Networks of morphological relations
Extended abstract

Sean A. Fulop
Dept. of Linguistics, Fresno State University

There has been a considerable amount of research on computerized methods of inducing morphology. The majority of these efforts have lately been directed toward "unsupervised" techniques, which in this context means that some morphological knowledge is induced from raw text (Hammarström and Borin 2011). "Morphological knowledge" is usually taken to mean something that a linguistic analysis would yield, including a lexicon of roots and affixes, and perhaps an organization into paradigms.

In this paper a different approach to morphology induction is presented, which is founded on an unusual model of morphological knowledge. The model is Whole Word Morphology (WWM), which is due originally to (Ford and Singh 1991). The usual way of modeling morphology involves linguistic elements called *morphemes,* which are supposed to be the "smallest meaningful units." This model is fraught with difficulties, and is probably not an accurate cognitive model of morphology. Whole Word Morphology trades morphemes for relations among words, formalizing the latter as its morphlogical primitives.

In Whole-Word Morphology, any morphological relation can be represented by a rule of the following form:

$$|X|_{\alpha} \leftrightarrow |X'|_{\beta} \tag{1}$$

in which the following conventions are employed:

1. X and X' are schematic forms of classes of words belonging to categories α and β (with which specific words can be unified in form);
2. vertical brackets signify that orthographic forms are to be used, though phonological forms can serve as well;
3. $'$ represents all the form-related differences between X and X';
4. α and β are lexical categories;
5. $|X'|$ and $|X|$ are semantically related in a specifiable fashion.

The arrow $\leftrightarrow$ thus represents together a formal, category, and semantic correspondence between words called a *lexical correspondence* (Neuvel 2003). These are the primitives of WWM; once we have lexical correspondences in our toolbox, morphemes are no longer useful, and it can be argued that they are cognitively non-existent.

A lexical correspondence (LC) representing orthographic English word pairs like *receive, reception* is shown below. An LC can equally well show word forms phonologically if desired, and can accommodate almost any viewpoint on the syntactic categories including categorial grammar.

$$|*\#\#\text{ceive}|_{\mathrm{V}} \leftrightarrow |*\#\#\text{ception}|_{\mathrm{Ns}} \tag{2}$$

The '#' signs in the above stand for letters that must be instantiated but are not specified; the '$*$' symbol stands for a letter that is not specified and that may or may not be instantiated. The LC (2) can therefore be interpreted as follows:

There is a verb that ends with the sequence "ceive" preceded by no less than two and no more than three characters, if and only if there is a singular noun that ends with the sequence "ception" preceded by the same two or three characters.

In previous computational work on WWM, lexical correspondences were called *word-formation strategies* (Neuvel and Fulop 2002), and from an acquisition perspective that is exactly what they are. An LC like (2) shows how to form a noun when one already knows a verb of a certain form, and vice versa. There is an advantage in not getting concerned about what the roots and morphemes are—does it matter? WWM can be taken as a cognitive and computationally useful model of morphology, in which the lexical correspondence is primitive. Moreover, the induction of morphological knowledge in the form of the LCs is deterministic, and need involve no statistics. The morphological knowledge that is available in a language corpus comprises exactly the lexical correspondences that can be discovered by a deterministic algorithm, such as that presented by (Neuvel and Fulop 2002). A speaker (or system) can be said to "know the morphology" of a language if he can correctly form word w_a from first knowing any word w_b that is related by a lexical correspondence.

The above simplistic model of morphology would be wonderful, if only it were generally adequate. It was shown in (Neuvel 2003) that while the basic WWM model works perfectly well for languages with "one-slot" morphology, it breaks down in "polysynthetic" languages with longer words and rich morphology. Neuvel also outlined a solu-

tion in the form of a more complex WWM model, which we examine in the present paper.

Consider the following example LC in West Greenlandic Eskimo from (Neuvel 2003), written phonemically:

$$/X_1C_1C_2Y/ \leftrightarrow /X_1C_1\text{taq}C_2Y/ \quad (3)$$

in which the elements X_1 and Y are phonemic strings and C_i are unspecified consonants, the left side is a verb and the right side is the same verb with a habitual aspect. In real verbs such as /saniuqqutpuŋa/ 'I go by,' this much information is not sufficient—notice we still don't know where to insert /taq/ to form the habitual.

The key idea is to relate one LC to another, which then allows the indeterminacy to be resolved by unification. Consider another LC from West Greenlandic:

$$/X_2\text{puŋa}/ \leftrightarrow /X_2\text{tuq}/ \quad (4)$$

A further requirement is that LC (3) be deterministic for certain instantiations; e.g. when $/X_1C_1C_2Y/ = /\text{uqaqpuq}/$ 'he/she says.' This instantiation unifies with LC (4) by setting $X_1C_1 = X_2 = /\text{uqaq}/$, and this resolves the indeterminacy in LC (3)—we now see where /taq/ should be inserted to create /saniuqqut-taq-puŋa/.

Examining these "metamorphological" relations brings us to the following picture. Lexical correspondences within a language can be arranged into a network with two sorts of connections, one symmetric and one asymmetric. The symmetric relation is that wherein two LCs share one template. LCs naturally cluster into cliques in the network by this relation, which were called "constellations" in (Neuvel 2003). More crucial to the WWM model is the asymmetric relation wherein one LC "depends upon" such a clique. This is defined as follows:

Definition 1 *An LC α is said to be* dependent *on clique $\mathcal{C}$ iff some variable in the form template of α is resolved by unification with a word instantiating a template of some LC $\beta \in \mathcal{C}$.*

The interesting point is that, considering the indeterminate lexical correspondences in a language, all and only those which are dependent on some clique can be definitely resolved, and therefore learned. This picture of the morphological network in natural languages has implications for the limits of natural morphology, particularly considering the learnability question. A learning algorithm for morphological relations in complex networks will also be considered in the paper.

References

Ford, A., and Singh, R. 1991. Propédeutique morphologique. *Folia Linguistica* 25(3–4):549–575.

Hammarström, H., and Borin, L. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2):309–350.

Neuvel, S., and Fulop, S. A. 2002. Unsupervised learning of morphology without morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning 2002*. Association for Computational Linguistics.

Neuvel, S. 2003. *Metamorphology*. Ph.D. Dissertation, The University of Chicago.