14

Consonant Gradation in Estonian and Sámi: Two-Level Solution

TROND TROSTERUD AND HELI UIBO

14.1 Introduction

Koskenniemi's two-level morphology was the first practical general model in the history of computational linguistics for the analysis of morphologically complex languages. In this article we will reconsider one of the key innovations in Koskenniemi (1983), namely the treatment of consonant gradation in finite state transducers. We will look not at Finnish, but at two languages with a more extensive consonant gradation system, namely Estonian and Sámi. The goal of the paper is to demonstrate two different ways of modeling consonant gradation in a finite state morphological system - lexical and morphophonological. We will also compare the resulting systems by their computational complexity and human-readability.

Consonant gradation is rare among the languages of the world, but stem alternation in itself is not, and the treatment of consonant gradation can readily be transferred to other stem alternation phenomena. Koskenniemi's original idea was to see stem alternation as an agglutinative phenomenon. Consider the example (14.1), showing a two-level representation of stem alternation.

$$ehTe\$: ehe \tag{14.1}$$

Here the $ sign is a quasi-suffix, introduced to trigger consonant gradation in the stem. Two-level rules decide the correspondence of T to surface phonemes t or 0 (empty symbol), based on the context, specifically, according to the presence or absence of the symbol $ in the right context.

Another type of rules for handling stem alternations that can be compiled

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

into finite state automata is the method of sequentially ordered replace rules presented in Karttunen (1994) which have the format shown in (14.2):

$$a -> \ b \,||\, LC \ \ RC \tag{14.2}$$

The rule should be interpreted like "*a* is replaced by *b* in the left context *LC* and right context *RC*". The main practical preference of the replace rules compared to two-level rules is that they can handle a segment consisting of several characters as a whole, whereas handling the change of a character string by two-level rules requires several rules to be co-ordinated, one for each character alternation. This is, for instance, useful for building the additive forms for some inflection types in Estonian where the lemma has duration I, but the singular additive form has duration III. This case is especially difficult for stops, the III grade of which is not built by simple doubling but the double consonant is different from the corresponding I grade phoneme, cf. the nominative and short illative (additive) forms *rida:ritta* ’row’, *tuba:tuppa* ’room’, *nägu:näkku* ’face’.

This kind of change can be handled by **one** replace rule (14.3) but requires **two** two-level rules. In two-level rule system it also requires the introduction of new lexical symbols (=, 2) which invoke the rules in (Figure 14.1).

$$d -> \ tt \,||\, V \_ V2 :; \tag{14.3}$$

```
WeakStop:Stop <=> _ %=: (StemVowel:) 2:;
     where WeakStop in (g G b B d D)
               Stop in (k k p p t t)
     matched;

%=:Stop <=> WeakStop: _ (StemVowel:) 2:;
     where WeakStop in (g G b B d D)
               Stop in (k k p p t t)
     matched;
```

FIGURE 1 Handling grade alternation I-III by two-level rules

On the other hand, the strong preference of two-level rules is that there should not been defined any ordering on them as they work independently from each other. We have made some experiments with replace rules in the early stage of building Estonian finite state morphology. Based on our experience it was quite difficult to write a consistent replace rule set, as some higher-priority rules often spoiled the contexts for some of the lower-priority rules. Writing a **consistent two-level rule set** turned out to be considerably easier.

These advantages / disadvantages have the same source. As the empty symbols are coded as real zeros in the lexical level of the two-level model, the two-level rules (as finite state transducers) can be intersected. And this approach does not allow "unequal" changes like b -> pp, where one character is replaced by two. Replace rules can replace a segment of arbitrary length by another segment of arbitrary length. On the other hand, replace rules cannot be intersected, but should be applied sequentially instead.

More or less the same advantages have also been pointed to in Karttunen (2001):

> From the current point of view, two-level rules have many interesting properties. They are symbol-to-symbol constraints, not string-to-string relations like general rewrite rules. Two-level rules make it possible to directly constrain deletion and epenthesis sites because the zero is an ordinary symbol. Two-level rules enable the linguist to refer to the input and the output context in the same constraint.

