*Technical Reports on Language Technology*
*Report No 87, 2022*
*http://www.njas.helsinki.fi/salama*

# Converting Standard Finnish to Kitee dialect[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi
*DOI: 10.13140/RG.2.2.30856.49920*

**Abstract**

The report describes the process, how Standard Finnish can be converted to Kitee dialect, a speech variety spoken in Northern Karelia. Because the dialect does not have a standard written form, the output of the conversion process is more or less on ad hoc basis. Yet it gives an approximate picture of how each word should be pronounced. The paper discusses possible approaches for implementation and describes the selected approach in detail.

**Key Words:** *morphological analysis, dialectology.*

## 1 Introduction

The incentive for constructing the dialect converter came from discussions among my class mates about the proper pronunciation of Northern Karelian dialect. No agreement was found, of course, because the concept 'dialect' is a continuum. The speech varieties change gradually when we move from one place to another. Therefore, I do not claim that what I have implemented here is a correct Northern Karelian dialect. I only claim that it is a speech variety spoken in the 1950'ies in Närsäkkälä, a village of Kitee on the Russian border. In a neigbouring village the speech may already be a bit different.

The reader may wonder what is the point in attempting to construct such a converter. The answer is: it is just sheer fun. But the actual implementation was very far from fun, as it turned out later. The process also revealed many interesting features in the development of dialects. I will discuss some of them below.

## 2 Approaches for constructing a dialect converter

At least two approaches come to mind for implementing the dialect converter. Because the conversion process is not actual language translation, some lighter methods would probably do the job.

One could consider implementing it by identifying typical sound changes and constructing a rule set for making the needed changes. If the changes would always take

place in the same environment, this would be the easiest solution. However, such rules easily produce overgeneration and apply to wordforms where they should not apply. A dialect is not a monolithic structure, where global rules apply. Therefore, this approach was abandoned.

I took another approach, where I made use of language analysis, so that I could consider each morpheme separately, and rewrite the morpheme if needed. This required enhancing the morphological lexicon, so that the input morpheme was copied to the output section, so that it was possible to modify it there to meet the requirements of the dialect.

After modifying the lexicon there were two alternatives for writing the conversion rules.

In one method, all such entries that required conversion were retrieved into a separate file, which was then modified into a Beta rule file. The actual conversion in the morphological lexicon was then made using Beta rules. Using sufficient context in conversion rules, it was possible to make needed changes into the morphological lexicon.

I soon found out that it was more secure to modify the required entries directly in the morphological lexicon, although it required that each entry in the lexicon needed to be checked separately. This dull work was somewhat alleviated when part of changes was made using macros.

Rewriting morphemes also concerned all inflection morphemes. In Finnish, there are hundreds of inflection classes, if we count the inflection of verbs and nominals (nouns, adjectives, pronouns and numerals). Each entry also in these sub-lexicons needed checking and possibly modification.

It was possible to make most of the needed modifications into the morphological lexicon, but not all. The problem is that phonological changes do not take place inside morphemes only. Many changes take place on points, where morpheme boundaries are crossed. Therefore, there was need to work out a post-analysis module, where it was possible to implement cross-morpheme changes.

I use the example in (1) to demonstrate the conversion phases needed.

(1)
Minä tahdon syödä banaania. (I want to eat banana.)

When we analyse the sentence using the enhanced lexicon, we get the result as in (2).

(2)
```
"<*minä>"
      "minä" PERSPRON CAP SG1:m NOM:ie
"<tahdon>"
      "tahtoa" V VMOD V52-Fo R:tah PRES SG1:on
"<syödä>"
      "syödä" V TRV V64y-f R:s Clit-f INF:yyvvä
"<banaania>"
      "banaani" N N6 R:pannaan SG PAR:ii
"<.>"
      "." **CLB
```

The analyser prints all output morphemes as they are defined in the lexicon. The lexical entries that are involved in the result in (2) are shown in (3).

(3)

```
m PersSG1-2 "minä SG1:m";
inä #f " NOM:ie";

tah V52-Fo "tahtoa VMOD V52-Fo R:tah";
don Clit " PRES SG1:on";

s V64y-f "syödä TRV V64y-f R:s";
yödä Clit-f " Clit-f INF:yyvvä";

banaan N6 "banaani N6 R:pannaan";
ia POS-an " SG PAR:ii";
```

In the beginning of the line is the lexical morpheme of the standard language, and in the end, preceded by a colon ':', is the form in Kitee dialect. The conversions are as in (4).

