The segmentation problem in morphology learning

Christopher D. Manning
University of Sydney
cmanning@mail.usyd.edu.au

Recently there has been a large literature on various approaches to learning morphology, and the success and cognitive plausibility of different approaches (Rumelhart and McClelland (1986), MacWhinney and Leinbach (1991) arguing for connectionist models, Pinker and Prince (1988), Lachter and Bever (1988), Marcus et al. (1992) arguing against connectionist models, Ling and Marinov (1993), Ling (1994) using ID3/C4.5 decision trees, and Mooney and Califf (1995, 1996) using inductive logic programming/decision lists, among others). However – except for a couple of forays into German – this literature has been exclusively concerned with the learning of the English past tense. This has not worried some. Ling is happy to describe it as "a landmark task". But while the English past tense has some interesting features in its combination of regular rules with semi-productive strong verb patterns, it is in many other respects a very trivial morphological system – reflecting the generally vestigal nature of inflectional morphology within modern English.

In this paper, I briefly discuss some experiments on learning morphological forms in languages with much richer morphological paradigms. Such languages are common throughout much of the globe (from Latin and Greek to Inuit and Cashinahua or Anmajere and Kayardild – to finish with some Australian examples). Attempting to learn morphology in languages with rich morphology raises quite different problems from those discussed in the work above, issues discussed – if rather naively and unsatisfactorily from a computational viewpoint – in earlier work such as Pinker (1984), MacWhinney (1978) and Peters (1983). Foremost among these is the *segmentation problem* of how one cuts the complex morphological forms into bits with meanings identified. Note that I assume here that the child has already figured out the meanings of words. This is a big assumption, but it is reasonable for a model to focus on one aspect of the learning problem – and at any rate the learning task is still much broader and more realistic than that attempted by the recent English past tense literature. It may not even be unrealistic; see Pinker (1984:29–30) for a general defense of assuming some form of "semantic bootstrapping" and MacWhinney (1978:70–71) who for arguments for the learning of word meanings before gaining a productive understanding of them ("it appears that the use of inflections in amalgams is stabilized semantically before these amalgams are analyzed morphologically"). Thus the learning task which I am attempting to address could be stated thus:

> Given a set of words and a representation of their meanings, determine an internalized representation that will allow heard and (regular) unheard forms to be successfully predicted and parsed.

The segmentation problem is difficult

There are many morphological issues that make the segmentation problem difficult. If a learner works on-line, then it has to be careful not to be sent down wrong tracks. For example, suppose a learner already knows that the English past tense is regularly /t/ after a voiceless sound. If it encounters the past of *burst* we would expect the following analysis to be generated:

(1)

Meaning	*Word*	*Stem*	*Tense*
[PRED: burst, TENSE: PAST]	burst	**burs**	**t**

The stem is wrongly found to be /burs/, and could only perhaps be fixed after observing the present and deciding that some form of reanalysis is necessary.

Many languages have *fusional morphology* where one morpheme expresses multiple semantic

Christopher D. Manning (1998) **The segmentation problem in morphology learning.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning,* ACL, pp 299-305.

components. This means that looking for a consistent phonetic exponent for one meaning component will be in vain. For example, consider tense in the following data from Pocomchi:

(2)	'to see'	Present	Past
	[SUBJ: I, OBJ: YOU]	tiwil	šatwil
	[SUBJ: I, OBJ: THEM]	kiwil	šiwil

The account of MacWhinney (1978) does not address fusional morhology, Pinker (1984) attempts to but various flaws in his proposed segmentation procedures mean that fusional morphology is frequently mishandled, and due to the simplicity of the English past tense task, none of the more recent work addresses this problem.

Further problems are created by inflectional classes (declensions or conjugations). For example, if one starts with a bunch of words in the Latin ablative singular:

(3)	mensā	table.ABL.SG
	servō	slave.ABL.SG
	urbe	city.ABL.SG
	manu	hand.ABL.SG
	rē	thing.ABL.SG

Then there is no (fusional) morpheme that expresses ablative singular. It has different allomorphs for different inflectional classes.