From a formal point of view there is no substantive difference: a cascade of rewrite rules and a set of parallel two-level constraints are just two different ways to decompose a complex regular relation into a set of simpler relations that are easier to understand and manipulate Karttunen (2001). Thus, it is more like matter of taste, i.e. what kind of rule system seems better to grasp for the individual language engineers. We have opted for two-level rules but that does not mean we exclude the possibility of using replace rules at all. At the moment only two-level rules are used in the description of Estonian morphology but it is possible to code b -> pp, d -> tt etc. rules as replace rules instead.

In their historical overview of the development of finite state transducers, Karttunen (2001) pointed out that one problem with two-level transducers in the early years was connected to hardware limitations:

> It was also known from the beginning that a set of two-level transducers could be merged into a single one (...) by intersecting them. The problem is that in both cases the resulting single transducer is typically huge compared to the sizes of the original rule networks. Composition and intersection are exponential in the worst case. That is, the number of states in the resulting network may be the product of the sizes of the operands. Although the worst case is purely theoretical, in practice it turned out that intersecting large two-level rule systems was either impossible or impractical on the computers available in the early 90s.

In the present article we would like to look at the efficiency issue of two-level rule systems again, in the light of the processor speed of contemporary computers.

14.2 Method

In addition to giving a descriptive overview of the finite state systems of Estonian Northern and Lule Sámi, we are also going to give some characteristic numbers in order to measure the rule sets and lexicons of each system, thereby givingsome ground for comparison.

We will compare the morphological transducers of Northern Sámi, Lule Sámi and Estonian. Linguistically, we may say that Estonian and Sámi are similar as regard to the number of stem variants for the words with consonant gradation - usually two, but in some cases there are even four. While comparing we have to bear in mind that the systems are in different stage of development, as regard to their lexical coverage.

In section 14.5.1, we will use the following units of measure:

- number of records per lexical unit in stem lexicon;
- number of continuation lexica per lexical unit;
- number of states and arcs in the resulting morphological transducer (which is the composition of lexical and rule transducers);
- time of compilation of the rule (and lexicon) transducer.

14.3 Consonant gradation types in Estonian and Sámi

14.3.1 Consonant gradation types in Estonian

There are three different phoneme durations in Estonian (I - short, II - long and III - extra long). In written form the durations II and III are identical (written as a double vowel/consonant or a cluster of 2-5 consonants or two vowels), except for the stops where there are three different written forms as well (I - *g, b, d,* II - *k, p, t,* III - *kk, pp, tt*). There are two principally different consonant gradation types in Estonian - qualitative and quantitative.

1) Qualitative changes

1a) deletion of a stop (g, b, d, k, t) or s (table 1).

1b) assimilation (*kandma : kannan* ’to carry’, *vars : varre* ’stalk’);

1c) replacement of a weak stop by rules b:v, d:j, g:j (*kaebama : kaevata* ’to complain’, *rada : raja* ’path’, *märg : märja* ’wet’);

arg	:	ara	fearful	käskida	:	käsin	to order
tuba	:	toa	room	ehte	:	ehe	adornment
vedama	:	vean	to transport	mesi	:	mee	honey
uskuda	:	usun	to believe				

TABLE 1 Deletion of g, b, d, k, t, s

Additionally, in some inflection types with the gradation type 1a) the singular additive form is in duration III (cf. the examples presented in section

14.1). The same occurs in another inflection type where in all other cases the stem remains unchanged (sg nom *pesa* ’nest’, sg gen *pesa*, sg part *pesa*, sg addit *pessa*).

