Two-Level Morphology with Composition

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen

Xerox Palo Alto Research Center
Center for the Study of Language and Information
Stanford University

1. Limitations of "Kimmo" systems

The advent of two-level morphology (Koskenniemi [1], Karttunen [2], Antworth [3], Ritchie *et al.* [4]) has made it relatively easy to develop adequate morphological (or at least morphographical) descriptions for natural languages, clearly superior to earlier "cut-and-paste" approaches to morphology. Most of the existing "Kimmo" systems developed within this paradigm consist of

- linked lexicons stored as annotated letter trees
- morphological information on the leaf nodes of trees
- transducers that encode morphological alternations

An analysis of an inflected word form is produced by mapping the input form to a sequence of lexical forms through the transducers and by composing some output from the annotations on the leaf nodes of the lexical paths that were traversed.

Comprehensive morphological descriptions of this type have been developed for several languages including Finnish, Swedish, Russian, English, Swahili, and Arabic. Although they have several good features, these Kimmo-systems also have some limitations. The ones we want to address in this paper are the following:

(1) **Lexical representations tend to be arbitrary.** Because it is difficult to write and test two-level systems that map between pairs of radically dissimilar forms, lexical representations in existing two-level analyzers tend to stay close to the surface forms.

This is not a problem for morphologically simple languages like English because, for most words, inflected forms are very similar to the canonical dictionary entry. Except for a small number of irregular verbs and nouns, it is not difficult to create a two-level description for English in which lexical forms coincide with the canonical citation forms found in a dictionary.

However, current analyzers for morphologically more complex languages (Finnish and Russian, for example) are not as satisfying in this respect. In these systems, lexical forms typically contain diacritic markers and special symbols; they are not real words in the language. For example, in Finnish the lexical counterpart of *otin* 'I took' might be rendered as *otTa1I1n*, where *T*, *a1*, and *I1* are an arbitrary encoding of morphological alternations that determine the allomorphs of the stem and the past tense morpheme. The canonical citation form *ottaa* 'to take' is composed from annotations on the leaf nodes of the letter trees that are linked to match the input. It is not in any direct way related to the lexical form produced by the transducers.

(2) **Morphological categories are not directly encoded as part of the lexical form.** Instead of morphemes like *Plural* or *Past*, we typically see suffix strings like +*s*, and +ed, which do not by themselves indicate what morpheme they express. Different realizations of the same morphological category are often represented as different even on the lexical side.

These characteristics lead to some undesirable consequences:

I. **Generation is more cumbersome and less efficient than analysis.** Because the information about morphological categories is available only on the leaf nodes of the trees, many paths through the structure may have to be tried before the right one is found. Some ways around this problem have been invented (Barton [5]) but in practice their use is limited.

II. **Annotated letter trees cannot be minimized.** Although letter trees, annotated with morphological information, are a kind of finite-state network, they cannot be minimized because all the information associated with the leaf nodes would get lost when the branching tails are merged.

The approach that we describe in this paper overcomes these problems and allows a representation of morphological information that maps more easily to the representation found in traditional lexicons. On this basis we have constructed morphological analyzers for English and French (with Carol Neidle) at Xerox PARC.

2. Desiderata

We follow two simple principles:

(1) **Inflected forms of the same word are mapped to the same canonical dictionary form.** This applies to both regular and irregular forms. For example, in our English analyzer the surface forms *happier* and *better* are directly matched with the lexical forms *happy* and *good*, respectively, rather than some nonwords.

(2) **Morphological categories are represented as part of the lexical form.** Instead of encoding morphological categories such as *Plural, Comparative, 1stPerson* as annotations on strings that realize them, we include them directly in the lexical representation. Consequently, our two-level representation of *happier* and *better* are:

lexical level	happy	+Comp	+Adj
surface level	happi	er	0
lexical level	good	+Comp	+Adj
surface level	bett	er	0

Figure 1

The stems are presented as the lemmas found in a dictionary, followed by morphological tags. *0* serves here as the epsilon symbol. Because there is no need to have other annotations on the lexicon trees, problems I and II in Section 1 have been eliminated. Lexical forms are always sequences of morphemes in their canonical representation.

The only obstacle to this approach is that the rules that constrain the surface realization of lexical forms become more difficult to write when there is little or no similarity between the two levels of representation. Designing such rules and understanding their interactions is a hard task even with the computational assistance provided by a complete compiler for the two-level formalism (Karttunen et al. [6]).