However, if the learning procedure just looks at stem-specific paradigms in isolation, and then compares the results to see if they happen to be similar (as Pinker (1984) suggested), there is nothing to make the learner hunt out similarity, to look deeper for alternative analyses that would expose common underlying structure (much as a linguist does). It is only this latter sort of approach that will allow us to postulate general phonological rules. Although a symbolic morphology learner presumably must start with stem-specific paradigms, we need to have a counterbalancing principle of *paradigm economy* (Carstairs 1988), which collapses together stem-specific paradigms where possible, even when this wasn't the obvious analysis at first. For example, consider the consonant-stem declension of Greek or Latin (the examples here are from Koiné Greek). If we see the forms:

(4)	himas	thong.NOM.SG
	himanta	thong.ACC.SG
	himantos	thong.GEN.SG

then (if it were not for any prior knowledge of Greek or Latin), the obvious analysis would be:

(5)	hima-	[pred: thong]
	-s	[case: nom, num: sg]
	-nta	[case: acc, num: sg]
	-ntos	[case: gen, num: sg]

and we will find other words that appear to decline similarly. However, when we see a reasonable collection of words of another kind:

(6)	skolops	stake.NOM.SG
	skolopa	stake.ACC.SG
	skolopos	stake.GEN.SG

we can decide it would be better to reanalyze the forms above thus:

(7)	hima-	[pred: thong] / ___ s
	himant-	[pred: thong] / elsewhere
	skolop-	[pred: stake]
	-s	[case: nom, num: sg]
	-a	[case: acc, num: sg]
	-os	[case: gen, num: sg]

The key to discovering the phonological rule that deletes alveolars before /s/ is a notion of paradigm economy that suggests the reanalysis shown in (7).

For identifying allomorphs of morphemes, Pinker (1984) depends heavily on a notion of "phonetic material in common". However, he merely suggests that the definition of this notion should be drawn from an appropriate theory of phonology. But in general a theory of phonology cannot just take two words and tell one what their "phonetic material in common" is. To consider an example from Latin nouns raised by Braine (1987), given the noun forms *ordo* and *ordinem*, the phonetic material in common is going to be *ord*. It requires a more sophisticated level of theory formation to determine that the desired root form for this word is actually *ordin*. Even in simpler cases of sandhi (word internal phonological changes), it will not be immediately apparent what the stem of a word (or other morphemes within it) is. Consider the Japanese verb forms in (8):

(8)	nomu	drink (present)
	nonda	drank (past)
	nomitai	want to drink
	nomimasu	drink (present honorific)

Is the stem 'drink' *no*, *nom*, or even *nomi*? Such a question cannot in general be answered simply using a notion of common phonetic material, but must be answered in terms of a broader understanding of the paradigmatic system of the language as a whole.

MacWhinney (1978) does provide an explicit, if simplistic, theory of phonetic similarity. In it, parts of words match only if they are string identical. But this notion is insufficient to account for not only sandhi effects but also many of the phenomena that inspired autosegmental phonology, that is, melodies being stretched or squashed to fit onto a skeleton. In particular, consider vowel lengthening of the sort shown in (9), from Hungarian:[1]

(9)

	SG	PL
water	viiz	vizek
fire	tüüz	tüzek
bird	madaar	madarak

It is clearly necessary for a learner to be able to identify the stems of these words as *viz*, *tüz* and *madar*, despite the fact that they are not segmentally identical in their two appearances. This will never happen if segments are simply matched one-for-one.

We see that getting a start on the segmentation problem seems to have two main components: working out what the allomorphs and/or underlying forms in the data are and working out the environments in which different allomorphs occur. For the first segmentation problem, we saw that neither allomorphs nor especially underlying forms can be correctly determined by just looking for "phonetic material in common". Indeed, we determined the stronger result that appropriate stems often cannot be determined by looking at a stem-specific paradigm at all, but can only be determined by comparisons across the morphological system, invoking some notion of paradigm economy. For the second problem, we can use existing classification techniques, which have been explored in the English past tense work. For example, one can use ID3, as I do here, as an algorithm that can find conditioning features while still being reasonably tolerant of noise (that is, irregular forms) in the data.

