Morphological Interfaces to Dictionaries

Michael Maxwell
Linguistic Data Consortium
3600 Market St, Suite 810
Philadelphia, PA 19104
Maxwell@ldc.upenn.edu

William Poser
Linguistics, University of Pennsylvania
3600 Market St, Suite 501
Philadelphia, PA 19104
billposer@alum.mit.edu

Abstract

Languages with complex morphologies present difficulties for dictionaries users. One solution to this problem is to use a morphological parser for lookup of morphologically complex words, including fully inflected words, without the user needing to explicitly know the morphology. We discuss the sorts of morphologies which cause the greatest need for such an interface.

1 Introduction

When it comes to dictionaries, not all languages are created equal. Quite apart from the fact that more effort has been put into lexicography for some languages than for others, languages vary in how they lend themselves to word look up.

Generations of English-speaking students have been told, when they were uncertain how to spell a word, to "look it up in the dictionary." How one is supposed to look up a word when one does not already know how to spell it has been the source of much distress for those same students.

For English, the chief obstacle to dictionary lookup is the orthography.[1] But for other languages, the structure of the language itself is the problem, and in particular the language's morphology. Unless the dictionary user is explicitly familiar with that morphology, determining the citation form of a given word can be quite difficult.

One solution, at least for electronic dictionaries, is to create an interface that uses a morphological parser to find the root or stem of the full word provided by the user, and then automatically look up that form (or its citation form), thereby shifting the need to explicitly know the morphology (and the choice of citation form) from the user to the computer. Such interfaces are described in Breidt and Feldweg 1997, Prószéky and Kis 2002, Streiter, Knapp, Voltmer et al. 2004, etc.

But building a morphological parser is a non-trivial task, and a simpler solution—where possible—would be preferable. In this paper, we discuss the sort of morphology that makes a parser interface especially desirable.

2 Morphology and Citation Forms

We need to clarify here that we are concerned with how difficult dictionary lookup is for "average" users, that is, for users who may not be *overtly* familiar with the morphology of the language. Linguists and (usually) language teachers are often familiar enough with the morphology that they can compute the citation form from any arbitrary inflected form, but many other users will not be able to do so.

Certain kinds of morphology can make it difficult for average users to find the citation form of an inflected word. Usually this is inflectional morphology, simply because forms related by derivational morphology are often given separate listings. However, some languages have productive and regular derivational morphology, so that forms related by derivational affixation may not in fact be listed. Furthermore, in some languages (such as Athabaskan languages, see below) the derivational and inflectional morphology interact so as to make finding a citation form especially difficult. Finally, the boundary between inflectional and derivational morphology is not always clear—whether to a linguist or to the end user.

We now turn to the specifics of how morphology can impede dictionary lookup. For languages with any degree of morphology, one form of the paradigm for a given part of speech is usually chosen as the citation form. Problems may arise for words which lack the chosen form (e.g. *pluralia tantum* words, such as *scissors*). In any case, users must generally be told what form to look for.

Of course, for languages with only a small amount of productive morphology, it does not take much sophistication to come up with a citation form from an inflected form. For English, the

[1] Orthography presents difficulties for dictionary lookup is difficult in languages which are not written alphabetically, and for which the lexical entries therefore cannot be alphabetized. While we do not address that issue in this paper, specialized front ends to dictionaries have been used for lookup in such languages; see e.g. Bilac, Baldwin and Tanaka 2002.

citation form of an inflected verb is generally found by stripping off one of a handful of suffixes (and sometimes undoing other spelling rules). Irregular verbs present complications, but their frequency makes them unlikely candidates for lookup, except by language learners. At any rate, irregular words can be placed in minor entries, separately alphabetized from the major entries, and cross-referencing the latter.

In practice, users may not even need to know how to remove the suffixes, since when searching for *walks* or *walking* they will find *walk*, and generally make the connection.

