Networks of morphological relations

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

Sylvain Neuvel
MOZ, Montréal

Abstract

Whole Word Morphology does away with morphemes, instead representing all morphology as relations among sets of words, which we call *lexical correspondences.* This paper presents a more formal treatment of Whole Word Morphology than has been previously published, demonstrating how the morphological relations are mediated by unification with sequence variables. Examples from English are presented, as well as Eskimo, the latter providing an example of a highly complex *polysynthetic* lexicon. The lexical correspondences of Eskimo are operative through their interconnection in a network using a symmetric and an asymmetric relation. Finally, a learning algorithm for deriving lexical correspondences from an annotated lexicon is presented.

1 Introduction

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.

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 these do not bode well for its utility as a cognitive model of morphology. In this paper a different approach to morphology and its 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). Whole Word Morphology trades morphemes for relations among word sets, formalizing the latter as its morphlogical primitives which we call *lexical correspondences* (LCs). One goal of this paper is to provide a more formal outline of WWM than has been published previously, which we do in Sections 2 and 3. This treatment shows that the main mechanism of the important relations in WWM is a generalized sort of unification with both individual and sequence variables.

Going further, 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 solution in the form of a more complex WWM model, which we put into our formal framework in section 4. The main problem with polysynthetic languages (here Eskimo is used as an example) is that the lexical correspondences may often be indeterminate—having no unique unifier.

From this treatment it emerges that lexical correspondences for a complex lexicon can be considerd as a network with two sorts of connections, one symmetric and one asymmetric. The symmetric relation is that wherein two LCs share one template, and LCs naturally cluster into cliques by this relation. More crucial to the WWM model is the asymmetric relation wherein one LC "depends upon" such a clique. An interesting point is that, considering the indeterminate lexical correspondences from a lexicon, 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 deriving lexical correspondences from an annotated lexicon is presented in section 6. As yet it is designed only for the simpler lexicons such as English, but it can certainly be expanded to function on polysynthetic lexicons and facilitate learning the morphological networks thereof.

2 Preliminaries

In order to formalize the ideas of Whole Word Morphology (here for the first time, it seems), we require a scheme of term unification involving individual and "sequence" variables, following e.g. (Kutsia 2007; Kutsia, Levy, and Villaret 2007). This will bear great similarity to the established study of *word equations,* e.g. (Abdulrab and Pécuchet 1989; Schulz 1993). Firstly, let us recall from basic linguistics that a *word* in a language, though escaping our firm definition, is essentially a sequence of elements which we traditionally call *consonants* and *vowels.* These two sorts of elements apply equally well to a word in its orthographic form in an al-

phabetic writing system, or its phonological form expressed using phones or phonemes.

For the purpose of our formalization, we require a two-sorted alphabet of individual constants $\mathcal{A} = \mathcal{C} \cup \mathcal{V}$. Let us specify that we also have individual variables *Var* in each of these two sorts, so that $Var = Var_C \cup Var_V$. For our convenience, the constants of both sorts will be written using standard letters of the English alphabet when we elect to use standard orthography, or using International Phonetic Alphabet characters when writing words in phonological form. In any case, vowel variables will be denoted with V_i where i is a positive integer, and consonant variables will be denoted C_i. In addition to the individual constants and variables, we also require a set Ξ of *sequence variables* denoted $X_i, Y_i, Z_i \ldots$ which will be defined as objects that can unify with sequences of things. Sequence constants are also possible to include, but will be redundant since they end up equivalent to a sequence of individual constants.

We can next define a *template* as a mixed sequence of variables (either sequence or individual) or constants. We have no need to define terms in the usual way because we do not require any function symbols. Technically, to be perfectly formal we are using a single function with flexible arity to define a template, and then a *word* is defined as a ground template (a sequence of constants).

A *substitution* σ is defined as a mapping from Var_C to $Var_C \cup \mathcal{C}$, from Var_V to $Var_V \cup \mathcal{V}$, and from sequence variables Ξ to finite possibly empty sequences of elements of $\mathcal{A} \cup Var \cup \Xi$. A substitution σ_1 is said, as usual, to *unify* templates $\mathcal{X}$ and $\mathcal{Y}$ just in case $\sigma_1(\mathcal{X}) \mapsto \mathcal{Y}$. In case $\mathcal{Y}$ is a word, a unifier is called a *match.*

Our aim here is to model *morphology* of a natural language as a system of relations among lexical entries. A *lexicon* $\mathcal{L}$ for a language is a set of lexical entries, where each lexical entry e_i is a data structure consisting of a word w_i (strictly defined in form only) together with some information about it. We can be somewhat noncommittal about this information, but it at least includes a syntactic category and a meaning. Let us also be vague about how precisely to specify these, so that the model of morphology is maximally flexible in accommodating linguistic frameworks.