An implemented symbolic morphology learner

My model works from being given pairs of a surface (allophonic) form and a representation of its meaning (this essentially consists of just encoding a word's position within paradigmatic dimensions of contrast, by giving it a meaning such as [PRED: apple, NUM: SG, CASE: ACC]). It works essentially as an affix-stripping model of morphological processing with a back-end environment categorization system based on the ID3 algorithm. My model and indeed all the models mentioned above, connectionist and symbolic alike, assume that morphemes and words can be satisfactorily represented as a linear sequence of segments. This flies in the face of much recent work in phonology (e.g., Goldsmith 1990), but works for 90% of languages, and is a useful simplifying assumption at this stage. However, I will introduce mechanisms that allow conditioning by nearest consonants or vowels, and the stretching of melodies, which actually allow us to capture some (though not all) of the features of an autosegmental analysis.

The model I will present here, like all English past tense models, is one of *conditioned allomorphy* that attempts to provide a solution to the two problems mentioned at the end of the last section: determining what the allomorphs of morphemes are and the environments where they occur. This is still somewhat less than a complete theory of phonology. So long as productive phonological changes are confined to inflectional endings, such a theory is in fact sufficient. However, if productive phonological rules change stems, then something more is needed: one must postulate phonological rules that can then be applied to generate the allomorphs of newly heard stems. This last task is not attempted here. However, it seems reasonable to suppose that this is a higher-order inductive step that would build on the results of a theory of learning conditioned allomorphs.

Chopping words into morphemes Words and their paradigmatic meanings are collected until a reasonable percentage of the forms for a particular stem-specific paradigm have been seen.

At this point a stem-specific paradigm is analyzed. The model (heuristically) determines likely candidates for the first or last morph in all words that contain the appropriate semantic feature (features are here things like TENSE or SUBJ.NUM) by looking at words that share a certain feature, and seeing if they are all phonetically similar at one end or the other. For each such candidate in turn, the model determines candidate guesses for each morpheme that expresses this feature.

The model uses both similarity matching between all words sharing a morpheme, and difference matching from the other end with all words

[1] In Hungarian orthography long vowels are indicated by acute accents, but here I write them as double vowels, roughly approximating the phonetic input to the child.

that have the same meaning except in the value of the morpheme in question to determine candidate morpheme values, as indicated in (10):

(10) a. Given *carries* and *carried*, one can attempt to learn [PRED: CARRY] by similarity matching.

b. Again given *carries* and *carried*, one can attempt to learn either PAST or PRES.3SG by difference matching (since the rest of the morphemes in these words are identical).

In the presence of word internal sandhi, using both same and difference matching will generally serve to delimit the boundary region wherein sandhi effects are occurring, and the model considers the possibility of a morpheme break anywhere within this sandhi region. For example, given the following data:

(11)

	'foot'	'house'
'my'	kepina	yotna
'your'	kepika	yotda

the program determines /a/ as a value for 'your' by same matching, but /ka/ and /da/ by difference matching by looking at the two forms for 'foot' and 'house' respectively. These two boundary points mark out the sandhi region within which the value of 'your' must be found (i.e., it is either /a/ of /Ca/ for some consonant).

To determine whether two strings of segments might reasonably be two allomorphs of a morpheme, the model uses a similarity condition. This is measured by counting a mismatch in phonological features. The model uses fairly standard phonological features (based on those in Halle and Clements 1983). This requirement of surface similarity between morphs is similar to, but weaker than having a Unique Underlier Condition. Across different word-specific paradigms, the form of a morpheme can vary at will – the similarity condition only applies when analyzing a word-specific paradigm, or a group of such paradigms when attempting inflectional class formation. Within a paradigm, if a solution satisfying the similarity condition cannot be found, then fusional morphs must be postulated.

As well as allowing a certain amount of mismatch of features between 'matching' segments, the similarity matcher was also built to handle the stretching (or squeezing) of melodies. When a segment occurs multiple times in one form, the matching routine will nondeterministically attempt to match any number of copies of that segment in one word with the segment in other words. In this way the Hungarian stem allomorphs discussed in (9) can make it past the similarity condition.

