Complexity, Two-Level Morphology and Finnish

Kimmo Koskenniemi
Kenneth Ward Church
Coling 88

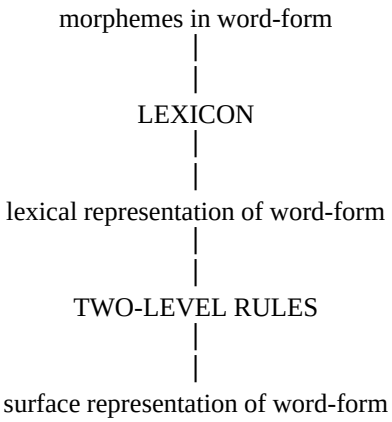
Abstract

Although, Two-Level Morphology has been found in practice to be an extremely efficient method for processing Finnish words on very small machines, [Barton86] has recently shown the method to be NP-hard. This paper will discuss Barton's theoretical argument and explain why it has not been a problem for us in practice.

1. The Two-Level Model

The two-level model provides a language independent framework for describing phonological and morphological phenomena associated with word inflection, derivation and compounding. The model can be expressed in terms of finite-state machines, and it is easy to implement. The model has, in fact, two aspects: (1) it is a linguistic formalism for describing phonological phenomena, and (2) it is a computational apparatus which implements descriptions of particular languages as operational systems capable of recognizing and generating word-forms.

The model consists of three representions (morphological, lexical and surface forms) and two systems (the lexicon and phonological rules) relating them:



The surface representation is typically a phonemic representation of word-form, but sometimes graphic or written forms are used instead. The lexical representation is an underlying (postulated) morphophonemic representation of of the word stem and affixes. These two representations need not be identical, and in case there are phonological alternations in the language, these representations are more or less different. The task of the two-level rule component is to account for any discrepancies between these representations.

The task of the lexicon component is two-fold. First, it specifies what kinds of lexical representations are possible according to the inventory of known words and their possible inflectional forms, plus derivations and compounds according to productive rules. The second task of the lexicon is to associate proper morphemes to lexical representations. The task of the lexicon component is considered to be universal.

Many languages can be quite well described with rather simple lexicon structures. The lexicon needed for Finnish is basically a set of sublexicons (for stems, case endings, possessive suffixes, clitic particles, tense of verbs, person, etc.). Each entry specifies all continuation lexicons which are possible after that morpheme. This scheme is equivalent to a (partly nondeterministic) finite state transition network.

Two-level rules compare lexical and surface representations. The partitive plural of the Finnish word lasi 'glass' is laseja. This form might be represented as a stem lasi plus a plural ending I plus a partitive ending A. The correspondence would be then be:

l a s i I A	lexical
l a s e j a	surface

There are three discrepancies here: the stem final i is realized as e (and not as i like in singuler forms), the plural I is realized as j instead of i, and the partitive A is realized as the back vowel a (and not as front vowel ä). The first discrepancy is

described with a two-level rule:

i:e <=> _ I:

This states that lexical i is realized as surface e if and only if it is followed by a lexical I (the plural affix). The plural I itself is a bit different from other i's because it is realized as j if and only if it occurs between two surface vowels (let V denote the set of vowels):

I:j <=> :V _ :V

The realization of partitive A is an instance of Finnish vowel harmony, which causes endings to agreee in frontness or backness with stem vowels. Thus A has two possible realizations: it must be a back vowel iff there are back vowels in the stem:

[A:a | O:o | U:u] => :Vback :Vnonfront* _

The set Vback contains the back vowels a, o, and u whereas Vnonfront contains anything that does not have one of ä ö u¨ on surface.

Phonological two-level descriptions have been made for about twenty different languages up to now. Only about a third of them can be considered to be comprehensive. Typically a description consists of 7-40 rules (English and Classical Greek being the low and high extemes).

A special compiler is used for converting these rules into finite state transducers (Karttunen, Koskenniemi, and Kaplan, 1987). The resulting machines are similar to the ones that were hand compiled, eg. in (Koskenniemi, 1983).