2) Quantitative changes

2a) alternation of long and short geminate (table 2, I column);

2b) alternation of strong and weak stops (table 2, II column);

kk : k	pikk : pika	long	k : g	vilkuda : vilgub	twinkle
pp : p	sepp : sepa	smith	p : b	kubjas : kupja	taskmaster
tt : t	võtta : võtan	to take	t : d	kartma : kardan	to be afraid
ss : s	kirss : kirsi	cherry			

TABLE 2 Estonian quantitative gradation

Estonian differs from Finnish, where consonant gradation is a weakening process only, in also having some noun inflection types with strengthening quantitative consonant gradation, although the weakening consonant gradation is considered the main type of consonant gradation.

Weakening consonant gradation is defined as follows:

- nouns: sg nom (sg part) - strong grade, sg gen - weak grade
- verbs: supine (primary form) - strong grade, indicative mode present tense - weak grade

In the paradigms of words with strengthening consonant gradation the strong and weak grade stems occur just conversely.

The strengthening consonant gradation types of nouns are the following:

a) nouns that derived from a verb with consonant gradation, e.g.: *hinne : hinde* ’mark’ (verb *hindama* - *hinnata* - *hindan* ’to evaluate’)

b) nouns that end with s and are in weak grade in singular nominative, but singular genitive is in strong grade and the final s is deleted, e.g , *saabas : saapa* ’boot’.

c) nouns that end with vowel + r (*vaher : vahtra* ’maple’, *tütar : tütre* ’daughter’)

d) nouns that additionally to the gradating stem have stem final change e-me (*liige : liikme* ’member’, *võti : võtme* ’key’)

There are no verb inflection types with strengthening consonant gradation in Estonian.

14.3.2 Consonant gradation in Sámi

In essence, Sámi consonant gradation is a phenomenon quite similar to its Finnish and Estonian counterpart. The consonant cluster on the border of the final and antepenultimate syllables of the stem, may change, i.e. the consonant cluster has different grades. Typically, there are two grades, strong grade and weak grade. Sámi and Estonian are among the few gradation systems with three grades, but in most cases the grade alternation is binary, i.e. III-II or II-I. Thus, a grade II consonant cluster may be strong relative to a grade I cluster, and weak relatively to a grade III cluster.

Historically speaking, strong grade was found in the consonant at the onset of the final or penultimate open syllable in a stem, whereas the grade changed to weak when inflectional processes closed the final open syllable (and vice versa, for consonant-final stems). In the modern languages, the triggering environment for consonant gradation is (inflectional or derivational) morphology.

We will look at the consonant gradation pattern of Lule Sámi, and in some special cases also at the pattern found in Northern Sámi. Linguistically speaking, they have the same consonant gradation system, but this pattern is represented in different ways in the respective orthographies of the two languages. Since the automata presented in this article generalise over written language, rather than over phonological representations, Lule and Northern Sámi consonant gradation must be treated as being more different than they are in the spoken language.

One difference is that Lule Sámi makes more use of digraphs, i.e., instead of *š* they write *sj*. In the two-level morphology formalism, each alternating symbol must get its own rule (many-to-many alternation is not allowed), this calls for more rules than the Northern Sámi gradation. On the other hand, in Lule Sámi going from strong to weak grade is a uniform process, letters are either changed or deleted, whereas in Northern Sámi letters may be either changed, deleted or added. Compare the following parallel forms, where the strong-weak alternation is denoted as xy : xyy in Northern Sámi, and as xyy : xy in Lule Sámi. Linguistically speaking, the gradation is in both cases of the same type *xøy:xy*, where ø = schwa.

‘stone’	nominative	genitive
Northern Sami	geađgi	geađggi
Lule Sámi	giergge	gierge

TABLE 3 Northern Sámi xy : xyy and Lule Sámi xyy : xy

We will first look at quantitative gradation, and then at qualitative gradation. Finally, we will look at a mixed type.