Definition 1 *A* lexical correspondence *(abbrev. LC) is a pair of non-ground templates (not words)* $\langle \mathcal{X}, \mathcal{Y} \rangle$ *obeying the conditions below. A lexical correspondence holds in lexicon* $\mathcal{L}$ *if we can formulate a pair of sets of lexical entries* $\langle le_{\mathcal{X}}, le_{\mathcal{Y}} \rangle$ *such that:*

1. $le_{\mathcal{X}}$ *and* $le_{\mathcal{Y}}$ *contain an equal number (at least two) of entries.*
2. *For the set* $le_{\mathcal{X}}$*, there is a match* σ_i *for each word* w_i *from each lexical entry such that* $\sigma_i(\mathcal{X}) \mapsto w_i$*.*
3. *For the set* $le_{\mathcal{Y}}$*, each word* w_i *from each lexical entry is such that* $\sigma_i(\mathcal{Y}) \mapsto w_i$*.*
4. *All lexical entries in* $le_{\mathcal{X}}$ *have the same syntactic category assignment, and respectively for* $le_{\mathcal{Y}}$*.*
5. *For each entry* $e_i \in le_{\mathcal{X}}$ *there is a unique corresponding entry* $e_j \in le_{\mathcal{Y}}$ *such that one specified semantic relation describes the correspondence of meaning between all the pairs. Succinctly we can think of this as a* semantic bijection *between* $le_{\mathcal{X}}$ *and* $le_{\mathcal{Y}}$*.*
6. *Templates* $\mathcal{X}$ *and* $\mathcal{Y}$ *must have at least one common element.*

3 Morphology with lexical correspondences

Definition 2 *Let us define a lexical correspondence* $\langle \mathcal{X}, \mathcal{Y} \rangle$ *for the English lexicon written orthographically, in the following way:*

$$\mathcal{X} = X_1 ceive$$
$$\mathcal{Y} = X_1 ception$$

This correpondence does hold of the English lexicon, since we find a set $le_{\mathcal{X}}$ including members *receive, deceive, perceive, conceive,* and corresponding set $le_{\mathcal{Y}}$ including members *reception, deception, perception, conception,* and these sets together with the templates obey all the conditions of Def. 1. The required matching unifiers exist, for sequence variable X_1. The members of $le_{\mathcal{X}}$ are verbs, while each corresponding member of $le_{\mathcal{Y}}$ is a noun referring to an instance of the act meant by the verb.

The above type of lexical correspondence was called a *word-formation strategy* by (Neuvel and Fulop 2002), because it is *deterministic.* Given a word from $le_{\mathcal{X}}$ such as *deceive,* only one word can possibly be formed from template $\mathcal{Y}$ by means of the matching unifier, namely *deception.* This important property is typical of lexical correspondences in lexicons which express such "one-slot" morphology—in which only one piece or edge of a word is changed to create a related word.

4 Morphology with polysynthesis

Neuvel (Neuvel 2003) analyzed the morphology of West Greenlandic Eskimo as a quintessential case of so-called "polysynthetic" morphology—roughly this means languages with long words and numerous slots for meaningful material to be added. For instance, one lexical entry consists of the verb /saniuqqutpuŋa/ 'I go by' (using IPA characters for Eskimo) while a related one consists of the verb /saniuqquttaqpuŋa/ 'I usually go by,' putting it into a habitual aspect.

Definition 3 *Let us define a lexical correspondence* $\langle \mathcal{X}, \mathcal{Y} \rangle$ *for the Eskimo lexicon written phonemically, which governs the habitual aspect in a wide range of words:*

$$\mathcal{X} = X_1 Y$$
$$\mathcal{Y} = X_1 taqY$$

This type of lexical correspondence has been called *multivalent* (Neuvel 2003) because its templates contain more than one sequence variable, and this generally makes the word equations nondeterministic.