2. Barton's Challenge

[Barton86] poses a challenge to find the constraint that makes words of a natural language easy to process:

''The Kimmo algorithms contain the seeds of complexity, for local evidence does not always show how to construct a lexical-surface correspondence that will satisfy the constraints expressed in a set of two-level automata. These seeds can be exploited in mathematical reductions to show that two-level automata can describe computationally difficult problems in a very natural way. It follows that the finite-state two-level framework itself cannot guarantee computational efficiency. If the words of natural languages are easy to analyze, the efficiency of processing must result from some additional property that natural languages have, beyond those that are captured in the two-level model. Otherwise, computationally difficult problems might turn up in the two-level automata for some natural language, just as they do in the artificially constructed languages here. In fact, the reductions are abstractly modeled on the Kimmo treatment of harmony processes and other long-distance dependencies in natural languages.'' [Barton86, p56]

We suggest that words of natural languages are easy to analyze because morphological grammars are small. As Barton shows, two-level complexity grows rapidly with the number of harmony processes. But, fortunately, natural languages don't have very many harmony processes.

Any single language seems to have at most two harmony processes:

- zero (most, ie. some 88 % of languages),
- one (Uralic, Tungusic, Sahaptian) or
- two (most Altaic languages)

Even in principle, a three dimensional vowel harmony is rather improbable, because it would lead to a total (or almost total) collapse of distinctions between vowels. In most languages there are not enough distinctive features in vowels to make a four-way harmony even possible. We have not found any reliable accounts for more than two harmony-like processes in a single language.

Normally, most complexity results describe space/time costs as a function of the size of the input. Claims in support of the two-level model are generally of this form; speed is generally measured in terms of numbers of letters processed per second. Barton's result is somewhat non-standard; it describes costs as a function of the size of the grammar (or more precisely, the number of harmony processes). Complexity results generally don't discuss the ''grammar constant'' because any particular grammar has just a fixed (and very small number) of rules (such as harmony processes), and thus it isn't very helpful to know how the algorithm would perform if there were more, because there aren't.

If phonological grammars were large and complex, there could be efficiency problems because processing time does depend on the size and structure of the grammar. However, since phonological grammars tend to be relatively small (when compared with the size of the input), it is fairly safe to adopt the grammar constant assumption.

3. Barton's Reduction

Let us consider the satisfaction reduction in [Barton86]. Barton used a grammar like the one below to reduce two-level generation to the satisfaction problem.

In this artificial grammar, it is assumed that there are an arbitrary number of harmony processes over the letters: a, b, c, d, e, f, ...; each letter must correspond to either T (truth) or F (falsehood), consistently throughout the word. This reduction is a generalization of harmony processes which are common in certain families of natural languages. In these languages, stem (and affix) vowels must agree in one or more of the following distinctive features:

- Front/back vowels (palatal, velar harmony), eg. in Uralic and Turcic languages. (Replaced by consonantal palatalization in Karaite, a Turcic language.)
- Rounded/unrounded vowels (labial harmony), eg. in Turcic languages
- Tongue height, eg. Tungusic languages
- Nasalization, and
- Pharyngealization eg. emphatic consonants and vowels in semitic languages

Some processes are classified as umlaut rather than vowel harmonies, but behave similarly. One, still different but relevant process, has been reported in Takelma (Sapir 1922). There, a suffixal /a/ is replaced with an /i/, if the following suffix contains /i/. This rule derives [ikumininininink] from underlying /ikumanananink/.

It may be a mistake to classify all of these processes as vowel harmonies, and if so, it only strengthens the claim that languages don't have very many vowel harmony processes.