Quantitative alternation involves fricatives, liquids and voiced nasals, 34 alternating pairs in Northern Sámi and 57 alternating pairs in Lule Sámi. In one subtype of the qualitative gradation, grade I is written with single consonant, and grade II with double consonant. In the standard orthography, grade

III is also written with double consonant. In earlier orthography, grade III was written with an apostrophe, and in order to give a linguistically adequate representation, our transducer also accounts for III-II gradation, although it is not visible in output mode. In future applications involving text-to-speech, the orthographically invisible III-II alternation will become relevant, as we via disambiguation will be able to predict the correct grade of nominative (grade III) and accusative/genitive (grade II) nouns, although they are written identically. An example from Northern Sámi is *oad̄'d̄i : oad̄d̄it : oad̄án* 'sleeper : to sleep : I sleep'.

A single, nongeminate letter may also be deleted. There are no examples in Sámi of an intervocalic consonant being deleted (of the Estonian type $tuba : toa$, but consonants that are part of consonant clusters may be deleted, as in the Lule Sámi pairs *jiegńa : jieńa* 'ice Nom:Gen', *spádnjo : spánjo* 'birch forest Nom:Gen'. Due to orthographical convention, one qualitative alternation in Lule Sámi is written as if it were a quantitative one, *htj:tj, hts:ts*, cf. *biehtse : bietse*, 'spruce Nom:Gen'. The other alternation belonging to this type are treated in section 14.4.2 below.

There are several types of qualitative consonant gradation. The simplest case is found in Lule Sámi, where one letter is changed into another one, like in *oakse : oavse* 'branch Nom:Gen' and *bákte : bávte* 'cliff Nom:Gen'. Only k undergoes this change, in consonant clusters with *s, t, tj* and *ts*. One-consonant changes are found in Northern Sámi as well, but with more complex context, as for the *rbm:rpm, rdn:rtn, rgη:rkη*, pairs, e.g. *fierbmi : fierpmi* 'net (Nom:Gen)'.

A different type of qualitative alternation is the $xx : yy$ type voice alternation where voiced stops and affricates change into unvoiced stops and affricates. The pairs are *bb/pp, dd/tt, gg/kk*, in Northern Sámi also *ddj/dj, zz/cc, žž/čč*, cf. Lule Sámi *oabbá : oappá* 'sister Nom : Gen'. Lule Sámi has the three latter alternations, but due to the different orthographical principles, they are written as *dtj:ttj, dts:tts*, and pattern with the *ks:vs* alternation, as far as two-level rules are concerned.

One type of qualitative consonant gradation is preaspirated stops and affricates change into their voiced counterparts. The II-I pairs in Northern Sámi are *hp:b, ht:d̄, hk:g, hdj:j, hc:z, hč:ž,* the Lule Sámi ones are *hp:b, ht:d, hk:g*. In grade III, the stops from grade II are doubled. Cf. the Northern Sámi series *ohcci : ohcat : ozan*, 'searcher : to search : I search', a three-grade inflection pattern. A corresponding example for Lule Sámi would be *jåhtte : jåhtet : jådåv* 'mover : to move : I move'.

14.3.3 Similarities and differences between Sámi and Estonian

We see that Sámi consonant gradation is more complex and variable than Estonian consonant gradation but there also exist a common part. Compared to the more well-known Finnish gradation, Sámi and Estonian are both more complex: More letters are involved in the alternation, a larger part of the lexicon is affected by the alternation. Contrary to Finnish, both Sámi and Estonian have a 3-way opposition, where the strongest grade III in certain cases is invisible in writing. The bulk of the alternation involves a binary opposition III/II or II/I, but there are also cases of III/II/I alternation within a single paradigm.

Typologically speaking, it is no accident that Estonian and Sámi differ from Finnish in another respect as well: Due to several apocopy processes, the segmental morphology has been shortened, and in many cases even disappeared (as in the important Genitive case). This has given consonant gradation a more prominent position in the grammar of Estonian and Sámi. Seen from a computational point of view, this typological difference is of no importance. In all three languages, consonant gradation is a non-segmental morphological operation, which must be triggered by elements introduced via the morphological process.

