Suffixing Preferences as a Consequence of Probabilistic Reasoning

Yoshihiko Asao
University at Buffalo

It has been known that suffixes are typologically more common than prefixes (Greenberg, 1963; Hawkins & Gilligan, 1988). This paper provides a new psycholinguistic account for this phenomenon based on a probabilistic model of morpheme segmentation.

Previous psycholinguistic accounts include Greenberg (1957) and Cutler et al. (1985). These accounts require specific assumptions about differences between syntax and morphology, and/or between stems and affixes. For example, in order to Cutler et al.'s account to work, we have to assume that (i) the parser needs the lexical information carried by a stem earlier than the grammatical information carried by an affix, and that (ii) there is no such preference beyond a word boundary, in order to account for the fact that a similar asymmetry is not found in syntax.

This paper proposes an alternative approach: the timely recognition of prefixes is difficult. We discuss that this prediction follows from distributional facts such as morpheme length and frequency, as well as the assumption that probabilistic reasoning is at work in language comprehension (Chater & Manning, 2006).

The following simulation studies were conducted in order to demonstrate our argument. First, a bigram model of English morphemes was constructed using the CELEX2 Lexical Database (Baayen, Piepenbrock, & Gulikers, 1996) and Google Web NGram. The model does *not* distinguish word-internal morpheme boundaries from word boundaries, so that we can avoid making a priori assumptions about differences between morphology and syntax. With this model, a simulation program was designed to incrementally reconstruct the most probable messages given partial phonetic inputs. This is equivalent to finding the most probable path in the lattice of possible morpheme sequences, as illustrated in Figure 1.

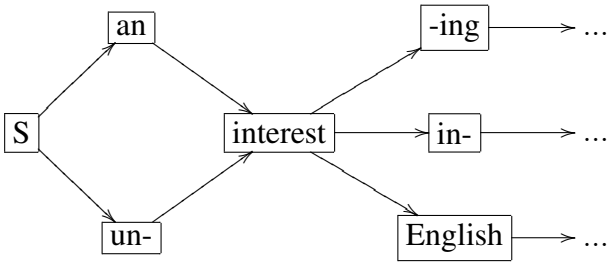

Figure 1: An example of lattice structure for the input /ənɪntərestɪŋ/

The probability ("confidence") that the model assigned to the existence of that morpheme boundary was computed and used as an indicator of the easiness of the morpheme boundary detection. For example, when an input starts with /hi/, 'he' is the most likely morpheme that matches this input with the probability of 93%. Assuming that the original message in fact starts with the word 'he', the confidence for the word boundary following 'he' will be 93%.

The simulation was run over 1,000 randomly generated sentences. In order to make sure that our results are not due to some peculiarity of English, the same procedure was applied to the data from the following two languages:

(1) a. **Reverse English**
A hypothetical language in which everything is the same as English except phonemes are given backwards per each sentence. Because English is a strongly suffixing language in the sense of Dryer (2011), Reverse English is a strongly prefixing language.

b. **Walman**
A Papuan language that is classified as 'weakly prefixing' (Dryer, 2011). The data was made available courtesy of Matthew Dryer.

Table 1 summarizes the results from the three languages. Morpheme boundaries were classified into three types: word boundary, boundary after prefix, and boundary before suffix.

Table 1: Confidence (%) by different types of morpheme boundaries

language	word boundary	boundary after prefix	boundary before suffix
English	90.7	47.9	82.9
Reverse English	87.4	51.6	87.5
Walman	82.4	74.1	75.5

As can be seen from Table 1, in English the mean confidence of boundaries after prefix is much lower than that of word boundaries or boundaries before suffix, suggesting the difficulty associated with prefixes. The results from Reverse English are surprisingly similar to those from English. The confidence for the boundaries after prefixes remained low at 51.6%, despite the fact that there were more prefixes than suffixes. In Walman, on the other hand, the confidence for boundaries after prefix is similar to that for boundaries before suffix.

We argue that the difficulty in detecting prefixes is caused by a combination of three factors: length, predictability, and phonological word-boundary cues.

(2) a. **Length**

The boundary after a short morpheme is hard to identify, because a short morpheme is more likely to match a part of other morphemes by chance (cf. Laudanna and Burani 1995).