As the distance between lexical and surface form increases, the mapping is easier to describe by allowing one or more intermediate levels of representation. The solution we adopted combines the two-level rule formalism with the cascade model of finite-state morphology discussed by Kaplan & Kay [7].

3. Composition of two-level rules

Our formal understanding of finite-state morphology is based on the demonstrations that both rewriting rules and two-level rules denote regular relations on strings (Kaplan [9]). The correspondence between regular relations and finite-state transducers and the closure properties of regular relations provide the computational and mathematical tools that our approach depends on. One of the earliest results of finite-state morphology is the observation

that regular relations are closed under composition (Johnson [8], Kaplan&Kay [7], Kaplan [9]). Consequently, a single transducer can be constructed whose behavior is exactly the same as a set of transducers arranged in an ordered feeding cascade:

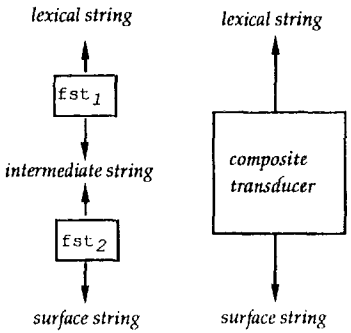

Figure 2

This observation was originally made about transducers corresponding to phonological rewrite rules, but it applies to regular relations or transducers no matter how they are specified. Although regular relations in general are not closed under intersection, the subclass of relations denoted by standard two-level rules is closed under this operation (Kaplan [9]). Thus `fst1` and `fst2` in Figure 2 may represent either a single two-level rule or the intersection of any number of rules.

When the relationship between lexical and surface forms is complex, the descriptive task of setting up rules that relate the two levels can be simplified by decomposing the complex relation to a series of less opaque matches. For efficient recognition and generation, the resulting cascade can be reduced to a single transducer. Although it would be possible in principle to produce the same single transducer directly from two-level rules, we have found many cases in our descriptions of English and French where the composition approach is not only easier but also linguistically more justified. We describe one such case in detail.

4. French compound plurals

French plurals can be formed in a variety of ways. Some of the most common patterns are illustrated in Figure 3.

We omit here the actual two-level rules; what Figure 3 illustrates is simply the joint effect of several rules that constrain the realization of the plural morpheme and the shape of the stem in regular nouns. Note that the constraints here are local; the stem and the plural morpheme are in a fixed position with respect to each other.

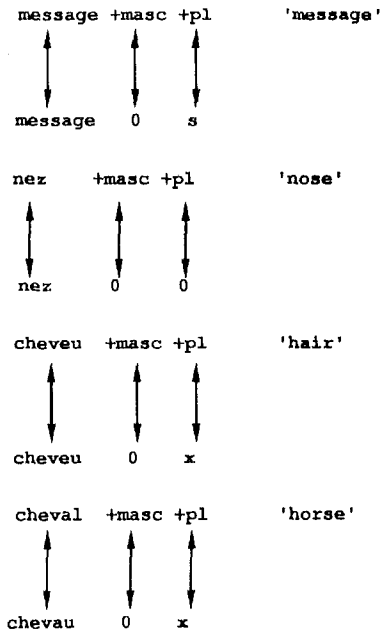

Figure 3

In compound nouns and adjectives, several patterns are possible: (1) only the first part of the compound is marked for the plural, (2) both are, (3) none are or (4)

only the last is. The possible patterns and some examples are given in Figure 4.

The interesting cases are those in which the first part needs to be pluralized. In a simple two-level system, the information about plural formation summarized in Figure 3 would have to be rewritten and adapted so that the rules could apply over a distance in the position just before the hyphen.

No plural marking at all

un je-ne-sais-quoi 'a certain something'
des je-ne-sais-quoi

Plural marking on the first compound

un chef-d'oeuvre 'masterpiece'
des chefs-d'oeuvre

Plural marking on the second compound

une mini-jupe 'mini-skirt'
des mini-jupes

Plural marking on both compounds

une porte-fenêtre 'French window'
des portes-fenêtres

Figure 4

The simple rules for regular plural formation illustrated in Figure 3 do not work for first parts of compounds because the affected elements are not in the same configuration relative to each other. Although it is possible to modify the rules, the new versions would be rather complicated and do not capture the simple fact that the plurals *portes* and *fenêtres* in *portes-fenêtres* in themselves are regular, the only thing that is special about the word is that plurality is expressed in both parts of the compound.