From the computational point of view, a more important difference is the non-existence of the stem final vowel in the lemmas of some noun types of Estonian: the stem vowel appears in inflected (genitive) stem but not in lemma stem. For handling this phenomenon, the two-level model provides us with a sufficient toolset. It is possible either to include the stem final vowel to into the lexical representation of the stem and force it to be deleted for singular nominative. And it is also possible to add the stem vowel to the inflected stem in a continuation lexicon. The morphological description of Estonian uses the second approach.

In the following section we will see how the consonant gradation processes have been described by the means of finite state morphology. The research have been done independently, thus coming to the similar solutions is incidental. And in some cases we have used different means to describe similar processes.

14.4 The finite state description of Estonian and Sámi morphology

14.4.1 Two-level morphology of Estonian

The morphological description of Estonian has been built up, lead by the principles of two-level morphology model (Koskenniemi (1983)). It consists of a network of lexicons and a set of two-level rules.

The two-levelness of the model means that the lexical representations of morphemes are maintained in the lexicons and the task of two-level rules is to "translate" the lexical forms into the surface forms and vice versa. The lexical forms may contain information about the phoneme alternations, about the structure of the word form (morpheme and compound boundaries) etc.

The most optimized system of inflection types of Estonian Viks (1992) includes 38 types - 26 noun types and 12 verb types. 14 noun types and 10 verb types are the types which have some kind of consonant gradation (including the types where the gradation is not visible in the written form). As the system deals with written language, only the inflection types with qualitative changes in stem and types with quantitative stop gradation are of our interest. There are 15 such inflection types according to Viks (1992) in Estonian – 10 noun types and 5 verb types.

The main principle in describing Estonian consonant gradation has been to keep the lexical representation as readable and meaningful as possible. We have used the capital letters K, P, T, G, B, D, S to mark the phonemes which undergo some kind of change (deletion, assimilation) in the inflection processes. Additionally, the character $ is used to mark the weak grade (similarly to Koskenniemi (1983)).

```
 Lexicon Nimisona                               Lexicon 18
hamBa  07_S-0;   poisS  23_I;                      TP_18at;
jalG   22_A;     riD=a  18_Adt_PlPV;            :$ TP_18an;
jıG=i  18_Adt;   siGa   18_PlPV;
laD=u  18_Adt;   sıD=a  18_Adt_PlPV;
laG=i  18_Adt;   teGu   18;
luG=u  18_Adt;   tiGu   18;
maDu   18;       tikK   22_U;
maG=u  18_Adt;   tekK   22_I;
manDEr 03_I;     tuB=a  18_Adt_PlPV;
paTj   24;       vahTEr 03_A;
```

TABLE 4 Presentation of the stems with consonant gradation in the root lexicon

Note that there is only lexical representation given in the root lexicon, not as it is usually done in lexical transducers (e.g. $tigu + S : tiGu18$) where lemma and morphological information are given on the left side of the transducer and lexical representation of the word-form is on the right side. The considerations for this kind of solutions are discussed in Uibo (2005). The capital vowels are subject to deletion synchronously with grade alternation. The symbol = is used to mark the consonants that are subject to gemination when building singular additive (corresponding rules given in section 14.3.1). The next level of lexicons (Lexicon 18 in Table 4) divides the word-forms between strong and weak grade referring to the corresponding continuation lexicons TP_18at and TP_18an. For the weak grade stems the $ sign is added at the end of the stem.

The two-level rules are convenient to handle phoneme alternations, concerning only one phoneme. If the stem change is more complex (e.g. *idu:eo*), then it can be handled analytically.

Let us consider an inflection type in Estonian, which is characterized by weakening stem inflection (the deletion of phoneme *b, d, g* or *s*) and also changes in the immediate neighborhood of the disappeared consonant - the lowering of the surrounding vowels.