(4)

```
m > m
inä > ie

tah > tah
don > on

s > s
yödä > yyvvä

banaan > pannaan
ia > ii
```

When morphemes are converted into desired form, they are concatenated into words. The final result is in (5).

(5)
*Mie tahon syyvvä pannaanii.*

We can note that the personal pronouns are likely to change, usually into a shorter form, or into a form that is easy to pronounce. There are two sonoral consonants that cannot be pronounced in Kitee dialect. These are the alveolar *d* and labial *b*. The corresponding stop consonants *t* and *p* can be pronounced without problem. There are various methods for implementing the pronunciation of *b* and *d*. In our example *d* vas simply omitted in one case. In another case, the consonant *v* in double form substitutes *d*, and the vowel structure around changes. The consonant *b* was substituted with its stop variant *p*.

We also note that when a consonant is preceded by a vowel and followed by a double vowel, the consonant is doubled. This is a very common phenomenon, and I tested whether it is a global rule by writing a corresponding script. The result showed that the rule is not global, and I had to remove it.

## 3 Observations in constructing the converter

Because there was need to copy the morphemes to the output section, the lexicons became much larger than the standard lexicons are. The Finnish lexicon was already split into two files, because the implementation of compound nouns was included into the system, and the analyser was not able to run the lexicon in one piece. Now it turned out that even two lexicons do not work properly, and I had to construct a third and fourth lexicon and move part of words there. With these modifications the system became operative.

Another problem was the combination of morphemes. As the modifications were made to entries in the lexicon, it was not sure that the modifications made in inflection lexicons applied correctly to all words of that class. The conventions in the dialect do not necessarily follow the same rules in all words of the same class.

The third problem was that there are differences in how old and commonly used words behave compared with newly introduced words. The older words undergo more drastic modifications than the newer words. This can be seen, for example, in the implementation of *d*. In older words it has such realisations as v, j, and omission. Some consonant clusters cause also other realisations. In newer words, *d* is simply substituted with *p*.

A peculiar feature in Kitee dialect is that it does not allow any word-initial consonant clusters. All consonants except for the last consonant in the cluster are stripped off, no matter how ugly the result sounds, When we add to this the change of *d* into *t*, *b* into *p*, and *g* into *k*, we get very interesting wordforms. If the consonant cluster is in the middle of the word, it may, or at least part of it, be retained, or changed into another cluster. This is seen, for example, in the family name Wetterstrand as it is pronounced in Närsäkkälä. There is a cluster of four consonants, and as such impossible to pronounce in Närsäkkälä. It is pronounced Vestekranni without hesitation, retaining very little of the original form.

Why have they given such non-pronounceable names to innocent people? Other good examples of family names in Närsäkkälä are in (6).

(6)
Gröhn > Ryöni
Björn > Pyörmy
Ek > Ekki
Gerlander > Kerlanteri
Sirén > Sirreeni
Brander > Ranteri
Stepanoff > Tepanohvi

If the purpose to give such names was to civilize people, the result is disastrous. Poor people, who cannot even pronounce their name properly! In fact, they don't even try. Why were they not allowed to use such names as Tuppurainen and Tappurainen, which they could pronounce without difficulty?

In Ostrobotnia, western Finland, boasting is customary, not shameful at all. A farmer boasted that he has three cars, Ratsun (Datsun), Rotke (Dodge), and Riiseli-mersu (Diesel-Mercedes). They substitute the word-initial *d* with *r*. In Närsäkkälä, such prosperous farmers do not exist. If there would, the farmer might say that he has Tatsun, Totke and Tiissel-mersu.

The consonant *g* is substituted with its stop variant *k*, except for in the cluster *ng*, where it is retained (6).

(6)
`Guggenheim > Kukkenheimi`
`kangas > kangas`

**4 Post-processing**

Because the word forms in target dialect were constructed by concatenating the rewritten morphemes, the result is not always correct. For example, doubling of the consonant may be needed, if the result form has two vowels after the consonant, and the consonant is preceded by a vowel. Such processes must be controlled after the word-form has been concatenated. Examples are in (7).

(7)
`apua > ap + uu > apuu > appuu` (help)
`ikään > i + kä + än > ikään > ikkään` (age)
`ihottumaa > ihottum + oo > ihottumoo > ihottummoo` (rash)
`samaa > sam + oo > samoo > sammoo` (same)
`hyvyys > hyv + yys > hyvyys > hyvvyys` (goodness)

Kitee dialect has also a convention that a vowel is repeated in a consonant cluster after the first consonant (8).

