JANNE BONDI JOHANNESSEN

Is Two-level Morphology a Morphological Model?

Abstract

This paper contains a close look at Koskenniemi's Two-level morphology from a linguistic point of view. The model will be compared to three other traditional, linguistic morphological models, IA, IP and WP. It will be shown that there are linguistic phenomena that can hardly be handled by some of the just mentioned models, and not at all in a linguistically satisfactory manner by the Two-level morphology.

1 Introduction

Koskenniemi's Two-level morphology (TM) has become well known since it was developed in 1983. One reason for this is probably that it is one of the few models within computational linguistics that has taken morphology seriously. To store full wordforms, inflected and derivated, in the lexicon may be possible for a language like English, with relatively poor morphology. But Koskenniemi saw that for Finnish, where a single verb can have between 12.000 and 18.000 different graphemic forms (included clitics), such a solution would not work. If the American computational linguists had been Red Indians speaking the Cherokee-language Oneida, instead of white and English-speaking, then they too would probably have developed a morphological model that could handle their verbs with up to 100.000 forms each.

I assume the Two-level morphology to be well-known, and I will thus only give a very short description of it, before I procede to the main task; to compare the Two-level morphology with other morphological models, and to see if this model can be said to be a morphological model.

2 A Short Description of Two-level Morphology

The Two-level morphology is designed to perform both analysis and synthesis on the basis of more or less the same data. It has at its disposal a rule module and

Proceedings of NODALIDA 1989, pages 51-59

a lexicon module. The rule module takes care of one-segment correspondences, mostly phonological ones. The lexicon module may consist of several lexicons, one or more for stems and others for affixes. From each lexicon there is a pointer to the next possible lexicons. The entries in the lexicon may look different from their surface representation, which the rule module takes care of. (1) and (2) are examples of lexicon entries in two sublexicons for Norwegian:

(1)
```
LEXICON Nouns
vintEr  /MNoun  Lexeme=WINTER
gutt    /MNoun  Lexeme=WINTER
```

(2)
```
LEXICON /MNounSg
0       /Genitive   Num=sg/Defin=ind/Gender=m/.
en      /Genitive   Num=sg/Defin=def/Gender=m/.
```

The information that we get about a word-form that is analyzed, is the information that is accumulated through all the lexicons that have been consulted. Thus if we analyze *vinteren*, we get the information from both the stem- and the suffix lexicon:

(3) `vinteren: Lexeme=WINTER  Num=sg/Defin=def/Gender=m/`

(This accumulation is the reason for the seeming zero-inflectional morph that is apparent in (2). It is not meant as a suffix, it is just there to ensure that the information about singular and indefinite is collected. This information could not have been represented in the stem lexicon, even if the stem is identical to the word form of indefinite singular, because the stem lexicon also points to lexicons for plural and definite forms. Since information is accumulated on its way through the lexicons, we would, if we had given the singular indefinite information in the stem lexicon, have gotten absurd results like *vintrene = singular, plural, indefinite, definite.* In other words: The stem lexicon can only include information that is common for all the wordforms belonging to one lexeme.)

The lexical form of the entry we have looked at is `vintEr` (1), but the surface representation should be as in (4), of course:

(4) vinter

The default alphabet then includes a lexical `E` that corresponds to a surface `e`, (`E:e`), (in addition to the usual `e:e`). The reason for this cumbersome representation is that *vinter* and many other Norwegian lexemes go through a morphophonemic change that deletes the `e` before certain morphological endings:

(5)
```
Singular:  vinter - vinteren
Plural:    vint_rer - vint_rene
```

If we want to have the same lexical entry for all wordforms of one and the same lexeme, which is obviously the most satisfactory solution from a linguistic point of view, we have to make the 'e' which can go away, a little different from other 'e'-s that can not be deleted (e.g. in *vinteren*), so that we can later formulate a rule that refers only to the appropriate 'e'. Only then can we keep one lexical entry for this lexeme, `vintEr`, instead of two, e.g. as in (6):

(6)
```
vinter      /MNounSg      Lexeme=WINTER
vintr       /MNounPl      Lexeme=WINTER
```

We then formulate a rule that overrules the lexical default values:

(7)
```
"E-deletion in stem before plural"
E:0 <= _ Liquid PlSuffix ;
```

(The rule context consists of names that refer to certain sets and definitions that we have predefined.)