Example list of words belonging to the type is given in Table 5.

madu	:	mao	snake	lugu	:	loo	story
siga	:	sea	pig	käsi	:	käe	hand
pidu	:	peo	party	nuga	:	noa	knife
tegu	:	teo	action	süsi	:	söe	coal
uba	:	oa	bean				

TABLE 5 Weakening stem inflection in Estonian

A rule for handling the deletion in Table 5 is found in Figure 2:

```
"b,d,g,s deletion"   ($ marks the weak grade)
  LV:0 <=> Vowel: _ Vowel: $:;
```

FIGURE 2 "b,d,g,s deletion"

The immediate right and left contexts of the deletion rule (figure 2) are identical (`Vowel:`), they refer to any underlying vowel. The rule for vowel lowering (figure 3) has two distinct contexts: the vowel lowering may occur before (*siga : sea*) or after (*madu : mao*) the consonant gradation, in the first case the consonant gradation is part of the right context, and in the latter case it is part of the left context.

```
HVow:LVow <=> Bgn _ LV: StemVow: %$: ;
              Bgn Vow: LV: _ %$: ;
              where HVow in (u ü i)
                    LVow in (o õ e)
                    matched ;
```

FIGURE 3 "Vowel lowering"

In the paper Uibo (2000) the stem flexion types and the discovery process of rules have been discussed in details. The most problematic morphophoneme in Estonian is *D* which may correspond to five different surface phonemes in weak grade: $D : 0$, $D : l$, $D : n$, $D : r$ and $D : j$. There the only way was to differentiate the correspondences by very detailed context. And luckily, the contexts do not overlap.

The number of consonant gradation rules in Estonian two-level morphology is 16 - this is the number of different lexical-surface character pairs that correspond to the weak grade (strong grade is considered default and weak grade - marked).

14.4.2 Finite state morphology of Sámi

Sámi consonant gradation is intertwined with many other morphophonological processes, such as stem vowel alternation and diphthong simplification. We use dummy elements to trigger the different morphophonological processes.

We will present two approaches to representing the consonant gradation types with two-level automata, the process-wise and the segment-wise approach, respectively. We will concentrate upon three types of alternations: The quantitative alternation (*ff:f*), the voicing alternation (*bb:pp*) and the seemingly inverted alternation *ig:igg* for the Northern Sámi schwa alternation.

14.4.3 Process-wise vs. segment-wise alternation

In the two-level formalism, we may generalise over either consonant gradation type (i.e., over context) or over alternating letter. We illustrate both options with an example from Lule Sámi. First we give a rule for the Lule Sámi consonant alternation $rgg : rg$, the rule in 4 (as a rule collapsing the 19 different consonant gradation patterns of this type that can be found in Lule Sámi). We may note that *g* takes part in another consonant gradation pattern as well, in the *g:ń* pattern in Figure 5, with 3 other consonant pairs.

```
Cx:0 <=> Vow: Cx _ Cy Vow ( StemCns: ) WeG: ;
where Cx in ( b d d g k l l l m m n p p s s s s ń ń )
      Cy in ( m j n ń n d j t b p d s t k m n t g k )
      matched ;
```

FIGURE 4 "Gradation Series 1, III-II, three-letter patterns"

```
Cx:0 <=> Vow: _ Cy Vow ( StemCns: ) WeG: ;
where Cx in ( b d d g )
      Cy in ( m j n ń )
      matched ;
```

FIGURE 5 "Gradation Series 1, II-I, two-letter patterns"

Alternatively, one may choose to analyse the alternation in question not as a generalisation over multiple types, but as a generation over multiple contexts, i.e. write one rule for each of the 10 consonant that are involved in the 23 alternation patterns of the 2 rules above. The result, for g, may be seen in Figure6:

```
g:0 <=> Vow: ( g ) _ ń Vow: (StemCns: ) WeG: ,
        Vow: [ j|l|r|v ] _ g Vow: ( StemCns: ) WeG: ;
```

FIGURE 6 “Consonant gradation g:0”

The Lule Sámi consonant gradation was analysed in both ways. Ordered according to context, the set contains 20 rules, ordered according to alternating consonant, it contains 25 rules. When ordered according to alternating consonant, each rule contains appr 4 subrules, thus the total number of rules in the latter approach is 74.