If a language is exclusively suffixing, not even a large amount of inflectional affixation need stand in the way of lookup. If the user cannot figure out the citation form of a word, he can simply look up the first few letters to find the entry. Thus, even languages like Turkish or Quechua often pose little problem for lookup. (Nevertheless, for some users, it may not be obvious that the citation form thus found corresponds to the inflected word, see e.g. Corris, Manning, Poetsch et al. 2004: 47.)

More problematic for lookup is prefixation.[2] Since dictionary words are usually alphabetized from the beginning of the word to the end (left to right in most writing systems), in theory the user would have to strip prefixes before doing lookup. An obvious work-around would be to alphabetize words in (exclusively) prefixing languages from right to left. Alternatively, the dictionary could provide an index alphabetized from right to left, where the user could find the citation form, then look up that form in the main part of the dictionary. To our knowledge, this solution has not been employed, although this may be due to the paucity of exclusively prefixing languages.

The reverse alphabetization solution would not work for languages which employ both prefixing and suffixing, such as Tzeltal (Mayan). But even here the situation is not too bad if the number of prefixes is small, as in fact it is in Tzeltal: the common prefixes are *h-/k-*, *a-/aw-*, and *s-/y-*, and stripping these probably does not present much of a problem to most users of the *Vocabulario tzeltal de Bachajon* (Slocum and Gerdel 1965).

The real problem for languages having both prefixes and suffixes arises when the language has a large number of prefixes, or when the language productively employs compounding or incorporation, which can have the same effect for dictionary lookup as productive prefixation. German is a notorious example of the difficulties occasioned by compounding, and Nahuatl is an example of a language having incorporation.

In Nahuatl, indefinite direct objects can often be incorporated into the verb: *chi:lkwa* 'to eat chili' is composed of the verb stem *kwa* 'to eat', preceded by the incorporated noun *chi:l* 'chili'. The naïve user may succeed in finding the incorporated noun in a printed dictionary, but may be at a loss to decipher the rest of the word, since *kwa* is not a noun suffix in Nahuatl.[3]

A greater difficulty for the average dictionary user is nonconcatenative morphology, such as infixes, partial reduplication, and templatic morphology. In Tagalog, for example, there is an affix *-um-* marking actor focus, which is infixed following a word-initial consonant (Schachter and Otanes 1972). Furthermore, the Tagalog imperfective aspect is indicated by partial reduplication. Thus the word *bumibili* is a form of the verb root *bili* 'to buy', where the reduplication is *bi-*, and the infix *-um-* is stuck into the middle of this reduplicated syllable.

In some cases, the user can (or should!) be expected to understand this and deal with converting *bumibili*, say, to the appropriate citation form. And in fact dictionary writers often assist by providing partly inflected forms: in the case of Tagalog, for instance, citation forms generally include the focus affixes. But as the complexity of the morphology increases, relying on the user to guess the citation form from an inflected form becomes less of an option. At the same time, explicitly including multiple inflected forms in the dictionary becomes cumbersome, even impossible.

In the following subsections, we detail difficulties occasioned by the particular morphologies of Semitic and Athabaskan languages.

2.1 Semitic Languages

Arabic, like most other Semitic languages, employs templatic morphology, in which affixes composed of vowels can be interdigitated between consonants of the root. Affixation can also modify the root consonants, frequently by gemination. For example, a typical Arabic root *ktb* can appear in inflected forms as diverse as *katab*, *kattab*, *ktatab*, *ktanbab*, and *kutib* (Spencer 1991). Some of this morphology is derivational, and some inflectional, but it all poses a problem for users.

Moreover, Arabic is ordinarily written without many of the vowels. While this may ease the

[2] If the citation form is prefixed, this may also cause problems for alphabetization, since many words may fall into the same section of the alphabet. This problem is well-known, but is not the focus of our discussion.

[3] In practice, this problem in Nahuatl is ameliorated by the fact that incorporation is not highly productive. Therefore the most common cases of incorporation should arguably be listed in the dictionary.

problem caused by the interdigitated vowels, it means that the dictionary user may have more difficulty distinguishing root consonants from affixal consonants, since the vowels are not present in the written form to help parse the word.