3 How Can Two-level Morphology be Characterized When Compared to Other Linguistic Models for Morphological Description?

In our century traditionally there have been three models for morphological analysis; IA (Item and arrangement), IP (Item and Process) and WP (Word and paradigm). For a discussion of these models, see Hockett 1954, Matthews 1972, 1974, Robins 1970. A fourth model can also be mentioned, which I shall not go into here; NM (Natural morphology), see Wurzel 1982 or Bybee 1985. Below I shall compare the Two-level morphology with each of the three models. (The discussion will to a large degree be built on Johannessen 1988.) As they have not existed quite simultaneously, I will start with the oldest one and then end with the newer one.

3.1 Item and Arrangement

The main characteristic of this model is that there are minimal units, morphemes, that can be arranged in a number of ways to form bigger units. The morphemes are abstract units that are represented through their allomorphs. Since at the time of IA (approximately 1930–1950) the view held that syntax and morphology should be described in the same way; that there is ideally a one-to-one relationship between morpheme and allomorph, more precisely a relationship where one morpheme has one surface realization and vice versa:

(8) IA:

Morphemes:	Allomorphs:	'Word'	
{gutt} + {indef pl}	gutt-er	gutter	(= boys)
{hus} + {indef pl}	hus-0	hus	(= houses)

IA and TM have in common that the different elements are arranged lexically, as we see. But the elements of IA (morphemes) are abstract, so that in (8) we have the same second element in both words, it is just realized differently (different allomorphs). In TM on the other hand, the two plural formatives have nothing in common because of their different realization in the lexicon. In TM they are actually two different endings, since they are different graphemically:

(9) TM:

Lexical (stem) entries:	Lexical (affix) entries:	'Word'
gutt	er	gutter (= boys)
hus	0(nothing)	hus (= houses)

We do not see any morphosyntactic information here, since it is irrelevant for the model. The grammatical features that are present in the lexicon entries, can not be made use of by the rules. TM does not get past the concrete level of allomorphs, it can thus not be equivalent with the IA-model.

3.2 Item and Process

The IP model was popular until the 1960s. Like IA this too is a model based on the morpheme-allomorph distinction. The difference from the IA model is that the IP model allows processes, that is, it allows elements to undergo a metamorphosis to gain a shape different from the original one. This is possible both at phoneme and morpheme level. The model allows rules of both sorts.

When it comes to the process part, we can say that IP and TM have something in common. We have seen the rule part of TM, and even if there we deal with pairs of segments that correspond with each other in certain circumstances, the idea could be that the correspondence looks like a process. (In fact: The rule formalism is designed to take care of morphophonemic changes that are abundant in Finnish (vowel harmony and consonant gradation)). I can also cite Karlsson and Koskenniemi (1985:127): What is described by rules is "fairly natural one-segment modifications; mostly automatic, transparent, productive, exceptionless alternations between phonologically closely related single phonemes in predominantly phonological contexts." Phonological rules are then taken care of in TM. Morphological rules, on the other hand, i.e. processes that form e.g. plural word-forms from stems, are not possible in TM, which handles all formatives in the lexicon part.

Morphophonemic changes can thus be described in TM in a manner similar to IP (when we ignore the lack of morphemic level in TM) :

(10) IP:

Morphemes:	Allomorphs:	Morphophonemic rule:	'Word'
{bok} + {indef pl}	bok-er	o -> ø/ _ C er	bøker

(11) TM:

Lexical (stem) entry:	Lexical (affix) entry:	Two-level rule:	'Word:'
bok	er	o : ø/_ C er	bøker

The two models are similar so far, but only as long as the rule-context is purely phonological (graphemical). Morphological context is impossible in TM, but possible in IP :

(12) IP:

Morphemes:	Allomorphs:	Morphophonemic rule:	'Word'
{bok} + {indef pl}	bok-er	o -> ø / _ C {+pl ind}	bøker

The only way TM can use morphological information, is to make the morphological information 'phonological' by adding extra characters in the rule context. The extra character will then symbolize the morphological class or feature. E.g. can we put a dollar sign in front of the affix (which must of course also be present in the lexicon where the affix has its entry) or the morphological ending may itself get a different lexical shape, to satisfy the need for a context that can be morphologically unique:

(13) TM:
```
+ morphophonemic rule: o -> ø / _ C  $er
  or
+ morphophonemic rule: o -> ø / _ C  Er
```

Now one might want to reply that it is not important. But in natural language it is often necessary to distinguish between phonological and morphological conditioning. A number of Norwegian dialects have productive palatalization of /k/ and /g/ in front of noun suffixes, but not otherwise:

(14)
/stok/	(= stick)
/stoc-en/	(= the stick (nom.))
/drek-e/	(= (to) drink)
/stoc-a/	(= the stick (dative))
/tak-a/	(= thanked)

It would be a mistake to phonologize this type of morphophonological process. It is the morphological category of the suffix that conditions the alternation of the stem, and not the phonological shape.