The computational difference between the two is that ordered according to context, the rule set contains a large number of conflicting contexts, who must be resolved by the parser. The parser is good at it, but it takes time, a quarter of an hour on a not too fast machine, as a matter of fact. Comparing the compilation time between the two rule sets (on a 400 MHz Power Mac G4), we se a huge difference, cf. Table 6.

Rule set ordered	# of	# of	Compilation time		
according to:	rules	subrules	real	user	system
alternating consonant	4	16	0m11.013s	0m1.140s	0m0.240s
context	4	4	16m14.387s	3m8.250s	0m7.430s

TABLE 6 Compilation time

We discuss the compilation issue at the end of paragraph 14.5.1.

Schwa alternation

This alternation was represented in Table 3 above. In Lule Sámi, this alternation may be analysed in the same way as the quantitative alternation type $xyy : xy$ found in pairs like *liehppa:liehpa* ’shelter Nom:Gen’. The phonological realisation is different in the two cases, but from a computational point of view, this is irrelevant. In Northern Sámi, the gradation in question is written $xy : xyy$, and here this alternation must be analysed in a different way. The method chosen was to represent the strong grade underlyingly as $x'y$, and to replace the apostrophe with the consonant to the right in the weak grade, and then to prevent any apostrophe from the surface representation.

14.5 Comparing the treatments

14.5.1 Consonant gradation as lexeme property or as lexicon property

Originally, consonant gradation was a phonological alternation, which affected phonologically defined consonant clusters in phonologically defined environments. As we have seen, the environments have now become morphological, and must be treated as such. When it comes to the gradating consonants themselves, the situation is not that clear. Most of the phonologically appropriate stems undergo consonant gradation, but not all of them do. The gradation types are not equally regular, in Estonian, for example, the qualitative gradation forms a closed class, whereas the some types of the quantitative consonant gradation are regular and productive.

In principle, there are two ways of dealing with this:

1. Alternating and non-alternating consonant clusters are not distinguished in the lexicon, rather, they are directed to different sublexica, and treated differently there.
2. Alternating and non-alternating consonant clusters are pointed towards the same sublexica, hence they have the same morphology. The difference is found in the stem, where the consonant clusters are given different archiphonemes. Either the alternating or the non-alternating consonant may be given the special phoneme.

The Sámi words *goahti* ’hut’ and *stáhta* ’state’ both contain the consonant cluster -*ht*-. The former alternates with -*đ*-, and the latter does not. This difference may be handled in two ways, denoted a and b in Figure 7.

```
a. Directing gradating and non-gradating to different lexica
    LEXICON NounStems
    goahti GRADATING-BISYLL-NOUN  ;
    stáhta NONGRADATING-BISYLL-NOUN  ;

b.i One continuation lexica, but marking the gradating noun
    LEXICON NounStems
    goahti:goahTi  BISYLL-NOUN  ;
    stáhta  BISYLL-NOUN  ;

b.ii One continuation lexica, but marking the non-gradating noun
    LEXICON NounStems
    goahti  BISYLL-NOUN  ;
    stáhta:stáhTa  BISYLL-NOUN  ;
```

FIGURE 7 Two strategies for continuation lexica

In a., the subsequent lexicon GRADATING-BISYLL-NOUN would contain a consonant gradation trigger not present in NONGRADATING-BISYLL-NOUN. In b.i, we would have a morphophonological rule changing

T to đ, and another rule deleting h in front of a T:đpair. Solution b.ii would be the mirror image of b.i, having gradation as the default case, with a rule deleting all instances of t (but not T) in the context h_V, for the relevant word-forms, and with a rule rewriting T as t in all contexts.

Whether to choose b.i or b.ii is a matter of taste. If consonant gradation is the rule and not the exception, it is of course tempting to treat the exceptions as such. On the other hand side, giving gradation the special treatment makes it easier to control: Gradation occurs where we have said that it should, and nowhere else. Both Koskenniemi (1983) and the present treatment of Estonian thus chose the b.i option.