We avoid these complications by creating a cascade of two-level rules in which the first stage is only concerned with the plurals of compounds. It starts from a lexical form in which the words are marked for the pattern that they take and creates an intermediate level in which the information about number and gender is distributed over the agreeing parts. This is illustrated in Figure 5 for the masculine plural of *social-démocrate*, a word in which both parts get pluralized.

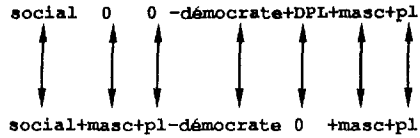

Figure 5

The effect of the first stage of rules is to copy the morphological tags from the end of the compound to the middle whenever the +DPL (double plural) diacritic is present.

The second layer of rules applies uniformly to simple nouns as well as compounds. In the case at hand, the two plurals in *sociaux-démocrates* are realized in the regular way, as shown in Figure 6.

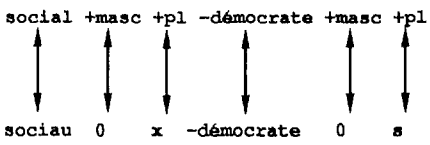

Figure 6

By first intersecting the rules in each set and then composing the results in the way shown in Figure 2, we end up with a transducer that eliminates the intermediate level altogether and maps the lexical representation directly to the correct surface form, and vice versa. Figure 7 illustrates the final result.

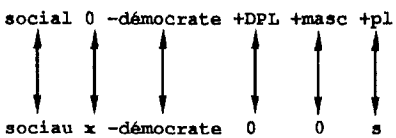

Figure 7

The representation in Figure 7 fulfills the desiderata laid out in Section 2 except that it contains a special diacritic +DPL that marks the behavior of *social-démocrate*, with respect to plural formation. In the next section, we show how that diacritic can be eliminated.

5. Composition with the lexicon

By choosing the canonical dictionary form as the lexical form in our English and French analyzers and by including morphological categories directly as part of that representation, we have eliminated the need for additional annotations in the lexical structure that are common in existing Kimmo systems. We can treat the letter tree as a simple finite-state network in which all morphological information is carried on the branches of the tree and not on the leaves.

Taking this idea one step further, we may think of the lexicon as a trivial first stage in a cascade of transducers that maps between the lexical and the surface levels. The second stage is the two-level rule system. In the case of our analyzers for English and French, the rule system starts out with three levels but reduces to two by intersection and composition. The final stage is the composition of the rule system with the lexicon.

This progression of pushing the original Kaplan & Kay [7] program to its logical conclusion is depicted in Figure 8.

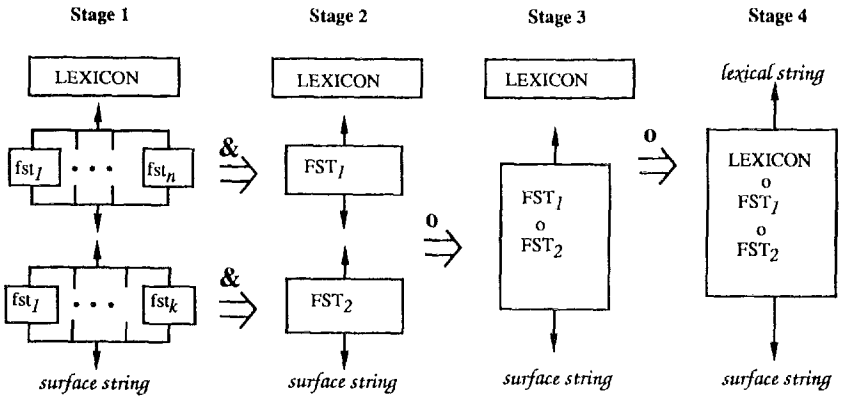

Figure 8

Figure 8 sketches the construction of our morphological analyzers for English and French. Arrows labeled with & represent intersection, arrows marked with o stand for composition. (We have simplified this picture slightly by omitting the composition of small bookkeeping relations that are necessary to model properly the interpretation of epsilon transitions in two-level rules.)