Barton's Satisfaction Grammar

```
T F , - a b c d e f
NULL 0
ANY  =
END

"a-consistency (or a-harmony)"   3  3
    a a =
    T F =
1:  2 3 1
2:  2 0 2
3:  0 3 3

"b-consistency (or b-harmony)"   3  3
    b b =
    T F =
1:  2 3 1
2:  2 0 2
3:  0 3 3

"c-consistency (or c-harmony)"   3  3
    c c =
    T F =
1:  2 3 1
2:  2 0 2
3:  0 3 3

d,e,f-consistency all follow the same pattern

"satisfaction"   3  4
    = = - ,
    T F - ,
1.  2 1 3 0
2:  2 2 2 1
3.  1 2 0 0
```

Empirically, we observe that generation time is linear with the length of the word and exponential with the number of harmony processes. That is, given Barton's Satisfaction grammar, words of the form *aaa...** are processed in time linear with the number of *a*s, but words for the form *abc...* are processed in time exponential with the number of different characters.

Linear with Input Length

Input	Steps in Generation
a	2
aa	4

aaa	6
aaaa	8
aaaaa	10

Exponential with Number of Harmony Processes

Input	Steps in Generation
a	2
ab	6
abc	14
abcd	30
abcde	62
abcdef	126

Barton showed that generating words in the two-level model with *n* harmony processes can be reduced to a satisfaction problem with *n* variables. Thus, it is not surprising to find that the two-level model takes time exponential with the number of harmony processes.[1]

1. **Most harmonies are progressive, ie. the harmony propagates from left to right. A few exceptions to this are mentioned in literature: Sahaptian (including Nez Perce), Luorawetlan (including Chuckchee), Diola Fogny, and Kalenjin languges. These are said to have so called dominant and recessive vowels where an occurrence of a dominant vowel in the stem or even in affixes causes the whole word to contain only dominant variants of vowels. We have found no references to languages with more than one harmony process combined with (potentially) regressive, or right-to-left direction.**

   Left-to-right harmony seems to have a virtually unlimited scope because, in addition to inflectional affixes, also derivational suffixes that can be recursively attached to the stem.

   Neither progressive nor regressive harmony-like processes cause any nondeterminism in recognition in the Two-Level Model. Even generation of word-forms with progressive harmonies is always quite deterministic. The only truly nondeterministic behavior with vowel harmonies occurs in the generation with regressive harmonies where there is no way to choose among possible realizations of prefix vowels until the word root is seen.

   An artifical (and almost maximal) example of the unbounded character of Finnish vowel harmony is the following where back harmony propagates from the verbal root (havai- 'observe') all the way to the last clitic particle (over seven derivational, one case ending and a possessive suffix):

   havai-nTO-llis-tU-ttA-mA-ttOm-UUTe-llA-nsA-kAAn
   havai nno llis tu tta ma ttom uude lla nsa kaan

4. Experience With Finnish

However, if there are only a fixed (and small) number of harmony processes, as there are in any natural language, then processing time is found to be linear with input length. This has been our experience as verified by the following experiment. We collected a word list and measured recognition time as a function of word length in characters. The word list is a combination of two samples from a Finnish newspaper corpus (seven issues of Helsingin Sanomat consisting of some 400,000 running words):

- all Finnish words with 17 or more letters in the whole corpus, plus
- some 700 words of running text from the same corpus.

(This construction produces very few words with 16 characters.)

Figure 1 plots recognition time (in steps) as a function of word length. Note that the relationship is well modeled by the linear regression line with a slope of 2.43 steps/letter. The data show no hint of an exponential relationship between processing time and word length.

One of the two outliers is ''lakiasiaintoimistoa,'' an 18 letter word that takes 206 steps (11.4 steps/letter). Part of the trouble can be attributed to ambiguity; this word happens to be two ways ambiguous. In addition, there is a false path ''laki+asia+into+imis...'' that consumes even more resources. The fit of the regression line can be improved considerably by removing these ambiguous words as illustrated in figure 2.

5. Conclusion

A disclaimer is in order. The two-level formalism does not guarantee efficient implementations as such; the formalism may be inappropriate for some problems (such as processing an unnatural language with hundreds or thousands of phonological processes). Moreover, the choice of two-level rules and lexical representations may affect performance. The formalism permits several styles of description (corresponding roughly to abstract, concrete or natural phonology, etc). Some may be more suitable than others for a

particular problem. More generally, finite state automata are not the solution to all problems; they are inadequate for some, and non-optimal for others.