3.3 Preliminary Summary

We have looked at two linguistic models for morphological analysis which both have the distinction morpheme-allomorph, i.e. which take the segmental side of natural languages very seriously. One of them, IP, is a little more flexible in that it accepts segmental changes triggered by some phonological or morphological feature.

When we compared TM to these models, we saw that it seems to be inspired by them. It too emphasizes the segmental side, through the linked lexicons. Also it seems to be inspired by the rule module of IP, although TM only allows "phonological" conditioning for the triggering of rules.

The serious defect of TM, however, is that it lacks a conceptual, morphological level. It operates only at the concrete, phonological (graphemic) one, which is of course the reason for the just mentioned phonological triggering.

We have seen that IA and IP are not fashionable today. The reason is that they are too limited to account for all facts about natural language. But the knowledge that morphological processes can be more than just elements arranged in a certain order is not new. E.g., the American linguist Edward Sapir in his book "Language" in 1921 distinguished between six different processes,

(i.e. ways of expressing morphological characteristics) where he included things like "internal modification of the radical or grammatical element", reduplication, accentual and quantitative processes.

It is this knowledge that led to a revision of morphology by Matthews in 1972.

3.4 Word and Paradigm

The WP is a model that attempts to take morphology seriously in that a grammatical feature can be realized in many different ways, like Sapir suggested. In this model the underlying representation is even more abstract than in the two preceding morpheme models. Any wordform is represented through its lexeme (an invariant representation of the word) with the grammatical (morphosyntactic) information represented as an unorded set:

(15) GUTT_{N} masculine, indefinite, plural

To reach the correct wordform, the stem, which is the starting point, can go through various processes:

(16) GUTT_{N} masculine, indefinite, plural:

Stem:	gutt
+ operation:	suffix -er
= Word	gutter

The number of processes is potentially infinite, the reason for this is that it is the word which is the basic unit in this model: If a wordform differs in more than one way from any other wordform in the same paradigm, then it goes through more than one process to reach its final shape. And all the processes will be exponents of the same grammatical (morphosyntactic) feature. We shall look more closely at three linguistic phenomena that are problematic for the two other linguistic models and for TM, but not for WP.

The first and most important difference between the WP model and the morpheme models is that while the morpheme models need a one-to-one relationship between morphological contents and its realization, WP accepts a many to one/one to many-relationship:

(17) BOK_{N} feminine, indefinite, plural

Stem:	bok
+ operation:	suffix -er
+ operation:	change stem vowel
= Word:	bøker

As we have seen previously, the morpheme models and TM necessarily must give priority to one of the realizations, and let the other(s) be conditioned by it. This poses strong constraints upon the linguist, who will have to give arbitrary priority to one realization, e.g. to let an affix trigger a vowel change.

A second problem is that while there still might be some universal claims about the priority of affixes to 'internal modifications', so that the first problem may look smaller, a worse case occurs when there are more than one *affix* that represents a grammatical feature. This is the case in German past participles:

(18) ge-sag-t

In the German case we might still argue to give priority to the suffix, though, since the other verbal features in the language are marked by suffixes, but consider the Kubachi dialect of Dargwa from the Northeast Caucasus, where each adjective agrees with the noun's gender and number both initially and finally, in addition to agreeing with number penultimately:

(19) Kubachi (dialect of Dargwa, Northeast Caucasus):
b-īk'a-zi-b qalč'e 'little bird'
d-īk'a-žu-d qulč'-ne 'little birds'
(The example is from Anderson 1988:32)

The morpheme models that we considered previously would have to give one affix priority over the other, or make use of the concept circumfix. The TM on the other hand, does not present a satisfactory solution.

It could represent the prefixed affixes in a separate lexicon, and have pointers from there to the further lexicons. This however would mean that the stem would occur as many times in the stem lexicon as there are prefixes in the language, since the suffixes have to agree with the prefixes. Other equally inelegant solutions are also possible.