The Sámi solutions presented here opt for alternative a. This gives a simpler stem lexicon but more complicated continuation lexica. And indeed, the Northern Sámi transducer has 250 continuation lexica for the noun, adjective and adverb complex, as compared to the somewhat lower 164 for Estonian. Note that the number of continuation lexica is also dependent upon the coverage of the transducer. The Lule Sámi transducer has a weaker coverage for derivational processes, and a somewhat more regular adjective declension pattern, and here the number of continuation lexica is 108.

For our Southern Sámi transducer we have chosen option b.i, and although the numbers cannot be compared directly (Southern Sámi does not have consonant gradation, but its Umlaut phenomenon is of compatible size and complexity), it contains only 29 continuation lexica.

	records per	root				Processor	Compilation time	
language	lex unit	lexicon	states	archs	paths	MHz	rule trans	lex trans
Estonian	109/55=1.98	400	1,940	5,009	circular	700	3s	0s
Lule S.	166/48=3.45	760	2,413	4,722	755,374	1,400	0.5s	1.6s
North. S.		75,294	95,652	258,350	circular	1,400	1m 6.2s	2m 38.9s
Lule S.						400	6.2s	12.2s
North S						400	5m 6.5s	7m 25.6s

TABLE 7 Comparing the compilation of Estonian and Sámi

In evaluating the results shown in Table 7, one has to be aware that the rule sets have been built based on different principles - Northern Sámi uses the context-oriented approach, whereas in the rule sets for Estonian and Lule Sámi each rule handles a concrete pair and lists all the possible contexts disjunctively on the right side of the rule. There are lots of formal conflicts in the Northern Sámi rule set, which is reflected in the compilation time. As long as compilation time is not a critical factor, the context-oriented approach of Northern Sámi is fine, but writing two-level rules relative to the alternations will reduce compilation time drastically.

14.6 Conclusion

The two-level morphology compiler *twolc* is fully capable of handling even large and complex grammars. However, in a longer perspective we could try to combine two-level and replace rules (another tool from the Xerox finite state package – *xfst* – can be used for that purpose), as some kind of rules are more convenient to be handled by replace rules.

Non-segmental morphology may be handled by abstract, segmental rule triggers.

If compilation time is a factor, then context conflicts should be resolved before compilation. Still, even for large rule systems compiled on machines as slow as 400 MHz, this only gives 7 min as compilation time. During a developmental phase this may be a nuisance, but it can be lived with. And better source code reduces the compilation time to seconds.

We have shown that the finite state system of lexicons and rules both of which are computationally finite state transducers is very flexible: the system builder can choose if (s)he wants to describe a certain phenomenon by rules or by lexicons. As a rule of thumb, stem changes are more likely to be described by rules and morpheme combination rules by lexicons, but as we have seen, some types of the stem changes can be more naturally described by continuation lexicons.

References

Karttunen, Lauri. 1994. Constructing lexical transducers. In *15th International Conference on Computational Linguistics (COLING-94)*, pages 406–411. Kyoto, Japan.

Karttunen, Lauri. 2001. A short history of two-level morphology. *Sámi diedalaš áigečála* 3:100–123.

Koskenniemi, Kimmo. 1983. *Two-level Morphology: A General Computational Model for Word-form Production and Generation*. Publications of the Department of General Linguistics, University of Helsinki. Helsinki: University of Helsinki.

Uibo, Heli. 2000. Kahetasemeline morfoloogiamudel eesti keele arvutimorfoloogia alusena. In *Tartu Ülikooli Üldkeeleteaduse Õppetooli toimetised 1: Arvutuslingvistikalt inimesele*, pages 37–72.

Uibo, Heli. 2005. Optimizing the finite-state description of Estonian morphology. In *15th Nordic Conference on Computational Linguistics, NoDaLiDa 2005*.

Viks, Ülle. 1992. *A concise morphological dictionary of Estonian*. Tallinn: Eesti keele instituut.