However, the two-level model has made a significant contribution. It has enabled the construction of a comprehensive, efficient and compact morphological recognizer of Finnish with broad coverage, an important practical achievement that had not been accomplished before the introduction of the two-level model. To better understand why the two-level model is able to achieve broad coverage of Finnish with modest computing resources, and where the two-level model might break down, it is important to analyze time and space performance very carefully. In so doing, certain idealizations will need to be introduced. For instance, we have found it helpful to consider recognition time as a function of word length. Other idealizations are possible. Barton has discussed generation time as a function of the number of harmony processes, and by implication, the number of phonological processes in general. This idealization, in our opinion, is not helpful; it confuses the picture by considering a host of artificial languages that bear little resemblence to reality. Natural languages do not have very many phonological processes, but they do have a comparatively large number of words.

References

Barton, E., 1986, ‘‘Computational Complexity in Two-Level Morphology, in 24th Annual Meeting of the Association for Computational Linguistics.’’

Barton, E., 1987, Berwick, R. and Ristad, E., ‘‘The Complexity of Two-Level Morphology,’’ chapter 5, in Computational Complexity and Natural Language, MIT Press, Cambridge, MA.

Karttunen, L., Koskenniemi, K., and Kaplan, R., 1987, ‘‘A Compiler for Two-level Phonological Rules,’’ in Dalrymple, M., Kaplan, R., Karttunen, L., Koskenniemi, K., Shaio, S., and Wescoat, M., ‘‘Tools for Morphological Analysis,’’ Report No. CSLI-87-108, Center for the Study of Language and Information, Stanford University.

Koskenniemi, K., 1983, ‘‘Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production,’’ Publications No. 11, University of Helsinki, Dept of General Linguistics, Hallituskatu 11-33, SF-00100 Helsinki 10, Finland.

Sapir, E., 1922, ‘‘The Takelma language of southwestern Oregon,’’ in F. Boas (ed.), ‘‘Handbook of American Indian Languages II,’’ pp. 1-296, BAE Bulletin 40 (II), Washington.

Ultan, R., 1973, ‘‘Some Reflections on Vowel Harmony.’’ In ‘‘Working Papers in Language Universals,’’ Number 12, November 1973, pp. 37-67. (Language Universals Project, Committee on Linguistics, Stanford University, Stanford, California.)

Vago, R., (ed.), 1980, Issues in Vowel Harmony. (John Benjamins, Amsterdam.)

