Computational Morphology and the Teaching of Indigenous Languages

Michael Gasser
School of Informatics
Indiana University

Abstract

Computer-assisted language learning is by now so common around the world as to be something of a default, and the teaching of the indigenous languages of the Americas is already benefiting from the new technology. **Intelligent** computer-assisted language learning relies on software that has relatively sophisticated models of the target language and/or the learner. An example is the use of a program that has an explicit model of some aspect of the grammar of the target language and can analyze or generate words or sentences. Many indigenous languages of the Americas are characterized by complex morphology, and morphology must play a significant role in the instruction of these languages. This paper describes how morphological analyzers and generators can handle the complex morphology of languages such as K'iche' and Quechua and discusses a potential application of this technology to the teaching of such languages.

Computer-Assisted Language Learning

In recent years, computers have become so important in language teaching that it hard to imagine a class without them. Students use computers to do exercises practicing what they have learned in the class, they access documents from the Internet, they interact with other learners or with native speakers of the target language on the Internet, and they write papers with word processing software that may be especially adapted to second language learning. The field of **computer-assisted language learning** (CALL) has its own conferences and its own journals, *CALICO Journal*, *Computer Assisted Language Learning*, and *Language Learning and Technology*. Computers are even a part of the language curriculum in relatively impoverished parts of the world, including regions where indigenous languages are taught as part of a bilingual curriculum or as second languages to legacy speakers or non-indigenous people.

More recently, advances in computational linguistics and artificial intelligence have led to what is referred to as *intelligent* computer-assisted language learning (ICALL; see Heift & Schulze, 2007 for an overview), which can also be considered a subfield within the area of intelligent tutoring systems and other applications of artificial intelligence (AI) and computational linguistics in education more generally (a field with its own journals, the *International Journal of Artificial Intelligence in Education* and the *Journal of Interactive Learning Research*). A common feature of applications of AI in education has been a model of the learner, based on the learner's errors; examples within ICALL are Heift (2001) and Amaral & Meurers (2007). Another possibility for ICALL is the incorporation of explicit models of the language being learned (e.g., Nagata, 2002). Such programs, or more specifically, a proposal for their incorporation into the teaching of indigenous languages, are the subject of the rest of this article.

Grammar in Second Language Instruction

Before discussing the potential use of intelligent software in indigenous language instruction, we need to consider briefly the role of grammar in language teaching more generally.

No issue has been more controversial in the field of second language teaching than that of how (or whether) grammar is to be taught. While the explicit teaching of grammar played a significant role in traditional language teaching, the influence of behaviorist psychological theory and later the ideas of Stephen Krashen (1985) called into question the value of making learners aware of the sorts of generalizations that linguists themselves make when analyzing a language. More recently, the question has been framed in terms of (1) whether second language learners benefit from focus on linguistic form, with or without the associated meaning, as opposed to focus exclusively on meaning, and (2) whether explicit instruction is superior to implicit. There is now an emerging consensus that focus on form is beneficial and that explicit instruction pays off (Norris & Ortega, 2000), though many details remain up in the air. After many years of debate and research, grammar again has a secure place in the language curriculum.

But what grammar? This depends on the language. For a language such as English, the focus has naturally been on syntax: how questions are formed, the structure of passive sentences, relative pronoun choice, etc. For a language such as Spanish, a significant amount of time would normally be dedicated to morphology, in addition to syntax: conjugation, paradigms for the different verb classes, adjective-noun agreement. Among the commonly taught languages, there are ones with more complex morphology than Spanish, for example, Russian. But experience with the teaching of Indo-European languages, or other commonly taught languages such as Chinese and Japanese, does not prepare one for the teaching of a language with genuinely complex morphology, such as Swahili, Amharic, Quechua, or Nahuatl.

The Morphology of the Indigenous Languages of the Americas

Greenberg's attempt to group all of the indigenous languages of the Americas into three large families has never met with general acceptance. Most historical linguists would assign these languages to dozens of families, with many languages still considered isolates. Despite this diversity, there are some quite striking tendencies, some of which appear to be areal features, that is, the result of contact for long periods during which languages have borrowed features from one another.