Traditional Arabic dictionaries have been root based; that is, the head word of a lexical entry is the root, with all derivational and inflectional morphology removed. Listed derived forms appear as subentries under a given root (and inflected forms which must be listed are generally included as variant forms within the subentry for a given derived form). Because of the difficulty 'undoing' Arabic derivational and inflectional morphology poses for the average user, so-called "alphabetic" dictionaries have become increasingly popular. In an alphabetic dictionary, derived forms serve as headwords, so that alphabetization is done over the entire set of lexemes, whether root or stem.

Root-based dictionaries and alphabetic dictionaries each have strengths and weaknesses. A root-based dictionary gathers the information on related forms into one place, rather than scattering it throughout the dictionary, as is the case for an alphabetic dictionary. On the other hand, a root-based dictionary requires a much more explicit understanding of Arabic morphology than many users possess. Even so, finding the citation form of an irregular plural or an irregular verb in an alphabetic dictionary can be a daunting task.

In summary, Arabic morphology forces the dictionary writer to choose between a root-based format and an alphabetic format; both approaches have their disadvantages. Similar problems obtain for other languages with templatic morphologies. Fortunately, these problems can be overcome by interposing a morphological parser between the user and the electronic dictionary.

2.2 Athabaskan Languages

The difficulties that Athabaskan languages pose for dictionary lookup have been detailed in Poser 2002; here we give an outline of the problem for one such language, Carrier.

Like other Athabaskan languages, Carrier is predominantly prefixing, with verbs carrying numerous prefixes. Each verb can have tens or even hundreds of thousands of forms. But the sheer number of verb forms is not all that different from other agglutinative languages such as Finnish or Turkish. The real problem is that Carrier prefixes are a *mixture* of inflectional and derivational morphemes, with the derivational affixes often appearing outside of inflectional affixes.

Furthermore, it is not infrequently the case that there are prefixes which obligatorily combine with a root in a certain meaning. In effect, Athabaskan languages have discontinuous verb stems.[4] For instance, the Carrier verb "to be red" consists of the root *k'un* with the "valence prefix" *l-* immediately preceding the root and the prefix *d-* several positions to the left, giving forms like:

dilk'un	"you (sg.) are red"
duzk'un	"I am red"
hudulk'un	"they are red"
hudutilk'un	"they will be red"

Note that some subject markers follow the *d-* while others precede it. Also notice that the allomorphy sometimes collapses two affixes into a single segment (s+l → z in *duzk'un*).

For dictionaries, the implication is that there is in general no contiguous or invariant portion of the verb that can serve as the citation form. The morphology is primarily prefixal, but the existence of extensive stem variation and some suffixation means that the stem is not a good citation form, and that ordering forms form right-to-left will not keep related forms close together. Worse, the phonological material that contributes the basic meaning of the word is not, in general, contiguous. This means that any citation form will not be easily extracted by an unsophisticated user. Moreover, no simple sorting will keep related forms together.

Worse, many verb roots are highly abstract, so that a form can only be given an English translation on the basis of the root together with one or more prefixes. Examples are found in the "classificatory" verbs. For example, the verb root meaning "to give" takes distinct derivational affixes depending on the type of object being handled: ball-shaped objects, names, houses, non-count objects, long rigid objects, contents of open containers, liquids, fluffy "stuff"—and these classifiers may not be adjacent to the root.

In light of the difficult of dictionary lookup in Athabaskan languages, one approach has been to list, for each verb, a single form, as in the major dictionary of Navajo (Young and William Morgan 1987). However, this requires the user to be able to analyze a verb form and convert it to the citation form. This is a non-trivial task even for fluent native speakers; it is difficult or impossible for language learners. Indeed, the problem of dictionary use for Navajo is so acute that the Diné (Navajo) College has instituted a one semester course "Navajo Grammar and Applied Linguistics", which is largely devoted to teaching

[4] These are somewhat analogous to English verb-particle combinations such as "*bring* a matter *up*", in which the verbal inflection (and often a direct object) intervenes between the verb root and the particle. But the intervening inflectional morphology in Athabaskan is vastly more complex than that of English.

college-level native speakers of Navajo how to use the dictionary of their own language.