All Words

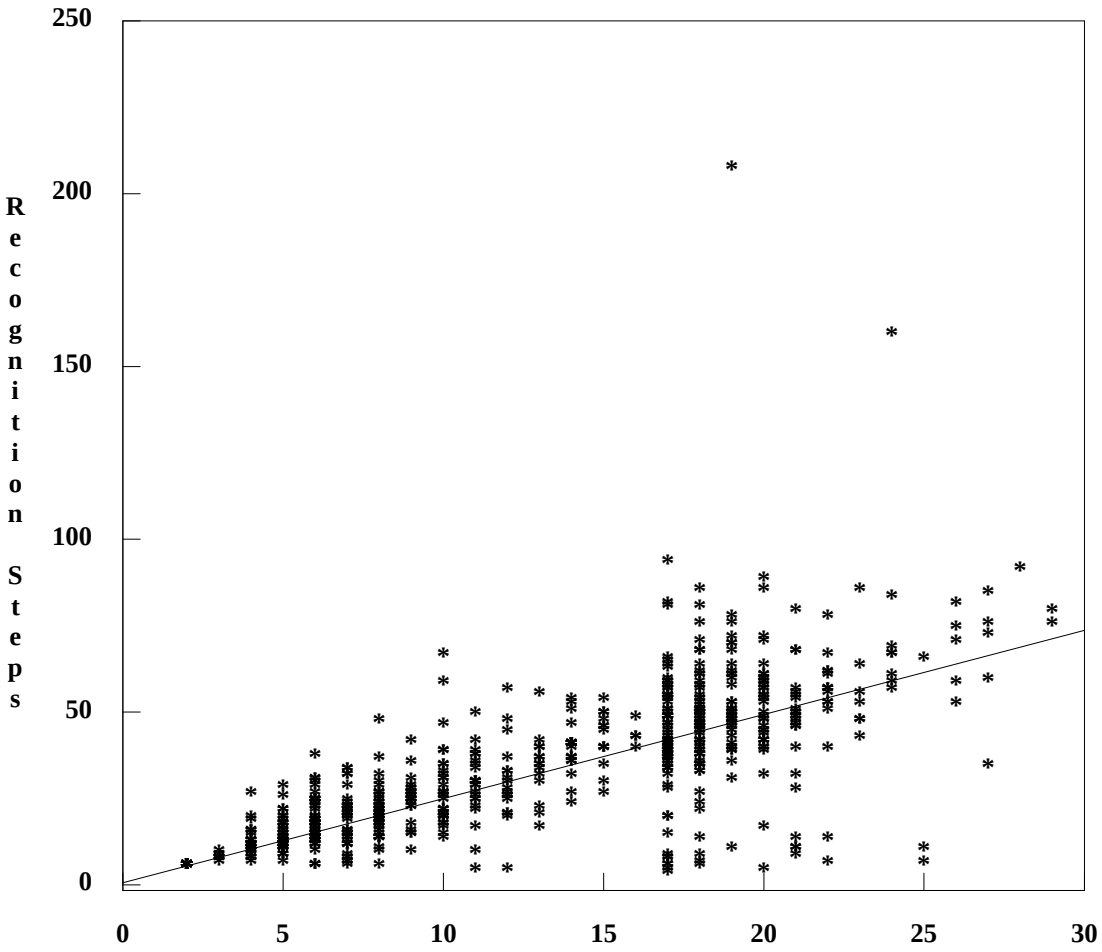


Figure 1

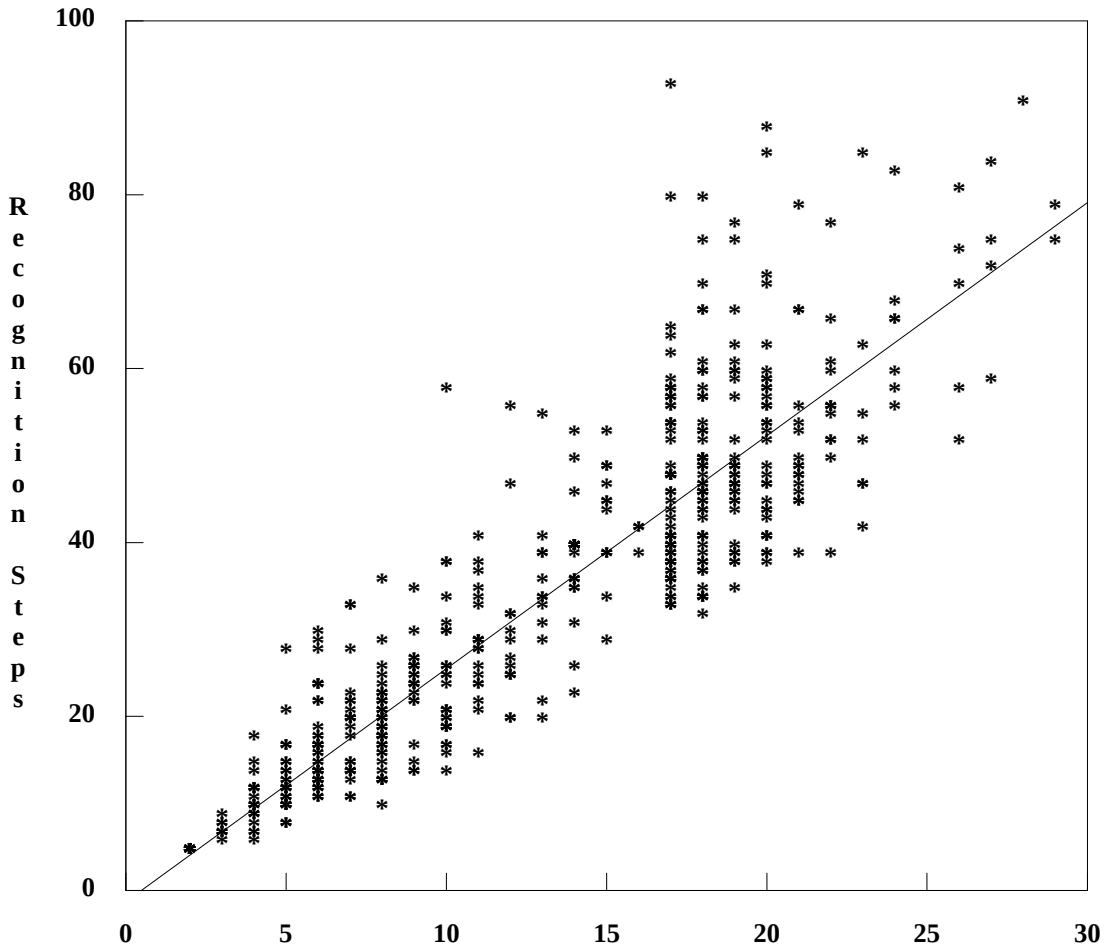


Figure 2

Appendix

	All Words		
Length	**Steps**	**Ratio**	**Instances**
2	4.00	2.00	26
3	6.38	2.13	8
4	10.69	2.67	29
5	12.71	2.54	42
6	16.22	2.70	60
7	16.62	2.37	39
8	19.08	2.39	48
9	22.92	2.55	25
10	26.28	2.63	32
11	27.13	2.47	23

12	27.94	2.33	18
13	32.21	2.48	14
14	38.15	2.73	26
15	40.38	2.69	13
16	41.75	2.61	4
17	40.95	2.41	82
18	44.25	2.46	65
19	53.17	2.80	40
20	49.64	2.48	36
21	42.30	2.01	23
22	49.67	2.26	15
23	54.86	2.39	7
24	76.00	3.17	8
25	26.00	1.04	3
26	66.00	2.54	5
27	63.80	2.36	5
28	90.00	3.21	1
29	76.00	2.62	2

For unambiguous words, the results are:

Unambiguous Words Only

Length	Steps	Ratio	N
2	4.00	2.00	26
3	6.38	2.13	8
4	9.65	2.41	20
5	11.89	2.38	28