Proposition 4 *Any template which has the form* $X\mathcal{Z}Y$ *with sequence variables surrounding a (possibly empty) subtemplate* $\mathcal{Z}$ *will match with any word with some part matching* $\mathcal{Z}$*.*

No proof is necessary, recalling that a sequence variable will match with the empty sequence. A problem from this fact arises with any word having more than one part matching $\mathcal{Z}$—the word equation then has no unique solution. E.g. examining the above Eskimo word meaning 'I go by,' we cannot determine where the sequence taq should be inserted.

Neuvel outlined some important facts about the Eskimo lexicon that can ultimately resolve the above matching problem. Chiefly, there are other lexical correspondences in Eskimo which are related to the LC of Def. 3 in an important way.

Definition 5 *Let us now define a lexical correspondence* $\langle \mathcal{X}, \mathcal{Y} \rangle$ *for Eskimo which governs the formation of nominal forms of the verbs:*

$$\mathcal{X} = X_2 puŋa$$
$$\mathcal{Y} = X_2 tuq$$

where $\mathcal{X}$ *is a template for 1st person verbs and* $\mathcal{Y}$ *is a template for nominalized verbs.*

The lexical set $le_{\mathcal{X}}$ for the above includes the entry /uqaqpuŋa/ 'I say,' for which the substitution $X_2 =$ uqaq provides the only match. This kind of LC may be called *monovalent* because there is only one sequence variable in the templates, with the result that the LC is deterministic.

More importantly, there are words which match the templates of both Def. 3 and Def. 5, including /saniuqqutpuŋa/. From Def. 5, the relevant match is provided by the substitution $X_2 =$ saniuqqut. We now use this as a constraint in the solution of our original problematic word equation

$$X_1 Y = \text{saniuqqutpuŋa}$$

Now by unifying variables where possible across the two LCs, $X_1 = X_2 =$ saniuqqut, so we can use Def. 3 with the additional constraint to create the word /saniuqquttaqpuŋa/. Notice how the utility of a multivalent LC was increased by unification with a deterministic LC. This can be recognized as an asymmetric "dependency" relation, defined in the sequel.

5 The lexical network of Eskimo

Following are three verb forms for "eating" in Eskimo: /nirivuŋa/ 'I eat'; /nirivutit/ 'you eat'; /nirivuq/ 'he/she eats'. The theory of Whole Word Morphology demands that every lexical correspondence is a binary relation. Thus these three words fall into the lexical sets of three interrelated LCs, defined here.

Definition 6 *Let us define the lexical correspondences which govern the relations between*

1. *1st and 2nd person verbs*

$$\mathcal{X} = X_1 C_1 uŋa; \qquad \mathcal{Y} = X_1 C_1 utit$$

2. *1st and 3rd person verbs*

$$\mathcal{X} = X_2 C_2 uŋa; \qquad \mathcal{Y} = X_2 C_2 uq$$

3. *3rd and 2nd person verbs*

$$\mathcal{X} = X_3 C_3 uq; \qquad \mathcal{Y} = X_3 C_3 utit$$

Examining these three LCs, we observe that templates of 1 and 2 are equivalent up to alphabetic variation (they don't merely unify), i.e. $\mathcal{X}_1 = \mathcal{X}_2$. Similarly, $\mathcal{Y}_2 = \mathcal{X}_3$ and $\mathcal{Y}_3 = \mathcal{Y}_1$, so the three LCs are cyclically connected by the symmetric relation of *sharing a template.* They thus form a *clique* within the entire network of Eskimo lexical correspondences, so far as we have seen.[1] The Eskimo lexicon contains numerous monovalent (deterministic) lexical correspondences, which tend to be interrelated in cliques like the above. Overall, however, the lexicon is a collection of entries and cliques, and may not generally be entirely connected by the relation of sharing a template.

Numerous lexical correspondences in Eskimo, however, are multivalent and nondeterministic. Knowledge of such a correspondence, such as Def. 3, is of little utility in the derivation of new or unknown words. We have shown above how unification with a monovalent LC can make a multivalent LC more useful, so long as the two LCs share a lexical entry. We next define this asymmetric "dependency" relation.