One such tendency, common to many regions within the Hemisphere, is morphological complexity, especially for verbs. Many of the languages of the Americas form words through extensive **agglutination**, whereby sequences of morphemes are strung together, each contributing a meaning to the whole. Another common means of combination is **reduplication**, whereby some portion of a root is copied.

In the area of **inflectional morphology**, verbs may go well beyond what is familiar with European languages. As with most Indo-European languages, verbs usually agree with their subjects in person and number (an important exception is the Oto-Manguean languages). However, in many cases they also agree with their objects, as in this Quechua example, in which both the subject ('you') and the object ('I') are marked on the verb.

1. *qhawa-wa-nki*
   see-1sOBJ-2sSBJ
   'you see me'

Verbs may also be explicitly inflected for tense-aspect-mood, which in Indo-European is often fused with the subject agreement morphemes. In this K'iche' example, the *x-* prefix indicates "completive" aspect.

2. *x-e'-opan-ik*
   COMPL-3pABS-arrive-INTR
   'they arrived'

However, it is in the area of **derivational morphology** that morphologically complex languages differ most from the commonly taught European languages. Again it is for verbs that the possibilities are the most extensive. Through affixation or reduplication, a verb root may be converted to an adjective or noun stem or to a range of verb stems with different meanings. A Quechua verb root, for example, can participate in dozens of different verb stems through a set of derivational suffixes conveying notions of manner or modality, and strings of several of these suffixes are possible within a single word. In the following word, the root is *wayk'u-* 'cook' (example from Herrero & Sánchez de Lozada, 1978):

3. ***wayk'u**rquysimuwankimanchu*
   'could you please go help me cook?'

Languages with this degree of morphological complexity are referred to as **polysynthetic**.

Finally, a significant number of indigenous languages of the Americas make use of **noun incorporation** within verbs, the inclusion of lexical morphemes representing one or more arguments of the verb. Here is an extreme example from the Uto-Aztecan language Huichol (Wixárika; Iturrioz & Gómez López, 2006); two lexical morphemes within the word are indicated in bold, *xakɨ́* 'dish' and *tui* 'bring'.

4. *mepɨtetatsiheti***xakɨ́***te***tui***rietsitɨaxɨaximekaitɨkaku*
   'however, they already wanted to start ordering us to **bring** them the **dishes**'

Morphologically complex languages are not simply languages that do the work that syntax does within words instead of between words. Very often they make distinctions that are not made in other languages. One important domain for derivational morphology is that by which different arguments of the verb (participants in the event referred to) are added, deleted, foregrounded, or backgrounded. Languages such as English and Spanish have only one systematic way to accomplish this: through the passive, by which the object of a transitive verb is foregrounded and the subject eliminated or backgrounded. A language like K'iche', on the other hand, has up to five options. These will be discussed below in the section *Learning about argument promotion, demotion, and addition in K'iche'.*

Languages with complex morphology have far more possible word forms than a learner could ever learn individually. In a language such as Quechua a given verb root can appear in well over a million word forms, most of which have little likelihood of actual occurrence but which speakers of the language are nonetheless expected to be able to produce and understand. The implications of these facts for language learning are that the learner must somehow master the compositional aspects of word structure. Studies of the acquisition of the polysynthetic language Inuktitut as a first language (Allen, 1994) show that children are well on their way to mastery of the complex aspects of word structure, including noun incorporation, by the time they are four years old.

What does this mean for the teaching of a morphologically complex language such as K'iche' or Quechua to second-language learners? Clearly, achieving even rudimentary proficiency in such languages requires morphological sophistication: the ability to combine lexical and grammatical information in the production of a novel word form and to parse a novel word form into its component morphemes. One way or another, learners need to be made aware of the structure of words and the significance of the different components of words.

Computer-assisted morphology instruction has not attracted much attention within CALL, probably because of the domination of English and other Indo-European languages. One example is GLOSSER, a vocabulary learning assistant for Dutch speakers learning French (Nerbonne, Dokter, & Smit, 1998). While reading in the target language, students can call up GLOSSER on an unfamiliar word, and it provides them with a morphological analysis of the word as well as with lexical information. Another example is the intelligent tutoring environment for Turkish developed by Güvenir and Oflazer (1994). Their system allows the user to specify a set of grammatical features and then finds all of the words in a corpus that are characterized by those features. Note that Turkish is an agglutinative language with morphological complexity comparable to many indigenous American languages. Both of these systems, like the proposal that is the focus of this paper, build on advances in the field of computational morphology. The next section provides an overview of this area of research.