The other major approach to dictionary making in Athabaskan languages is to list individual morphemes. In order to use such a dictionary, the user must be able to analyze the word into root and affixes. But the root may have many shapes. For example, the root meaning "to go around in a boat" takes forms such as *ke, koh, ki, kel, keł,* and *ke?*. Although there is a pattern to these changes, it is complex if not irregular. The resulting difficulty for dictionary lookup should be obvious.

A root-based lexicon has been published for Navajo (Young, Morgan and Midgette 1992). It has the virtue of being comprehensive, and of avoiding duplication. For example, the detailed meaning of a verb root can be explained only once, in the entry for that root, rather than in each of many entries for forms derived from that root.

The problem with this approach is that it requires even more grammatical knowledge on the part of the user than traditional Athabaskan dictionaries, together with an understanding of an elaborate process for analyzing forms, looking up their components, and constructing the meaning of the form from its components. As a result, while analytic dictionaries are useful for linguists, but most people, including both language learners and native speakers, find them very difficult.

In sum, the morphological structure of Athabaskan languages forces difficult choices on the dictionary writer, and results in a steep learning curve for the user. Again, this is the sort of language structure where a morphological interface can make a crucial difference.

3 Conclusion

We have outlined ways in which the structure of languages can make a morphological parser as a front end for dictionary lookup attractive.

There are more uses to such technology than just dictionary lookup. If the morphology engine is a transducer, it can be used for generation as well as for parsing. Such a bidirectional engine can be used to generate the paradigm of any stem. While this is of little interest to native speakers, it may be of great assistance to language learners.

Another application would be to provide what amounts to a virtual interlinear text with morpheme glosses for any text in electronic form. To be sure, this text would not be disambiguated, unless a knowledgeable user put forth the effort, or unless an automatic disambiguator (tagger) was provided. Nevertheless, interlinear text, even in an ambiguous form, could be a useful for linguists and perhaps language learners.

In sum, a morphological transducer connected to an electronic dictionary can provide valuable aid for both native speakers and language learners.

4 Acknowledgements

Our thanks to Tim Buckwalter and Mohamed Maamouri of the Linguistic Data Consortium and Jonathan Amith for their comments on earlier versions of this paper.

References

Bilac, S., T. Baldwin, et al. (2002). *Bringing the Dictionary to the User: the FOKS system*. COLING-2002.

Breidt, E. and H. Feldweg (1997). "Accessing Foreign Languages with COMPASS." *Machine Translation Journal, special issue on New Tools for Human Translators* **12**: 153-174.

Corris, M., C. Manning, et al. (2004). "How Useful and Usable are Dictionaries for Speakers of Australian Indigenous Languages?" *International Journal of Lexicography* **17**: 33-68.

Poser, W. J. (2002). Making Athabaskan Dictionaries Usable. *Proceedings of the Athabaskan Languages Conference*. G. Holton. Fairbanks, Alaska Native Language center, University of Alaska**:** 136-147.

Prószéky, G. and B. Kis (2002). *Context-Sensitive Electronic Dictionaries*. COLING-2002.

Schachter, P. and F. T. Otanes (1972). *Tagalog Reference Grammar*. Berkeley, University of California Press.

Slocum, M. and F. Gerdel (1965). *Vocabulario tzeltal de Bachajon*. Mexico, Summer Institute of Linguistics.

Spencer, A. (1991). *Morphological theory : an introduction to word structure in generative grammar*. Oxford, UK ; Cambridge, Mass., Basil Blackwell.

Streiter, O., J. Knapp, et al. (2004). *Bridging the Gap between Intentional and Incidental Vocabulary Acquisition*. ALLC/ ACH 2004, Göteborg University, Sweden.

Young, R. W., W. Morgan, et al. (1992). *Analytical Lexicon of Navajo*. Albuquerque, University of New Mexico Press.

Young, R. W. and S. William Morgan (1987). *The Navajo Language: a Grammar and Colloquial Dictionary.* Albuquerque, University of New Mexico Press.