Definition 7 *A multivalent lexical correspondence* $\langle \mathcal{X}_1, \mathcal{Y}_1 \rangle$ *is said to be* dependent *on a monovalent clique of LCs just in case for some LC* $\langle \mathcal{X}_2, \mathcal{Y}_2 \rangle$ *in the clique there is some lexical entry* $e_i \in le_{\mathcal{Y}_1}$ *or* $le_{\mathcal{X}_1}$ *such that* w_i *unifies with either* $\mathcal{X}_2$ *or* $\mathcal{Y}_2$*, and the word equation* $w_i = \mathcal{X}_2 = \mathcal{X}_1$*, respectively* $w_i = \mathcal{Y}_2 = \mathcal{Y}_1$*, has a unique solution (for whichever pair of templates unify across LCs).*

It seems likely to us that in a natural language with lexical correspondences that are multivalent and nondeterministic, each of these will have to depend on some (clique of) monovalent LCs.

6 Induction of lexical correspondences

In typical models of morphological learning, as surveyed by (Hammarström and Borin 2011), the input is raw text and the output is a linguistic analysis into roots and affixes, which all have to be put into underlying forms. This "morpheme acquisition problem" is quite complex, comprising numerous nontrivial subproblems, and is known to be NP-complete (Ristad 1994). Modeling morphology as a set of lexical correspondences makes for a completely different learning model, involving only a deterministic algorithm with no need for statistical learning. It is not so much induction as it is deduction—using a simple algorithm one can arrive at the final model without any nondeterministic steps or model selection. We thereby completely avoid acquiring morphemes, worrying about underlying forms etc. The main caveat is that the algorithm cannot be carried out on raw text; some annotation indicating the syntactic categories of the words is required at the very least, and information about the meanings is preferable to have.

[1] In reality this clique could be larger, but we are limited in our illustrations here.

An algorithm to perform this "deduction" of the morphology from part-of-speech tagged text was implemented in (Neuvel and Fulop 2002). Here follows a simplified and formalized outline of this method:

1. From known lexicon $\mathcal{L}$ select a pair of sufficiently similar words w_1, w_2 from lexical entries e_1, e_2; e.g. the two words have several constants in common, either in a sequence or some other pattern.
2. Record two sequences $\text{Diff}_1, \text{Diff}_2$ which comprise, respectively, all constants from w_1, w_2 which are not shared. Sequence or individual variables should be recorded as needed to mark the respective positions in w_1, w_2 which do have the same constants. E.g. the earlier case $w_1 =$ receive, $w_2 =$ reception will yield $\text{Diff}_1 = X\text{ive}$, $\text{Diff}_2 = X\text{ption}$.
3. Record two sequences Sim_1, Sim_2 comprising those constants which are shared between w_1, w_2, using different variables as needed to mark the positions of the respective elements in Diff_1, Diff_2. E.g. $\text{Sim}_1 = \text{rece}Y_1$, $\text{Sim}_2 = \text{rece}Y_2$.
4. Repeat the above steps until the entire known lexicon is considered, resulting in a database of structures comprising word pairs with sequence quartets $\text{Diff}_1, \text{Diff}_2, \text{Sim}_1, \text{Sim}_2$.
5. For each pair $\langle \text{Diff}_1, \text{Diff}_2 \rangle$ occurring at least twice in the database, collect all the corresponding word pairs which also have the same corresponding syntactic categories and the same semantic bijections. These word pairs will form the basis for deriving a lexical correspondence. E.g. for the example at hand we will collect English word pairs $\langle \text{receive}_V, \text{reception}_N \rangle$, $\langle \text{deceive}_V, \text{deception}_N \rangle$, $\langle \text{perceive}_V, \text{perception}_N \rangle$, $\langle \text{conceive}_V, \text{conception}_N \rangle$. Traditional parts of speech are shown as the syntactic category subscripts.
6. Compare all pairs of similarity sequences $\langle \text{Sim}_1, \text{Sim}_2 \rangle$ connected to the word pairs in the preceding collection. From this we derive one new sequence pair, in which all constants not shared in all the similarity sequences are substituted by the same variables in Sim_1 and Sim_2. E.g. the preceding collection of word pairs will yield the sequence pair $\text{Sim}'_1 = X_2\text{ce}Y_1$, $\text{Sim}'_2 = X_2\text{ce}Y_2$.
7. Finally we unify Diff_1 with Sim'_1 to get one template $\mathcal{X} = X_2\text{ceive}$ of a lexical correspondence, and unify Diff_2 with Sim'_2 to get the other template $\mathcal{Y} = X_2\text{ception}$. We now recognize the LC that was provided in Def. 2.