Computational Morphology

Computational morphology is the branch of computational linguistics concerned with word structure. Two kinds of processing are of interest: **morphological analysis**, by which a **surface** word form is analyzed into a **lexical** representation, consisting of the word's component morphemes or grammatical features, and **morphological generation**, by which a lexical representation is converted to a surface word form. Consider the K'iche' verb *kinawilo* 'you see me'. One sort of morphological analysis would simply segment the word into the morphemes that make it up.

5. *kinawilo* → k-in-aw-il-o

A more abstract output lexical representation would indicate the lexical and grammatical significance of the morphemes. One style of representation appropriate for this is a **feature structure description** (FSD), consisting of pairs of features and their values, where values are either atomic expressions like 'feminine' or 'false' or feature structure descriptions in their own right. The word *kinawilo* could be represented at the lexical level with the FSD shown on the right below.

6. *kinawilo* → [root='il',
 abs=[pers=1,num=sing],
 erg=[pers=2,num=sing,-form],
 tam=incmpl]

This represents the information that the root of the verb is *il*, the absolutive argument (direct object) is first person singular, the ergative argument (subject) is second person singular informal, and the tense-aspect-mode is incompletive. (For an explanation of absolutive and ergative case, see the section *Learning about argument promotion, demotion, and addition in K'iche'* below.)

Morphological generation works in the opposite direction:

7. [root='il', → *kinawilo*
 abs=[pers=1,num=sing],
 erg=[pers=2,num=sing,-form],
 tam=incmpl]

The information required to perform morphological analysis or generation is of three types. There may be a **lexicon** of roots or stems, which combine with grammatical morphemes to yield surface word forms. **Morphotactics** refers to constraints on the order and class of the morphemes making up a word within a particular category. For example, the morphotactics of K'iche' transitive verbs specifies the following minimal set of morphemes: tense/aspect/mood+ergative+absolutive+root. **Alternation rules** are responsible for the variation in the forms that morphemes take in the presence of other morphemes. For example, each of the K'iche' ergative prefixes has two forms, one used before consonants, the other before vowels. Together, knowledge of alternation rules, morphotactics, and the forms of roots or stems in the lexicon represent the morphology of a given language.

Computational morphology has been an active research area within computational linguistics since the 1970s, with the discovery that morphological analysis could be handled efficiently by **finite state transducers** (FST). An FST is a network of states and transitions between them, and an analysis of a word is a path through the network. Each of the transitions along the path specifies a correspondence between input characters (or phones) and output characters, and the transducer converts sequences of input characters to sequences of output characters. One very useful property of FSTs is that they can be inverted; the same transducer that implements analysis (surface to lexical representation) for a given rule also implements generation (lexical to surface representation) through simple reversal of the input and output characters. Another useful property is composition; a sequence of FSTs, converting a surface representation to a lexical representation with various intermediate stages, can be collapsed into a single FST which behaves equivalently to the original sequence of FSTs.

A more in-depth discussion of finite state morphology is beyond the scope of this paper (see Beesley & Karttunen, 2003 for an overview). The main point to note for our purposes is that a single FST can encode all of the information represented in the alternation rules, morphotactics, and root lexicons for a particular word class in a given language. The framework adopted in the system described here is an extension of the conventional FST framework whereby networks take as input and yield as output a feature structure description as well as a sequence of characters (Amtrup, 2003; Gasser, 2009). This allows FSTs to implement analysis and generation as illustrated above in sentences 6 and 7.

Computational Morphology and Language Teaching

As discussed above, the point of explicit grammar instruction is to make the student aware of the grammatical structure in utterances. For morphology, this means showing students how morphotactics and alternation rules combine to yield the surface forms of words in the target language. But it's one thing to illustrate structure and processes; it's another to find ways for students to practice them. Without the benefits of computational morphology, the best that can be offered is a set of canned examples. *Intelligent* computer-assisted morphology learning (ICAML) makes use of morphological analyzers and generators which embody generalizations, so it offers the possibility of processing completely unexpected forms produced by the student or collections of features generated by the student. That is, it is both robust — it responds to forms not in its database — and extensible — it easily accommodates new lexical items and new rules.