(8)
`halko > halako` (log)
`kilpa > kilipa` (race)
`hölmö > hölömö` (fool)
`helppo > heleppo` (easy)

It might be possible to handle these processes with global rules. However, the danger is that the rules will over-generate. Especially for newer words in language, the rules do not apply. Therefore, I have described these processes in the morphological dictionary, where it is possible to control the conversion for each word.

**5 Competing structures**

The consonant doubling rule in the environment VCVV has a strong precedence, although it cannot be applied globally. There are interesting cases, where the dialect has two alternatives for modifying the standard language. Examples are in (9).

(9)

```
kivääri > *kivvääri[2]
      kiväri

kritiikki > *rittiikki
      ritikki

musiikki > *mussiikki
      musikki

ovaali > *ovvaali
      ovali

pinaatti > *pinnaatti
      pinatti

poliitikko > *polliitikko
      politikko

poliisi > *polliisi
      polisi

pokaali > *pokkaali
      pokali

tomaatti > *tommaatti
      tomatti
```

In (9) above, there is a double vowel in Standard language, which would require that the preceding consonant is doubled. Instead of doing this, the dialect shortens the vowel cluster and thus avoids the doubling of the consonant.

Another very common structure occurs in nouns that are derived from verb stems. Examples are in (10).

(10)

```
ajoitus > *ajjoitus (timing)
      ajotus

majoitus > *majjoitus (accommodation)
      majotus

häväistys > *hävväistys (disgrace)
      hävästys

valaistus > *vallaistus
      valastus
```

[2] Asterisk '*' in front of the word here means that the word is ungrammatical.

```
kaavoitus > *kuavvoitus
      kuavotus

kukoistus > *kukkoistus
      kukostus
```

But his rule is not global, as we see in (11).

(11)

```
aseistus > asseistus (weaponry)
      *asestus

koneistus > konneistus (machination)
      *konestus

yleistys > ylleistys (generalisation)
      *ylestys
```

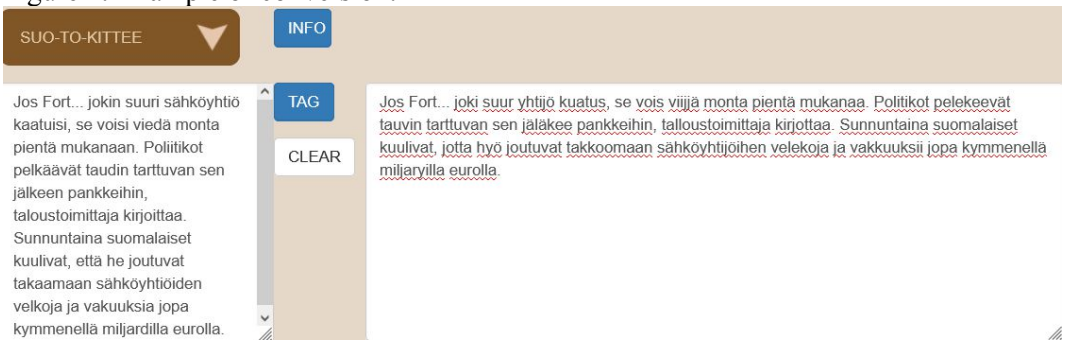
Another strategy for avoiding the application of the consonant doubling rule is to insert a consonant between vowels (12).

(12)

```
tatuointi > *tattuointi
      tatuvointi

ariointi > *arvviointi
      arvijointi

asiointi > *assiointi
      asijointi

kopiointi > *koppiointi
      kopijointi

partiointi > *parttiointi
      partijointi
```

Note that the intervening consonants *j* and *v* are not real consonants. They just help the speaker to survive the uneasy vowel sequence. At the same time, they help avoiding consonant doubling.
An extract from a news text is displayed in Figure 1. On the right side is the Kitee version.

Figure 1. Example of conversion:




## 6 Conclusion

It is tempting to construct a dialect converter using global rules, where certain sequences of characters in standard language are converted to meet the conventions of the target dialect. The paper shows that this is not possible, because the rules are not based only on character sequences. It depends on the age of the words in language, and also on their use frequencies. The older words are more likely to undergo heavy alterations than newer words. Also, frequent words change more easily than seldom used words. All this is very natural, and it is probably a common phenomenon in most languages.