Table 1 presents 14 LCs which were derived using an implementation of this algorithm operating on a tagged text of *Moby Dick* (Neuvel and Fulop 2002). This type of implementation only works for the "one-slot" morphology typical of English and the Indo-European languages. The algorithm above, however, is sufficiently general that it would be able learn the Eskimo LCs discussed previously given the properly annotated Eskimo language corpus. What it does not yet have the facility to learn is the dependency relation, which is all-important for the application to languages with a significant proportion of multivalent lexical correspondences.

Table 1: Lexical correspondences induced from *Moby Dick*

LCs		
1st template	2nd template	Examples
Xed_{PP}	Xe_{V}	baked/bake
Xed_{PP}	X_{V}	directed/direct
Xs_{Np}	X_{Ns}	helmets/helmet
Xing_{GER}	Xed_{PP}	walking/walked
Xing_{GER}	Xs_{V3s}	walking/walks
Xness_{Ns}	X_{ADJ}	short/shortness
Xly_{ADV}	X_{ADJ}	quick/quickly
Xest_{ADJ}	X_{ADJ}	hardest/hard
Xs_{V3s}	X_{V}	jumps/jump
Xer_{ADJ}	X_{ADJ}	harder/hard
$\text{Xless}_{\text{ADJ}}$	X_{Ns}	painless/pain
Xing_{GER}	Xy_{ADJ}	raining/rainy
Xed_{PP}	Xs_{V3s}	played/plays
Xings_{Np}	X_{V}	paintings/paint

Taking this step of induction does not seem, given Def. 7, to be a very difficult challenge going forward.

7 Conclusion

In this paper we have taken stock of previous work in the paradigm of Whole Word Morphology by updating, organizing, and formalizing its models and methods. Nothing of any great mathematical import has been achieved with this; rather, the main goal has been to show how a relatively simple model of natural morphology can be codified and learned by harnessing the mechanism of unification. This, in turn, expands on the theme that much of natural language can be learned using unification (Fulop 2010; 2011).

The Whole Word model of morphology, in which the lexical correspondence is primitive, has no need for morphemes in order to be expressively adequate. A further advantage, we would argue, is that this model is more likely to be cognitively relevant. The child acquisition literature is replete with findings which demonstrate children's propensity to learn morphology by "analogy," which is really the cognitive manifestation of a unification procedure. Moreover, there is little to no direct evidence that any person "knows" a specific morpheme in an abstract way, such as the abstract past tense suffix of English whose three pronunciations (so-called allomorphs) are [d, t, əd]. And going back to the *receive/reception* pairs discussed earlier, which morpheme would we need to know to accomplish this? A suffix *-tion* which attaches and simultaneously changes a vowel and substitutes a 'p' for a 'v'? Such notions may serve the Bloomfieldian linguist, but in natural language processing they generalize poorly, and in cognitive science they strain credulity.

References

Abdulrab, H., and Pécuchet, J.-P. 1989. Solving word equations. *Journal of Symbolic Computation* 8:499–521.

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

Fulop, S. A. 2010. Grammar induction by unification of type-logical lexicons. *Journal of Logic, Language and Information* 19:353–381.

Fulop, S. A. 2011. Erratum to: Grammar induction by unification of type-logical lexicons. *Journal of Logic, Language and Information* 20:135–136.

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

Kutsia, T.; Levy, J.; and Villaret, M. 2007. Sequence unification through Currying. In Baader, F., ed., *Term Rewriting and Applications*, volume 4533 of *LNCS*. Berlin: Springer. 288–302.

Kutsia, T. 2007. Solving equations with sequence variables and sequence functions. *Journal of Symbolic Computation* 42:352–388.

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.

Ristad, E. S. 1994. Complexity of morpheme acquisition. In Ristad, E. S., ed., *Language Computations*, volume 17 of *DIMACS*. American Mathematical Society. 185–198.

Schulz, K. U. 1993. Word unification and transformation of generalized equations. *Journal of Automated Reasoning* 11:149–184.