We have already seen above how computational morphology has been applied in two L2 teaching systems, as a means of analyzing unfamiliar words encountered by the student during reading (Nerbonne, Dokter, & Smit, 1998) and of locating words in a corpus that match a grammatical description provided by the student (Güvenir & Oflazer, 1994). In what follows, I describe another possible application of ICAML, one in which the student directly explores the ways in which different derivational suffixes can be applied to the same root, changing the meaning in subtle ways.

Automatic Morphological Processing in the Teaching of Indigenous Languages

A user interface for K'iche' and Quechua morphological processing

I have implemented a relatively complete morphological analyzer and generator for K'iche' verbs and a partial analyzer and generator for Quechua (a significantly more complex language with respect to morphology). The analyzers start with a surface form and output one or more analyses, each consisting of a verb root and representation of the grammatical structure of the word in the form of a feature structure description. Multiple analyses are appropriate in the case of ambiguous words. The generators start with a root and a feature structure representation of the grammatical structure of the word and output one or more surface words. Multiple words are appropriate if the grammatical description is compatible with multiple forms or if there are orthographic or phonological variants.

Figure 1 shows the current interface for Quechua. This would need to modified in several ways before it is tested on learners: the grammar behind the analysis would have to be brought in line with whatever approach is being taught, the text within the window should probably be in Quechua rather than Spanish, and the details of the interface would need to be appropriate for the learner's level. The interface allows the user to both analyze and generate words. Analysis starts with a word entered in the box at the top; clicking on the "Analizar" button then results in a root ("Raíz") and a sequence of morphemes in the yellow box ("Morfemas") and a grammatical analysis in the green box ("Propiedades gramaticales"). If the word entered is not grammatical, the message "No se puede analizar" appears in the box at the bottom of the window. The figure shows the system's analysis of the word *munanakunkichischu* 'you (pl.) don't love each other'.

