

Whole Word Morphologizer

Expanding the Word-Based Lexicon: A Non-Stochastic Computational Approach

Sylvain Neuvel, University of Chicago

If we take the enrichment of lexica to be not only the *raison d'être* of morphology but also the central issue of morphological theory, it seems reasonable to evaluate a theory of morphology, not on the means by which it represents recurring partials or lexical relations, but on its ability to generate new words based on a given lexicon. Within the framework of Whole Word Morphology (cf. Ford and Singh 1991, Ford *et al* 1997), I designed a small computer program that identifies morphological relations found in a lexicon and creates new words based on these relations. The purpose of this paper is to demonstrate the generative power of this *Whole Word Morphologizer* [henceforth WWM], to spell out the theory behind it and to discuss some of the theoretical issues that arose during its development.

Under the assumption that the morphology of a language resides exclusively in differences that are exploited in more than one pair of words within its lexicon (c.f. Neuvel and Singh 2000), WWM compares every word of a small lexicon (1000 to 5000 labeled phonemic or orthographic forms) and calculates the segmental differences found between them. Some of these differences occur more than once and are translated into bi-directional word-based morphological strategies that can be represented as:

$$(1) \quad /X/a \leftrightarrow /X'/b$$

where:

- a. $/X/a$ and $/X'/b$ are words and X and X' are abbreviations of the forms of classes of words belonging to categories a and b (with which specific words belonging to the right category can be unified or onto which they can be mapped).
- b. $'$ represents (all the) form-related differences between $/X/$ and $/X'/$
- c. a and b are categories that may be represented as feature-bundles
- d. the \leftrightarrow represents a bi-directional implication (if X , then X' and if X' , then X)
- e. X' is a semantic function of X (c.f. Ford *et al* 1997)

Each word in the lexicon is then mapped onto as many strategies as possible and contrasting new words are added to the lexicon. While some combinatorial restrictions are specified in each strategy, Whole Word Morphology, like most analogical models of word formation, is burdened by overgeneration. Just as, in componential morphology, two different morphemes can combine with different subclasses of stems but serve the same function, the relation between words of two categories is often expressed by two or more competing strategies. For example, when using the text of *Le petit prince* as its base lexicon, WWM produces two different strategies relating 2nd person verb forms to their infinitives. Given the verb *conjugues* 'conjugate, pres. 2nd sing.', one strategy produces the correct infinitive *conjuguer* while the other creates the word **conjuguere* (based on the relation between words like *fais/faire* 'do' and *vends/vendre* 'sell'). Using output from WWM as evidence, I argue that overgeneralization can be significantly reduced if the application of competing strategies is governed by a simple selection principle that allows only the most restrictive strategy to apply.

Brief References

Ford, A. & R. Singh. 1991. Propedeutique Morphologique. *Folia Linguistica*. 25.

Ford, A., R. Singh & G. Martohardjono. 1997. *Pace Panini*. New York: Peter Lang.

Neuvel, S. & R. Singh, 2000 *Vive la différence! What Morphology is About*. Presented at the 9th International Morphology Meeting, Vienna, Austria, February 25th 2000