When a proposed form has been found for each value of a feature (i.e., each case of a case feature or whatever), these affixes are then stripped from the correct end of all words that contain them, and the above analysis procedure can then be applied recursively to the remaining partial words. With luck, this procedure will correctly analyze words, but in cases of sandhi where the learner has had to make guesses, there may be mistakes. The model includes a number of obvious heuristics to tell it that a mistake has been made:

- If values have been assigned to all features, but there are still some segments left unassigned as a residue, then an error has occurred.
- If a stem is null an error has occurred. (Since most analysis is done on word-specific paradigms, which give no evidence of contrasting stems, this can be a useful heuristic.)
- An initial pass examines words that differ in one feature and if those words are different, the model notes that the values of the feature concerned must be different. If a solution then tries to assign an identical value to these different morphs then an error has occurred.

In cases of error, certain potential segmentations are eliminated (where multiple possible segmentations have been generated, as in the presence of sandhi effects). The limiting case is when no possible way of chopping the word into morphs succeeds. As mentioned above, this is indicative of fusion, which was defined as a last resort when there is no available analysis of multiple features into separate morphemes (allomorphs). In such cases all possible analyses should fail in this first phase, and the model will then recursively attempt higher level analyses that postulate first partially and then finally totally fusional analyses (so that, for example, instead of trying to find a morpheme representing each case of a CASE feature, the model will be trying to find a morpheme representing each value of the crossproduct of two or more features, for example a value for each case and number combination).

On completion of an analysis of this sort, the history of the morpheme stripping order can be reconstructed to give the morpheme order in words.

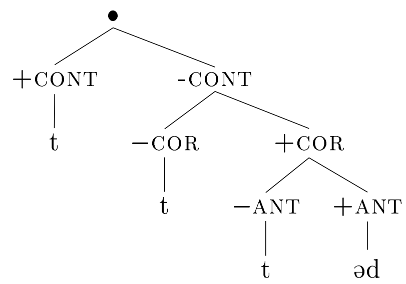
Additionally the program notes whether each feature appears to be compulsorily expressed or optional in the words that it has been trained on. No more subtle ordering information than this is currently learned.

Forming Inflectional Classes The above gives a plausible first attempt at a model that chops words into morphemes. But earlier, I argued that the correct chop point cannot always be discovered while looking at just a single word-specific paradigm. My program attempts to solve such problems by a process of inflectional class formation. After a second stem-specific paradigm has been analyzed, the model examines the two sets of endings that have been generated, and determines whether they are similar.[2] If the endings appear similar, the analysis procedure described above is then applied to words belonging to both stems simultaneously. If this analysis succeeds (proposing at most the same amount of fusion as when examining the stem-specific paradigms), then this reanalysis for the two words is recorded. Such a reanalysis can move the morpheme boundaries in cases such as the Greek consonant stem declension discussed above (4).

Learning phonological conditioning Once words are (hopefully correctly) segmented into morphemes, there may still be several allomorphs of a morpheme, and there remains the problem of determining which allomorph occurs when. The model assumes two possible forms of allomorph conditioning, phonological conditioning and lexical conditioning (where the stem lexeme determines which allomorph occurs), and uses a decision tree based learning system (with pruning) that can handle noisy input and disjunctive class descriptions is employed. To operate, the ID3 algorithm needs a list of possible features that can condition changes. The list used here is the following: an allomorph can be conditioned by any phonological feature (cons, son, ant, etc.) of any of the preceding or following segment or the preceding or following [−cons] or [−syl] segment. This captures autosegmental-phonology-like affects, since we are allowing the nearest consonant and vowel to also be 'adjacent' for the purposes of conditioning. If the decision tree fails to find phonological conditioning features, then lexical conditioning is assumed.

The output decision trees are then converted to something more similar to conventional phonological rules. However, in this model, all environments are surface conditions, so we cannot compact rule systems by using rule ordering (to selectively bleed/feed various rules). Instead a system of rule priorities was implemented, so that groups of rules form *default hierarchies* (Holland et al. 1986). This notion is the same as having elsewhere conditions on rules, as in the notion of disjunctive rule ordering. So, rather than having either the decision tree in (12a) or the equivalent rule set in (12b), the use of a default hierarchy lets us use the representation shown in (12c). Rules preceded by a number have a higher priority (equal to that number) and will apply in preference to other (usually but not necessarily more general) rules. Rules not preceded by a number can be regarded as having priority 1. Thus a word ending in a [−cont, +cor, +ant] sound will take the allomorph [əd], while all other sounds will receive the allomorph [t].