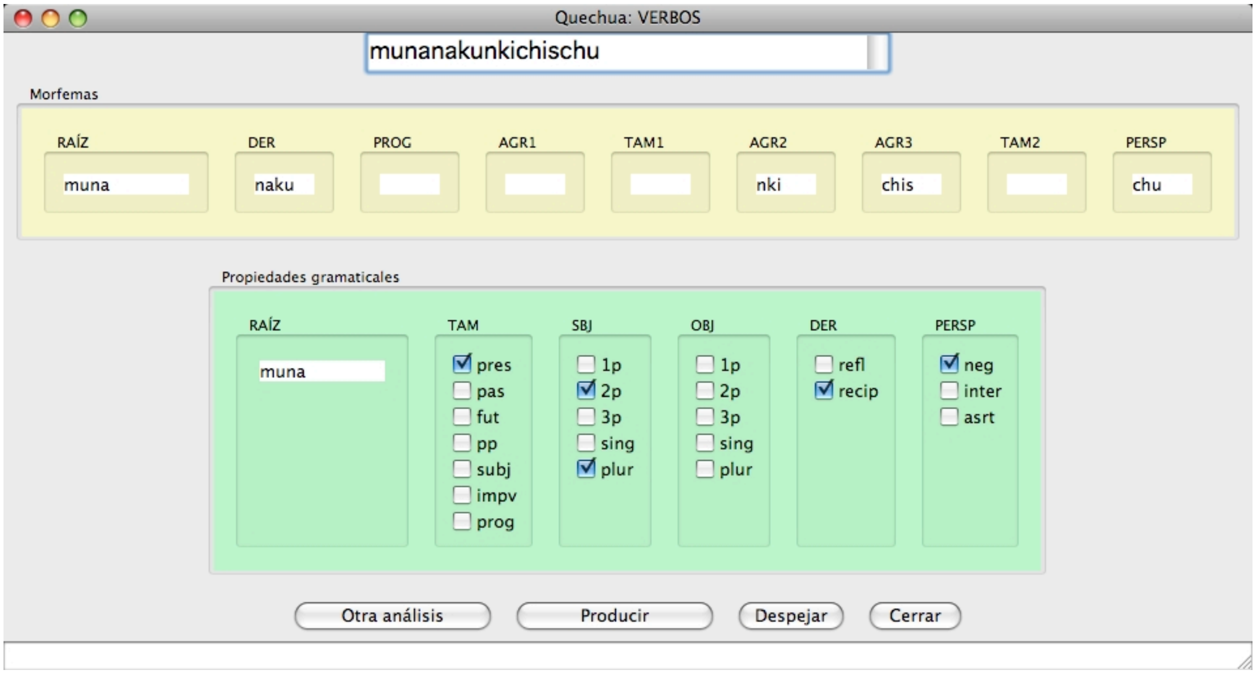

Figure 1. Morphological analysis of the Quechua verb *munanakunkichischu* 'you (pl.) don't love each other'

For morphological generation, the user enters a root and a set of grammatical features in the green box, and if the features are compatible, a surface word form appears in the box at the top of the window, and a sequence of morphemes appears in the yellow box below that.

In this form, the program could already be useful to students learning the language. Given an unfamiliar word, they could discover its structure or correct their own analysis. Given a root and a set of derivational and inflectional features, they could discover the correct form of the word or correct the word that they generate.

In the next section, I discuss how the program could also be used within a lesson on how different verbs derived from the same root are related to one another.

Learning about argument promotion, demotion, and addition in K'iche'

Unlike Indo-European languages and most of the other familiarly taught languages, Mayan languages (England, 2001) are **ergative** rather than accusative. Accusative languages distinguish two kinds of verb arguments, nominative, which are the subjects of both transitive and intransitive verbs, and accusative, which are the direct objects of transitive verbs. Ergative languages also distinguish two kinds of verb arguments, **absolutive**, which are the subjects of intransitive sentences and the direct objects of transitive sentences, and ergative, which are the subjects of transitive sentences. Consider the third person plural agreement marker on the verbs in these K'iche' sentences, shown in bold.

8. *K**e**qakunaj.* 'We cure them.'

9. *K**e**warik.* 'They sleep.'

Whereas in English, the pronoun is translated in the accusative form in the first example (*them*), in the second example, it is translated in the nominative form (*they*). In K'iche' both are in the absolutive form (-*e*-). The other third person plural agreement marker, the ergative form, *ki*, would appear in a sentence, such as the following, where it plays the role of the subject of a transitive sentence.

10. *Kuj**ki**kunaj.* 'They cure us.'

Most accusative languages offer a means, the **passive**, of promoting or foregrounding the direct object of transitive sentences and demoting, backgrounding, or deleting the subject. Many ergative languages, including Mayan languages, also offer this possibility. In K'iche' there are in fact two passive forms, with a slight distinction in meaning (López Ixcoy, 1997). One of the two possible passive forms for verbs in the class of *kuna-* 'cure' is formed with the suffix *-x*. In either case the direct object of the active transitive sentence remains in the absolutive case, in the following example, *-e-* 'they/them' (*-j* is a transitive suffix, *-ik* an intransitive suffix).

11. *Keqakunaj.* 'We cure them.'
    Kekunaxik. 'They are cured.'

As in English, the subject of the active sentence can appear as a prepositional phrase within the passive sentence (*they are cured by the doctors*), but we will ignore that possibility here.

Unlike accusative languages, many ergative languages, including the Mayan languages, also offer a further possibility for demoting or deleting an argument, the so-called **antipassive**: the *direct object* of a transitive verb, that is, the argument in absolutive case, is demoted. This results in an intransitive sentence with the original subject now in the absolutive case, in the following example, *-uj-* 'we/us', the suffix *–n* indicating antipassive for this class of verbs.

12. *Keqakunaj.* 'We cure them.'
    Kujkunanik. 'We cure (somebody).'

The details of the contexts that call for the antipassive in K'iche' (López Ixcoy, 1997) are quite subtle and complex; in fact some verbs have two different antipassive forms with slightly different pragmatic implications. These details will not concern us here.

If we think of the passive and antipassive as applying to a more simple active sentence, then both can be seen to convert a transitive to an intransitive sentence. The Mayan languages, like many other agglutinative languages from around the world, also have a way of achieving the reverse, the **causative**, which transitivizes an intransitive sentence by adding a causer argument. In K'iche' the usual causative suffix is *–(i)sa*. The subject of the intransitive sentence becomes the object of the transitive sentence; in an ergative language, this requires no change in case: in both sentences the argument is absolutive (*-e'-* 'they/them' in the following example).