Stage 1 consists of two parallel two-level rule systems arranged in a cascade, as illustrated in Section 4. In Stage 2, the rules on each level have been intersected to a single transducer. Stage 3 shows the composition of the two-level rule systems to a single transducer and Stage 4 represents the final result: a transducer that maps sequences of canonical dictionary forms and morphological categories to the corresponding surface forms, and vice versa. Although the conceptual picture is

quite straightforward, the actual computations to produce the structures can be resource intensive, in some cases quite impractical.

At the last stage, when the idiosyncratic behavior of particular lexical items has been taken into account in the composition of the lexicon with the rule transducers, all morphological diacritics such as the +DPL tag for French nouns with double plurals can be eliminated because the rules that depend on them have been applied. In full compliance with our desiderata in Section 2, the resulting transducer maps, among other things, *social-démocrate+masc+pl* directly to *sociaux-démocrates*, and vice versa.

6. Discussion

Finite-state morphology rests on the observation that ordinary morphological alternations involve regular relations. This is the basis of the early work by Kaplan and Kay [7] on converting ordered rewrite rules to a cascade of transducers and the parallel transducers of Koskenniemi's two-level model [1]. In recent times the two-level model has been more popular. It has turned out (Karttunen [10]) that parallel two-level constraints are even expressive enough to account for phenomena that require rule ordering in the classical phonological rule formalism. But there is no computational or theoretical reason to insist on two-level descriptions. Because the mathematical properties of rewrite and two-level rules are now well-understood (Kaplan [9], Ritchie [11]), we can compose any *n*-level description to just two levels. In our work on English and French morphology we came across many instances in which the introduction of an extra level is both practical and linguistically motivated. The case of French compound plurals is a typical example.

Our success in composing the rule system with the lexicon (Stage 4 in Figure 8) is due to a number of fortunate characteristics that morphological alternations and lexicons of natural languages seem to have even though they are not necessary or even probable from a formal point of view. We at least were surprised by some of our results. The most important of these delightful discoveries are:

(1) **Small case studies can be misleading.** The composition of a rule transducer against a lexicon containing a handful of words is so much larger than the input lexicon that one is tempted to conclude that the method can never succeed on a large scale. However, this blowup seems not to occur when the lexicon is already large.

(2) **Intersections and compositions of rule transducers tend to be large, but not nearly as large as they might be.** The result of intersecting a few dozen two-level rules may have thousands or tens of thousands of states, but not trillions as the worst case scenario predicts. Many rules tend to apply either in quite similar or in quite different environments. The finite state machinery can represent such patterns without multiplying state sets.

(3) **Composition with the lexicons reduces the complexity of rule interactions.** It might turn out that the composition of a large lexicon with an even larger rule transducer is bigger than either one of the input structures. In reality, the size of the result seems to be somewhere in the middle. The rules constrain the realization of all possible lexical forms. In the composition, their scope is restricted to just the forms that actually exist in the language. It turns out that this restriction makes the result smaller rather than larger even though the lexicon itself is a very irregular collection of forms.

The fact that it is possible to construct a lexical transducer for the whole language raises interesting theoretical issues. In linguistics it is commonly assumed that lexi-

cal entries and the rules for realizing them exist independently from one another. That assumption is, of course, also the starting point of the work that we are reporting about in this paper. The initial separation between the lexicon and the rules is useful in constructing a system for word recognition and generation. The rules are, in a sense, a decomposition of a very complex mapping between lexical and surface forms to a set of simpler relations that we can comprehend and manipulate. But in the construction of the final result individual rules and the distinct lexicon disappear. The rules play no role at all in the actual generation and recognition process. They are needed only for the purpose of enlarging the lexicon, although other acquisition methods can be envisioned. The rules are true generalizations about the two-level lexicon constructed with them but they are not a part of it.

In linguistics the psychological reality of rules is often taken to be established by the observation that a simple listing of all forms would be not only implausible but even impossible, given that the brain must have some storage limitations. The general organization of the system like the one we have described suggests that the role of rules might be quite different. Instead of being essential for the production and comprehension of speech, the rules that linguists are trying to discover may be—if they exist in the mind at all—only secondary reflections on the generalizations that can be encoded in the finite-state lexical structure itself.

References

[1] Koskenniemi, K. *Two-level Morphology. A General Computational Model for Word-Form Recognition and Production*. Department of General Linguistics, University of Helsinki. 1983.

[2] Karttunen, L. KIMMO: a General Morphological Processor. *Texas Linguistics Forum*, 22:217-228. 1983.