(12) a.



b. $[\text{tense: past}] \rightarrow \text{t} \;/\; \underset{\left[\begin{array}{ll}+ & \text{CONT}\end{array}\right]}{\text{X}}\ \_\_$

$[\text{tense: past}] \rightarrow \text{t} \;/\; \underset{\left[\begin{array}{ll}- & \text{CONT}\\ - & \text{COR}\end{array}\right]}{\text{X}}\ \_\_$

$[\text{tense: past}] \rightarrow \text{t} \;/\; \underset{\left[\begin{array}{ll}- & \text{CONT}\\ + & \text{COR}\\ - & \text{ANT}\end{array}\right]}{\text{X}}\ \_\_$

$[\text{tense: past}] \rightarrow \text{əd} \;/\; \underset{\left[\begin{array}{ll}- & \text{CONT}\\ + & \text{COR}\\ + & \text{ANT}\end{array}\right]}{\text{X}}\ \_\_$

c. [tense: past] → t

[2] The model uses an heuristic measure of similarity that focuses on the 'nucleus' of morphemes. That is, due to mistakes in segmentation, the margins of morphemes may well be different, but if they really belong to the same inflectional class, they should have a common core.

$$2\text{: [tense: past]} \rightarrow \text{əd} \;/\; \underset{\begin{bmatrix} - & \textsc{cont} \\ + & \textsc{cor} \\ + & \textsc{ant} \end{bmatrix}}{\mathrm{X}} \;\_\_$$

Finally the model includes a simple parser/generator which can use the rules learned by the preceding processes to parse and generate morphological forms.

Experimental Results

I've done small studies with my model on portions of the morphological systems of a number of languages. Provided the language phenomena stay within the bounds of what the model can cope with (i.e., avoiding semitic and similar templatic languages), it is a fairly robust learner. I assure the reader that my model can also learn the English past tense – essentially duplicating Ling's results, but actually doing the segmentation work rather than just a classification task. Here I will present a small study of the tense endings of Anmajere verbs (Anmajere is an Australian language; data is from Avery Andrews (p.c., 1989)). In addition, small studies have shown that the model can learn the following examples which I have mentioned previously:

- Latin nouns (this involves learning fusional case and number morphemes)
- Greek consonant stems (this involves learning paradigm reanalyses for data such as (4)
- Japanese verb morphology (word internal sandhi obscures morpheme size)
- Hungarian nouns (9) – this requires the stretching of melodies

Figure **??** shows verb forms from Anmajere. The digraphs *rr*, *rl*, *rn* and *rd* represent a single sound (a trill for *rr*, the rest are apical reflexives), using the usual orthography for Australian languages (Dixon 1980). While all these verbs are regular, they demonstrate more subtle phonological conditioning than in the English past tense. A final labial stop of verb stems is voiced or voiceless depending on the voicing of the first consonant (not the next sound) of the inflection (see the verbs 'depart' and 'leave alone'). In the inflections, the recent past has two allomorphs, having an apico-alveolar /n/ when the stem ends in a [−COR] consonant (for example, with *arlk-* 'yell'), and the retroflex /rn/ when it ends in a [+cor] consonant (for example, with *id-* 'cook').