13. *Ke'atinik.* 'They bathe.'
    Ke'awatinisaj. 'You bathe them.'

Thus we see that K'iche' offers up to five ways to demote, promote, add, or delete arguments of a verb: two kinds of passive, two kinds of antipassive, and causative. Learning how K'iche' "works" involves learning this system from the K'iche' perspective rather than in terms of how different source language expressions would be translated. A focus on form in this case would contrast sentences with simple underived verbs with their passive, antipassive, and causative counterparts, with respect to both form and meaning. Because the ergative-absolutive system itself should present unusual difficulties for a native speaker of an accusative language such as English or Spanish, it would be important to highlight how the case of the verb arguments shifts or remains the same in the different counterparts. In fact an exercise that highlights how causative, passive, and antipassive function would be an ideal way to become familiar with the ergative-absolutive system.

The morphological analyzer and generator for K'iche' offers the user the possibility of taking a given verb and transforming it through one of the three basic processes of causativization, passivization, and antipassivization. This could provide the basis for exercises on these processes and how they interact with ergative and absolutive case. Figure 2 shows the K'iche' interface after analyzing the word *keqakunaj* 'we cure them'. At this point the user can select one of the three features in the pink box labeled "Añadir propiedades". In each case the surface word form changes to the appropriate form referring to the same event or a similar event in the world; that is, the semantic roles of the participants are kept constant. Figure 3 shows the result of selecting antipassive given the situation in Figure 2; the surface word form has changed to *kujkunanik* 'we cure (somebody)', and the grammatical features have been adjusted to reflect this. That is, the argument in absolutive case is now first person plural, and there is no longer an ergative argument. “Otra producción” now appears on the second button because other variants of the word are possible.