[3] Antworth, E. L. *PC-KIMMO: a two-level processor for morphological analysis*. Occasional Publications in Academic Computing No. 16, Summer Institute of Linguistics, Dallas, Texas. 1990.

[4] Ritchie, G. D., G. J. Russell, A. W. Black, S. G. Pulman. *Computational Morphology. Practical Mechanisms for the English Lexicon*. The MIT Press, Cambridge, MA. 1991

[5] Barton, E., R. Berwick, E. Ristad. *Computational Complexity and Natural Language*. The MIT Press, Cambridge MA. 1987.

[6] Karttunen, L., K. Koskenniemi, and R. M. Kaplan. A Compiler for Two-level Phonological Rules. In Dalrymple, M. *et al. Tools for Morphological Analysis*. Center for the Study of Language and Information. Stanford University. Palo Alto. 1987.

[7] Kaplan, R. M. and M. Kay. Phonological rules and finite-state transducers [Abstract]. *Linguistic Society of America Meeting Handbook*. Fifty-sixth Annual Meeting, December 27-30, 1981. New York.

[8] Johnson, C. Douglas. *Formal Aspects of Phonological Description*. Mouton. The Hague. 1972.

[9] Kaplan, R. M. Regular models of phonological rule systems. Alvey Workshop on Parsing and Pattern Recognition. Oxford University, April 1988.

[10] Karttunen, Lauri. Finite-State Constraints. In the *Proceedings of the International Conference on Current Issues in Computational Linguistics*, June 10-14,

1991. Universiti Sains Malaysia, Penang, Malaysia. 1991.

[11] Ritchie, Graeme D. Languages Generated by Two-level Morphological Rules. Research Paper 496. Department of Artificial intelligence, University of Edinburgh, 1990. To appear in *Computational Linguistics.*

Résumé

Cet article décrit une nouvelle utilisation des transducteurs reguliers en analyse morphologique. Les systèmes Kimmo standard se composent d'un lexique en forme d'arborescence de caractères (trie) avec des sommets finals annotés d'information morphologique et un ensemble de transducteurs qui transcrivent les représentations lexicales en formes fléchies. Bienque ces systèmes soient supérieurs aux techniques antérieures d'analyse morphologique "cut-and-paste", ils ont un certain nombre de désavantages: les formes lexicales sont souvent arbitraires et différentes des lemmes d'un dictionnaire normal; l'analyse morphologique n'est pas encodée directement dans la forme lexicale. Le résultat est que la synthèse est souvent plus ardue que l'analyse et que les structures ne sont pas optimales.

Les analyseurs morphologiques construits à Xerox-PARC pour le français et l'anglais se basent sur deux principles simples: 1. les formes fléchies d'un même mot se basent sur un même lemme; 2. les catégories morphologiques font partie intégrante de la forme lexicale. Ainsi les formes lexicales sont toujours des séquences de morphemes. Il est difficile d'achever ces deux résultats désirables dans le cadre d'une description classique à deux niveaux parce que la distance entre les formes lexicales et les formes de surface est longue et très difficile à décrire avec un seul ensemble de régles phonologiques à deux niveaux.

Il est possible de résoudre ces problèmes en exploitant d'une façon plus approfondie les principes de la phonologie à deux niveaux. En guise d'exemple, l'article décrit une cascade de règles à deux niveaux qui permet une description simple du pluriel des mots composés en français. La première serie de règles insère des annotations de nombre et de genre après chaque élement de mots à pluriel double (*social-démocrate ~ sociaux-démocrates*), la séconde série de règles détermine la réalisation du nombre et du genre réquise par les racines.

Les caractéristiques mathématiques des transducteurs réguliers sont bien connues. Elles permettent la combinaison de transducteurs correspondant à des systèmes de règles à deux niveaux par composition et par intersection. Ainsi il est possible de réduire un système à niveaux multiples à un seul transducteur qui contrôle simultanément toutes les alternances morphologiques d'un langue. Vu que dans les lexiques anglais et français développés à Xerox toute l'information morphologique est codée directement avec le lemme, il est possible d'aller plus loin et de composer le lexique entier avec les règles. Le transducteur résultant, un lexique à deux niveaux, transcrit les formes lexicales directement en formes de surface et vice versa. Les règles ne sont utilisées que dans la phase de construction. L'analyse et la synthèse ne font usage que du transducteur lexical résultant.