Verb	Present	Recent past	Past
yell	arlkeme	arlken	arlkeke
depart	albeme	alben	alpeke
leave alone	imbeme	imben	impeke
hear	aweme	awen	aweke
cut	akeme	aken	akeke
speak	agkeme	agken	agkeke
cook	ideme	idern	
sit	aneme	anern	
take	ineme	inern	
shape	ardeme	ardern	
come out	arrademe	arradern	
run	arrjaneme	arrjanern	
chuck	iweme		
pop	ardeme		

Figure 1. Anmajere verb forms

My model can successfully learn these distributions, producing rules such as these:

(13) a. [pred: yell] → "arlk"

b. [pred: depart] → "alb" / $\_\_ \; \underset{[+\ \textsc{son}]}{\mathrm{C}}$

c. [pred: depart] → "alp" / $\_\_ \; \underset{[-\ \textsc{son}]}{\mathrm{C}}$

d. [tense: rec] → "ern" / $\underset{[+\ \textsc{cor}]}{\mathrm{X}} \; \_\_$

e. [tense: rec] → "en" / $\underset{[-\ \textsc{cor}]}{\mathrm{X}} \; \_\_$

f. [tense: pres] → "eme"

g. [tense: past] → "eke"

The model chose [+son] rather than [+voiced] as the distinguishing feature for the first alternation, which gives non-distinct results for the data that was given. This knowledge is sufficient for the model to be able to fill in the remaining entries in the above table (as a transfer test). However, as noted before, the model would need to go one stage further and learn universally applicable phonological rules to be able to extend its knowledge of stem allomorphy from known verbs to new or nonce verbs.

Conclusion

This work introduces a more substantial and realistic problem domain for morphology learning programs, and demonstrates a symbolic morphology learner that can learn an interesting range of the complex morphological systems found in the world's languages. On the other hand, it is not the final word, and more work still has to be done on generalizing its representations and algorithms so that it is capable of learning the morphology of all human languages.

REFERENCES

Archangeli, D. 1988. Aspects of underspecification theory. *Phonology* 5:183–208.

Braine, M. D. S. 1987. What is learned in acquiring word classes—a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*, 65–87. Hillsdale, NJ: Lawrence Erlbaum.

Carstairs, A. 1988. Nonconcatenative inflection and paradigm economy. In M. Hammond and M. Noonan (Eds.), *Theoretical Morphology: Approaches in Modern Linguistics*, 71–77. San Diego, CA: Academic Press.

Dixon, R. M. W. 1980. *The Languages of Australia.* Cambridge: Cambridge University Press.

Goldsmith, J. A. 1990. *Autosegmental and Metrical Phonology.* Oxford: Basil Blackwell.

Halle, M., and G. N. Clements. 1983. *Problem book in phonology.* Cambridge, MA: MIT Press.

Holland, J. H., K. J. Holyoak, R. E. Nisbett, and P. R. Thagard. 1986. *Induction: proceses of inference, learning and discovery.* Cambridge, MA: MIT Press.

Lachter, J., and T. G. Bever. 1988. The relation between linguistic structure and associative theories of language learning—a constructive critique of some connectionist learning models. *Cognition* 28:195–247.

Ling, C. X. 1994. Learning the past tense of english verbs: the symbolic pattern associator vs. connectionist models. *Journal of Artificial Intelligence Research* 1:209–229.

Ling, C. X., and M. Marinov. 1993. Answering the connectionist challenge: a symbolic model of learning the past tense of english verbs. *Cognition* 49:235–290.

MacWhinney, B. 1978. The acquisition of morphophonology. Monographs of the Society for Research in Child Development, 43 (1–2, Serial No. 174).

MacWhinney, B., and J. Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40:121–157.

Marcus, G. F., S. Pinker, M. Ullman, M. Hollander, T. J. Rosen, and F. Xu. 1992. *Overregularization in Language Acquisition.* Chicago, IL: University of Chicago Press.

Mooney, R. J., and M. E. Califf. 1995. Induction of first-order decision lists: Results on learning the past tense of english verbs. *Journal of Artificial Intelligence Research* 3:1–24.

Mooney, R. J., and M. E. Califf. 1996. Learning the past tense of english verbs using inductive logic programming. In S. Wermter, E. Riloff, and G. Scheler (Eds.), *Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing.* Springer Verlag.

Peters, A. M. 1983. *The units of language acquisition.* Cambridge: Cambridge University Press.

Pinker, S. 1984. *Language Learnability and Language Development.* Cambridge, MA: Harvard University Press.

Pinker, S., and A. Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193.

Rumelhart, D. E., and J. L. McClelland. 1986. On learning the past tenses of English verbs. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, 216–271. Cambridge, MA: MIT Press.