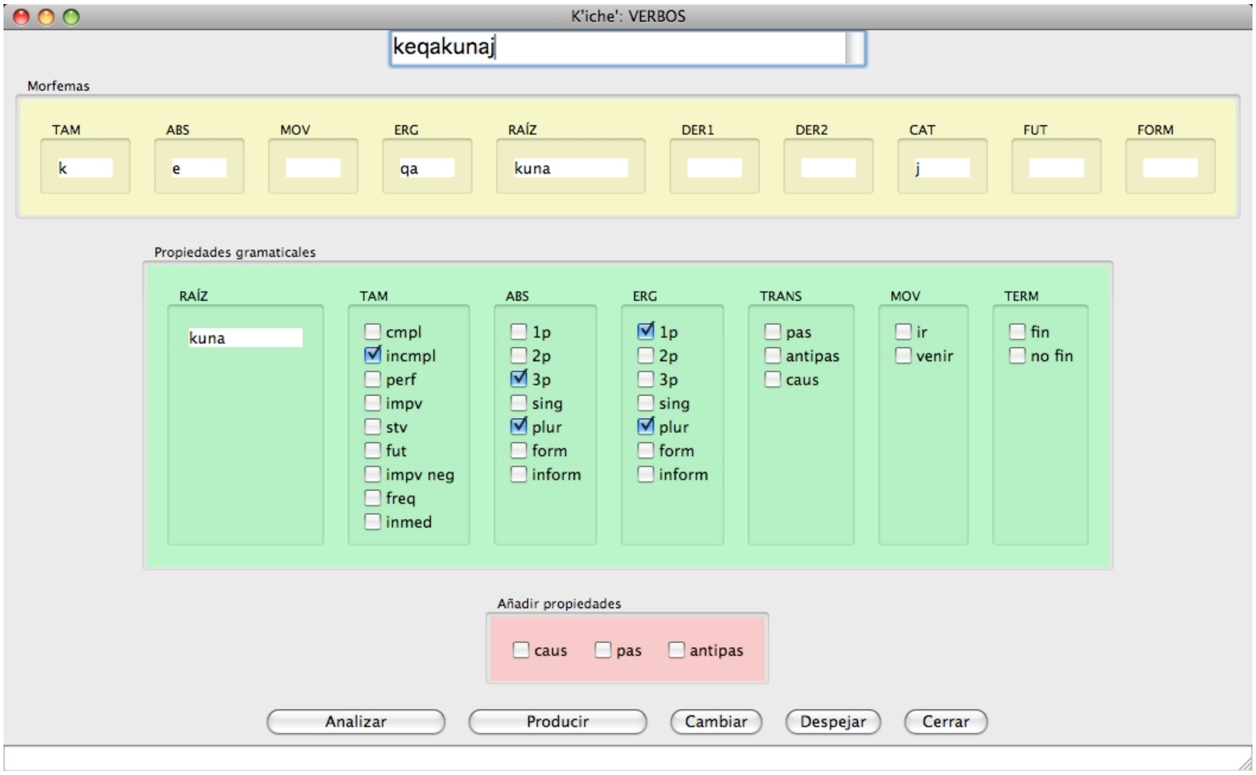

Figure 2. Analysis of the K’iche’ verb *keqakunaj* ‘we cure them’.

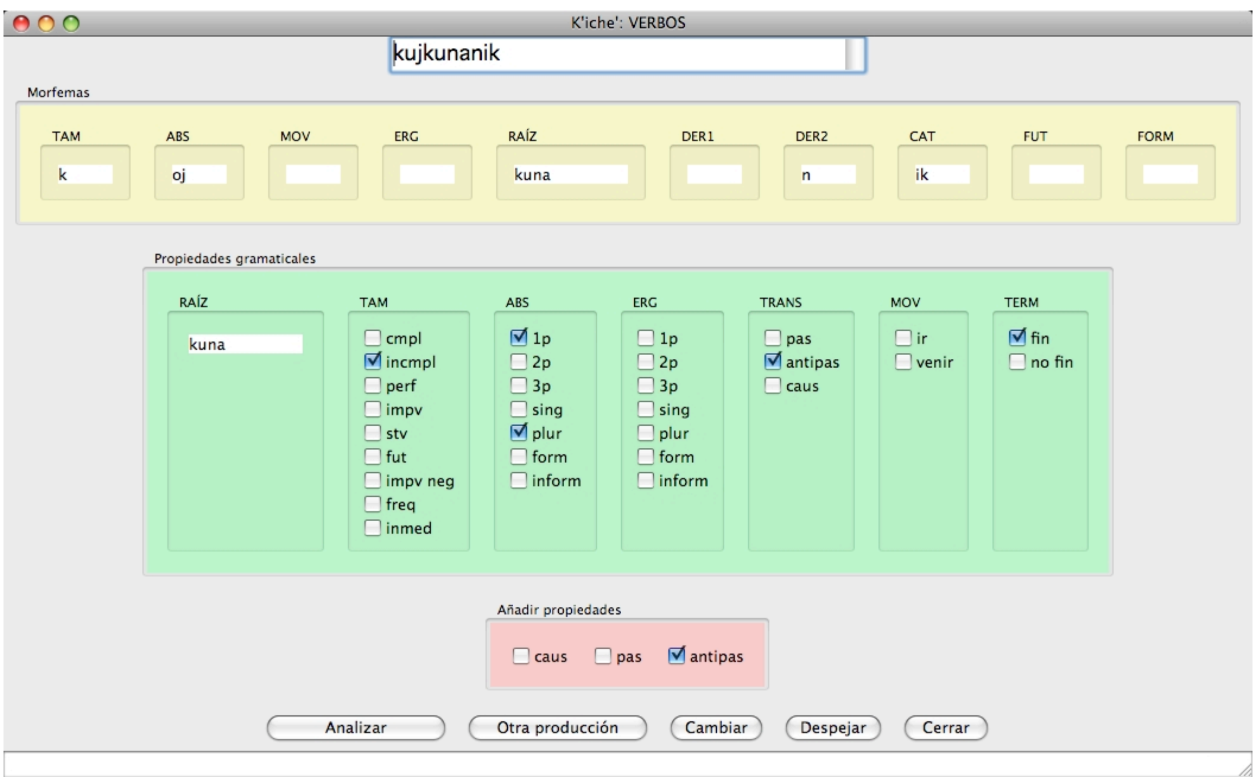

Figure 3. Conversion of the word in Figure 2 to its antipassive form *kujkunanik* ‘we cure (somebody)’

As in the Quechua example above, the details of the interface are not the point here; the interface would have to be adapted to suit a particular curriculum and a particular learner age group and level in any case. The point is to show how a system with knowledge of the structure of verbs and of how different verbs with the same root can refer to the same or similar events can provide students with useful information when they are learning these core properties of the language.