Because a stem is on average longer than an affix, this factor predicts that a prefix-stem boundary is more difficult to detect than a stem-suffix boundary. For example, consider the English word form *pipes* /pa^{j}ps/, which has the plural suffix *-s*. After you hear the first three segments /pa^{j}p/, you are very sure that the morpheme *pipe* is used, and you are ready to hear the next morpheme. On the other hand, in Reverse

English the equivalent word is /spaʲp/. When you hear *s*-, you cannot tell whether this is the plural prefix or a part of some other morpheme, and the morpheme boundary can be identified only retrospectively.

b. **Predictability**

A next morpheme is more predictable from context within a word than across a word boundary.

For example, in our English training data, the mean transition probability between morphemes within a word is as high as 47.8%, while the mean transition probability across a word boundary is 3.9%. This means that when a morpheme is not the first morpheme of a word, it is often highly predictable from its preceding context. This is arguably because morphology is characterized by fixed order, allomorphy, limited productivity, and lexical idiosyncrasy.

This suggests that a prefix is less predictable from the preceding context than a suffix, because a prefix is often the first morpheme of a word, while a suffix is by definition never the first morpheme of a word. This means the identification of a prefix is harder than that of a suffix, even when they have the same length.

c. **Phonological cues for word boundaries**

A word-internal morpheme boundary is harder to detect than a word boundary, because a word-internal morpheme boundary typically has fewer phonotactic and suprasegmental cues. For example, in English a phoneme sequence like /nl/ is rare in a single morpheme, and therefore it can be used as a cue of a morpheme boundary. However, an assimilation process, which occurs more often within a word than across a word boundary, has an effect of eliminating such low-frequency phoneme sequences.

Although the effects of phonotactics or suprasegmental cues were not directly implemented in our model, they did influence our results. For example, the prefix *in*- undergoes assimilation and ends up with /ɪ/ when it is followed by /n/, /m/, /l/ or /r/, as in *immature* and *illegal*, which makes the detection of the morpheme boundary more difficult. In our English simulation results, the mean confidence for the boundary after the prefix /ɪn/ was 61.4%, while the mean confidence for the morpheme boundary after /ɪ/ was very low at 0.7%.

This factor helps detect the morpheme at the end of a word. Because prefixes are by definition never at the end of a word prefixes, it provides another reason why a prefix is more difficult than other types of morphemes. This factor also predicts that a preposed grammatical word is not as difficult as a prefix.

The facilitation effects of these three factors can be summarized as in Table 2. We can see that every factor works against prefixes in favor of some other types of morphemes.

It must be noted that we do not claim that every prefix is difficult to process. For example, the detection of a prefix would not be a problem if it is very frequent. In fact, Walman, a

Table 2: facilitating factors by morpheme types

facilitates the recognition of ..	prefix	stem	suffix	grammatical word
length	–	✓	–	–
predictability	–	–	✓	–
phonological word-boundary cues	–	(✓)*	(✓)*	✓

*Facilitated only when they are at the end of a word

real language with more prefixes than suffixes, did not show the difficulty with prefixes. In Walman, the relatively high confidence for prefixes is due to the high frequency of person-number prefixes on verbs. *n-* (3SG.MASC), *w-* (3SG.FEM) and *y-* (3PL) account for the majority of prefixes in Walman, and the mean confidence values for the boundaries after them are high at 72.1%, 66.1% and 88.6% respectively. Getting rid of prefixes can be conceived as just one of the strategies that a language can take to avoid the processing difficulty.

The advantage of our approach is that there is no need to stipulate different processing mechanisms for syntax and morphology, nor different psycholinguistic statuses for stems and affixes. Rather, suffixing preferences simply follow from observable distributional facts of morphemes, plus the assumption that people can make probabilistic inferences in language comprehension.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *CELEX2*. Philadelphia: Linguistic Data Consortium.

Chater, N., & Manning, C. D. (2006, July). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*(7), 335–44.

Cutler, A., Hawkins, J. A., & Gilligan, G. (1985). The suffixing preference: a processing explanation. *Linguistics*, *23*, 723–758.

Dryer, M. (2011). Prefixing vs. Suffixing in Inflectional Morphology. In *The world atlas of language structures online* (chap. 26). Munich: Max Planck Digital Library.

Greenberg, J. (1957). *Essays in Linguistics*. Wenner-Gren Foundation.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In *Universal of language: Volume 2* (pp. 73–113).

Hawkins, J. A., & Gilligan, G. (1988). Prefixing and suffixing universals in relation to basic word order. *Lingua*, *74*, 219–259.

Laudanna, A., & Burani, C. (1995). Distributional properties of derivational affixes: implications for processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 345–364). Laurence Erlbaum Associates.