-C
ammatti|taito N1
ammatti|taitoinen N38
ammatti|tappaja N10
ammatti|tauti N5
ammatti|termi N5 f
ammatti|toveri N6
ammatti|tulo N1
ammatti|tutkinto N1-J
ammatti|työ N19 f
ammatti|työn|tekijä N12 f
ammatti|työ|väki N7-D f
ammatti|urheilija N12
ammatti|urheilu N2
ammatti|yhdistys N39
ammatti|yhdistys|liike N48-A f
ammatti|yhdistys|toiminta N9-J
ammatti|ylpeys N40 f
amme N48
öljy N1 f
öljy-yhtiö N3 f
öljy|alue N48
öljy|esiintymä N10 f
öljy|happo N1-B
öljy|hiekka N9-A
öljy|hoito N1 f
öljyinen N38 f
öljy|johto N1
öljy|kakku N1-A
öljy|kalvo N1
öljy|kamiina N13
öljy|kangas N41-G
öljy|kangas|takki N5-A
öljy|kanisteri N6
öljy|kankainen N38
öljy|kannu N1
öljy|kastike N48-A
öljy|kasvi N5
öljy|katastrofi N5
```

```
öljy|kenttä N10-C f
öljy|kerrostuma N10
öljy|lamppu N1-B
öljy|lautta N9-C
öljy|liuske N48
öljy|lähde N48 f f
öljy|läikkä N10-A f
öljy|läiskä N10 f
öljy|lämmitteinen N38 f
öljy|lämmitys N39 f
öljy|löytö N1 f f
öljy|maa N18
öljy|maalaus N39
öljy|maali N5
öljymäinen N38 f
öljyn|etsin N9-J f
öljyn|jalostamo N2
öljyn|jalostus N39
öljyn|jalostus|teollisuus N40
öljyn|keräys|alus N39
öljyn|kulutus N39
öljyn|paine N48
öljyn|paine|mittari N6
öljyn|poraaja N10
öljyn|poraus N39
öljyn|poraus|lautta N9-C
öljyn|suodatin N33-C
öljyn|torjunta N9-J
öljyn|tuonti N5-J
öljyn|tuotanto N1-J
öljyn|tuottaja N10
öljyn|tuottaja|maa N18
öljyn|vaihto N1
öljyn|viejä N10 f
öljyn|viejä|maa N18
öljyn|vienti N5-J f
öljy|palmu N1
öljy|pastelli N5
öljy|pisara N12
öljy|pitoinen N38
öljy|pohatta N14-C
öljy|pohja N10
öljy|pohjainen N38
öljy|politiikka N9-A
öljy|poltin N33-C
öljy|pumppu N1-B
öljy|puomi N5
öljy|puu N18
öljy|sampoo N20
öljy|satama N10
öljy|sora N10
öljy|sora|päällyste N48 f
öljy|sora|päällysteinen N38 f
öljy|sorastaa V53
öljy|sorastus N39
öljy|sora|tie N19 f
öljy|suo N26
öljy|säiliö N3 f
öljy|tahra N9
öljy|takki N5-A
öljy|tankkeri N6
öljy|tankki N5-A
öljy|teollisuus N40
öljy|tila N9
öljytä V52-C f
öljy|tuho N1
öljy|tuikku N1-A
öljy|tuote N48-C
öljy|tynnyri N6 f
öljytä V74 f
öljy|vahinko N1-G
öljy|valaistus N39
öljy|valo N1
öljy|varanto N1-J
öljy|varasto N2
öljy|vara N9
öljy|väki|rehu N1
öljy|väri N5 f
öljy|väri|liitu N1
öljy|väri|maalaus N39
öljyyntyä V52-J f
öljy|sheikki N5-A f
örinä N12 f
```

We note that all words, single words and compound words, were correctly marked. In manual checking, no marking mistakes were found in the result of marking the whole KOTUS list. The faultless marking was due to the manual construction of the SALAMA computational analyser for this purpose.

5 Assigning inflection codes on wordlists without codes

Above I have described how front/back vowel inflection codes can be assigned on the KOTUS wordlist that already has codes for inflection classes. The system that I have described can be applied also to such wordlists that have bare words without any inflection codes.

We can imagine that we want to construct a wordlist using a corpus of 100 million words as source text. We have an analyser that returns the base form of each word.[5] As a result we have an alphabetical list of the base forms of all words in the corpus, and information on the part-of-speech category of each form, but no inflection codes.

Our task is to add inflection codes to each word in the list, including front/back vowel inflection codes.

We take the enhanced KOTUS list described above. The list has full information on inflection of each word, and also the word boundaries of compound words are marked. This KOTUS list is converted into a Beta[6] rule file, such as in (9).

(9)