Conclusions

The indigenous languages of Latin America are very diverse, but the ones with the most speakers — Yucatec Maya and the different varieties of Nahuatl in Mexico, K'iche' in Guatemala, Aymara and the varieties of Quechua in the Andes, Guarani in Paraguay — are characterized by elaborate inflectional and derivational morphology. These languages, and many other indigenous languages, are being taught as second languages and, together with Spanish, in bilingual programs in more and more classrooms across Latin America. Research demonstrates that students learn complex grammar by being made aware of the complexity, by focusing on form as well as meaning. For morphologically complex languages, focus on form means, among other things, focusing on how verbs agree with their subjects and objects, how affixes convey different tense-aspect-mood categories, how derivational affixes and reduplication alter the meanings of roots in subtle and often language-specific ways.

Advances in computational morphology allow all of these formal distinctions (though not necessarily their precise semantics) to be encoded in programs that can analyze and generate words. In intelligent computer-assisted morphology learning, such programs can play an important role by providing students with a means of testing their knowledge of word structure, correcting their errors, and providing them with the appropriate form of a word when they are writing. The suggestions made regarding K'iche' in the last section, which to be sure still need considerable fleshing out as well as evaluation, are just one possibility. Another is a tool similar to the GLOSSER system of Nerbonne, Dokter, & Smit (1998) that would allow the student to click on an unfamiliar word in a text and have the system analyze it.

In summary, computers can best assist in teaching grammar when they know something about grammar. Computers can best assist in teaching languages such as K'iche', Quechua, and Guarani when they know something about their morphology.

References

Allen, S. (1994). *Aspects of argument structure acquisition in Inuktitut.* Amsterdam: John Benjamins Publishing.

Amaral, L., & Meurers, D. (2007). Conceptualizing student models for ICALL. In C. Conati and K. F. McCoy (Eds.), *User Modeling 2007: Proceedings of the Eleventh International Conference, Lecture Notes in Computer Science*. Wien, New York, Berlin: Springer.

Amtrup, J. (2003). Morphology in machine translation systems: efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, *18*, 213-235.

Beesley, K. R., & Karttunen, L. (2003). *Finite state morphology.* Stanford, CA, USA: CSLI Publications.

Chambers, A., & Davies, G. (Eds.) (2001). *ICT and language learning: a European perspective.* London: Taylor & Francis.

England, N. C. (2001). *Introducción a la gramática de los idiomas mayas*. Guatemala City: Cholsamaj.

Gasser, M. (2009). Semitic morphological analysis and generation using finite state transducers with feature structures. *Twelfth Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece.

Güvenir, H. A., & Oflazer, K. (1994). Using a corpus for teaching Turkish morphology. In L. Appelo and F. M. F. de Jong (Eds.), *Proceedings of the Seventh Twente Workshop on Language Technology* (pp. 28-40), Enschede, the Netherlands.

Heift, T. (2001). Intelligent Language Tutoring Systems for Grammar practice. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *6*(2). URL: `http://zif.spz.tu-darmstadt.de/jg-06-2/beitrag/heift2.htm`.

Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: parsers and pedagogues*. New York: Routledge.

Herrero, J., & Sánchez de Lozada, F. (1978). *Gramática quechua: estructura del quechua boliviano contemporaneo*. Cochabamba, Bolivia: Editorial Canelas.

Iturrioz, J. L., & Gómez López, P. (2006). *Gramática wixarika I.* Munich: LINCOM Studies in Native American Linguistics.

López Ixcoy, C. D. (1997). *Ri ukemiik ri k'ichee' chii': Gramática k'ichee'.* Guatemala City: Cholsamaj.

Krashen, S. (1985). *The input hypothesis.* London: Longman.

Nerbonne, J., Dokter, D., & Smit, P. (1998). Morphological processing and computer-assisted language learning. *Computer Assisted Language Learning*, *11*, 421-437.

Nagata, Noriko (2002). BANZAI: an application of natural language processing to Web based language learning. *CALICO Journal*, *19*(3), 583–599.

Norris, J. M. & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417-528.