But even if TM has problems in representing phenomena like the above, it can do it in an inelegant way. The third problem is more serious, however:

The third area is phenomena that are not 'segmental' in their nature. We recall Sapir who allowed internal modification as a means for representing morphological features. A typical example of this is the Germanic umlaut and ablaut, which in many cases is the sole distinguishing factor between two wordforms of the same lexeme:

(20) mann (indef sg)–menn (indef pl) (man–men)
se (infinitive)–så (preterite) (see–saw)
mouse (sg)–mice (pl)

Both the morpheme models that we looked at and TM have problems in describing such phenomena as vowel alternation, when it is not the biproduct of some other segmental, morphological process, but rather the main exponent of that morphological feature.

The main problem for TM is that all grammatical (morphosyntactic) information is only represented in the lexicon part of the system, and not in the rule part. If a two-level rule should take care of this information, it would 1) need a segment to trigger the rule, and 2) let the morphological information be accompanied by the trigger, and not by what is really the difference between the two forms—the vowel alternation:

(21) TM:

Lexical (stem) entry:	Lexical (affix) entry:	Two-level rule:	'Word':
`mann`	`-0`	`a:e <=> _ C* -0`	menn

4 Conclusion

The question which is the title of this paper, 'Is Two-level morphology a morphological model?', can from the previous discussion be answered quickly and clearly negatively. The reason for this is that a minimum to be demanded from a morphological model, is for it to accept morphological features and categories as primitives. In that way it could allow morphological conditioning on stem variation. But TM has to make use of artificial null-segments and other triggers, i.e. it has to make the originally morphological context 'phonological', segmental. Anything that is morphological—like the information in the lexicons—can be used only by the linguist, not by the model. In this way it does not have any possibility of making generalizations independently of phonological shape. It is even worse than the morpheme models, as it can not say that both a suffix and a vowel alternation could represent 'plural', e.g..

It therefore seems fair to say that TM is not a morphological model. There are, however, languages that can be well described by it, viz. the languages that are usually called agglutinative in traditional typology. These are languages like Finnish, whose morphology consists of easily separable affixes that each corresponds to one morphological unit. The phonological alternations in Finnish are not realizations of morphological features, only phonologically determined automatic alternations, which the two-level rules handles well. The important rule-part of the model makes the TM more phonological than morphological.

We may ask a last question: Is it important that a computational model has linguistic qualities? From the above discussion I think the answer should be positive.

References

Anderson, S.R. 1988. Inflection. In Hammond, M. & M. Noonan (eds.), *Theoretical Morphology*, 23–44. Academic Press Inc., San Diego.

Bybee, J. L. 1985. *Morphology. A Study of the Relation between Meaning and Form.* John Benjamins Publishing Company, Amsterdam.

Hockett, C.F. 1954. Two Models of Grammatical Description. *Word*, 10, 210–231. New York.

Johannessen, J. 1988. *Automatisk morfologisk analyse og syntese.* Lingvistisk institutt, Universitetet i Oslo, Oslo. Hovedoppgave.

Koskenniemi, Kimmo. 1983. Two-level Morphology: A General Computational Model for Word-form Recognition and Production. Publ. no. 11, Department of General Linguistics, University of Helsinki, Helsinki.

Källgren, G. 1983. Computerized Analysis and Synthesis of Finnish Nominals. In *Papers from the Seventh Scandinavian Conference of Linguistics*, 433–444. Publ. no. 10, Department of General Linguistics, University of Helsinki, Helsinki.

Karlsson, F. and K. Koskenniemi. 1985. A Process Model of Morphology and Lexicon. *Folia Linguistica*, 207–231, Haag.

Matthews, P.H. 1972. *Inflectional Morphology.* Cambridge University Press, Cambridge.

——1974. *Morphology. An Introduction to the Theory of Word-structure.* Cambridge University Press, Cambridge.

Rankin, I. 1986. SMORF—An Implementation of Hellberg's Morphology System. In *Papers from the Fifth Scandinavian Conference of Computational Linguistics*, 161–172. Publ. no. 15, Department of General Linguistics, University of Helsinki, Helsinki.

Robins, R.H. (ed.) 1970. *Diversions of Bloomsbury. Selected Writings of Linguistics.* North-Holland Publishing Company, Amsterdam.

Sapir, E. 1921. *Language.* Harcourt, Brace & World, New York.

Wurzel, W. U. 1982. *Phonologie - Morphonologie - Morphologie.* In *Linguistische Studien*, 93, Reihe A, Arbeitsberichte. Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft, Berlin.

Institutt for humanistisk informatikk,
Universitetet i Oslo,
P.b. 1102 Blindern, N-0317
Oslo 3.