6	15.80	2.63	40
7	18.09	2.58	22
8	18.26	2.28	38
9	22.62	2.51	21
10	24.12	2.41	25
11	27.84	2.53	19
12	29.31	2.44	13
13	33.42	2.57	12
14	37.64	2.69	25
15	41.67	2.78	12
16	40.00	2.50	3
17	45.83	2.70	59
18	47.43	2.63	49
19	49.67	2.61	30
20	52.52	2.63	31
21	51.62	2.46	16
22	55.67	2.53	12
23	54.86	2.39	7
24	64.29	2.68	7
26	66.00	2.54	5
27	71.50	2.65	4
28	90.00	3.21	1
29	76.00	2.62	2

Here are the data for words with two or more analyses (mostly two):

Ambiguous Words

Length	Steps	Ratio	N
4	13.87	3.47	8
5	15.50	3.10	12
6	20.06	3.34	16
7	20.70	2.96	10
8	26.25	3.28	8
9	30.00	3.33	3
10	37.67	3.77	6
11	42.00	3.82	2
12	35.00	2.92	1
13	35.00	2.69	1
14	51.00	3.64	1
16	47.00	2.94	1
17	61.00	3.59	4
18	59.80	3.32	5
19	92.20	4.85	5
20	55.50	2.77	2
21	54.00	2.57	1
24	158.00	6.58	1
25	64.00	2.56	1