```
! Adds inflection code to word lists
!
CHARACTER-SETS
B: BLANK #

LIMITOR: #
#: #

STATE-SETS
Start: 1

RULES
!                                      lc rc  sc     rs   mv  md
amatöörinoikeus; amatöörin|oikeus N40; 0  B   Start  0    5   1
amatöörinäyttelijä; amatööri|näyttelijä N12 f;
amatööriorkesteri; amatööri|orkesteri N6;
amatööriottelu; amatööri|ottelu N2;
amatööriurheilija; amatööri|urheilija N12;
amatööriurheilu; amatööri|urheilu N2;
amatööriys; amatööriys N40;
ambiofoni; ambio|foni N12;
ambiofoninen; ambio|foninen N38;
ambitio; ambitio N3;
ambitiokysymys; ambitio|kysymys N39 f;
```

---

[5] In SALAMA, I have implemented the morphological analyser so that it returns directly also the inflection codes for words. Therefore, the process described here is not needed if using SALAMA.

[6] The manual for using Beta can be found here:
https://github.com/koskenni/beta/blob/master/betaref.md

```
ambivalenssi; ambi|valenssi N5;
ambivalentti; ambi|valentti N5-C;
ambivalenttinen; ambi|valenttinen N38;
ambra; ambra N9;
ambrosia; ambrosia N12;
ambulanssi; ambulanssi N5;
ambulanssilentokone; ambulanssi|lento|kone N48;
...
```

The Beta rules, such as in (8) above, convert the plain words in base form into forms with inflection codes and put marks on word boundaries in compound words. The boundary marks can be removed if not needed.

Note that the left context (lc) is defined as zero (0), meaning that it is not defined. This means that the rules apply also to such compound words that have one or more members on the left than the word in the rule list has. The inflection code will be in any case the same because it is defined on the basis of the last member of the word.

The Beta rule file can be enhanced by adding such words that are missing in the KOTUS list.

Using Beta rules, it is also possible to add inflection codes to compound words in the KOTUS list. If the list is converted into a Beta rule file, and the left context of the words is left undefined, the system finds the last members of the compound words from the list and assigns the appropriate codes to the compound words.

6 Problems in assigning the inflection codes on KOTUS wordlist

The KOTUS wordlist has been encoded incrementally over the years, and inconsistencies in encoding are inevitable. In order to be usable in developing language processing tools, the encoding must be systematic. Therefore, the production of the properly and fully encoded wordlist is necessary. Nor only single words need to be encoded. Also compound words must be encoded, so that inflection patterns for each word can be controlled.

Therefore, systematizing the completing the encoding has taken much effort. As a result, each word in the list has an inflection code except for such words as *alku-* that are the beginning part of the compound word and need no encoding.

The parentheses around some sub-class codes have been removed and the format of sub-class encoding has been made uniform. Also commas sometimes used for separating codes have been replaced with spaces.

The front/back vowel inflection codes have been added only to such words that inflect. Therefore, adverbs, although many of them have been derived from adjectives, are without front/back codes.

In the KOTUS list, the words with two alternative part-of-speech classes have been listed only once. In the new implementation, both alternatives have separate entries. This solution has increased the list by about 1.300 new entries.

7 Discussion and conclusion

In computational processing of Finnish text, it is necessary to be able to identify the inflection type of words. Not only the inflection class is important, also it is important to

know whether the word has a font or back vowel inflection pattern. In single words, the rule is easy to construct. In compound words the problem is very difficult to resolve, because the inflection pattern depends on where the boundary before the last member of the compound is.

Traditionally the front/back inflection was implemented using two-level rules. The method works, where only single words occur. The problem becomes almost unresolvable in case of compound words.

Therefore, it becomes necessary to construct a system, where the boundaries between members of compound words are marked. Once this is achieved, we have a system where we can mark boundaries in word lists such as the KOTUS list. We also can mark boundaries in compound words in running text, because the analysis system identifies also inflected forms of compounds.