Machine Translation and the Future of Indigenous Languages

Mike Gasser
Indiana University

I Congreso Internacional de las Lenguas y Literaturas Indoamericanas

Temuco, Chile, October 2006

Abstract

The Linguistic Digital Divide frustrates the promise of the Digital Revolution by making information overwhelmingly available to speakers of a small number of favored languages, by denying speakers of disadvantaged languages a role in the the creation of knowledge in the global knowledge-based society, and by marginalizing these languages themselves within their own societies. In this paper, I argue that one way to address the Linguistic Digital Divide is through a large-scale translation project based on collaboration between computers and human translators. The project, called L3, relies in part on a sophisticated machine-translation system. This system makes use of on general-purpose tools that could be used by knowledgeable native speakers of languages to bring create systems that would perform and rudimentary translation and would then learn to improve as they are exposed to more data, as these become available. I discuss the case of the first specific project undertaken with the larger project, for translation between the Mayan language K'iche' and Spanish, focusing on problems related to morphology, and I consider ways in which translation systems for other Mayan languages could build on what is being developed for K'iche' by capitalizing on features that are shared between the languages.

Language and the Digital Revolution

The Digital Revolution

The Digital Revolution that began in the last half of the twentieth century and continues to unfold today has profound implications for science, culture, education, and political and economic life. Language plays a fundamental role in this revolution since most of the information being created and accessed is still in linguistic form, and people interact with the new technology mainly using language.

Particular languages and language more generally are also changed as a result of the Digital Revolution. They take on new functions as they become the means by which people express themselves and interact with each other and with machines in new ways (chat, text messaging, blogging, etc.). And as enormous quantities of text in particular languages are published and archived, the body of material that constitutes the cultural record of these languages (as well as a source for research) grows at an unprecedented rate.

The Global Digital Divide

However, as with previous technological revolutions that had significant cultural and socioeconomic consequences, the Digital Revolution is not reaching all populations in a uniform fashion. Among the early analysts of the process were those utopianists who believed that, because of the new possibilities for virtual communities that would bypass the existing power structures (Rheingold, 2001), the Digital Revolution would lead inexorably to a more democratic and just world. Contrary to these early hopes, however, it is possible to argue that on a global scale the Digital Revolution has actually increased some of the gaps that separate the more advantaged North from the less advantaged South. These gaps make up what has been called the "Global Digital Divide" (Chinn & Fairlie, 2004).

For the purposes of this paper, three aspects of the Global Digital Divide are of concern: (1) uneven access to information, (2) uneven participation in the global "knowledge-based society," and (3) the decline in linguistic and cultural diversity. The dramatic increase in availability of information that many, if not most, people in the Global North benefit from has been extended only to restricted elites within the Global South. At the same time, information and knowledge have taken on a new significance in economic and political life in an increasingly globalized world (David & Foray, 2003). This leaves large numbers of people more isolated than ever. The imbalance also works in the opposite direction, however. The digitally disenfranchised not only fail to benefit from the newly available stores of information; they are also unable to contribute to them. Thus the privileged fail to learn from the experiences of these non-participants and also fail to take their perspective into consideration in planning actions that may affect them. Finally, while the advent of the Internet encourages certain forms of diversity, the concentration of the resources of the digital world in a restricted number of societies may be accelerating certain forms of cultural homogenization within the digital community as well as the ongoing marginalization of many cultures and languages.

Inequities in these areas have received attention from the United Nations and its associated agency responsible for these matters, the International Telecommunications Union (ITU). In their World Summit on the Information Society (WSIS) meetings (ITU, 2003), they elaborated a series of "key principles" which included the following:

1. The ability for all to access and contribute information, ideas and knowledge is essential in an inclusive Information Society.

2. Cultural diversity is the common heritage of humankind. The Information Society should be founded on and stimulate respect for cultural identity, cultural and linguistic diversity, traditions and religions, and foster dialogue among cultures and civilizations.

The WSIS principles document concludes with a highly optimistic statement regarding the global potential of the Digital Revolution:

> All individuals can soon, if we take the necessary actions, together build a new Information Society based on shared knowledge and founded on global solidarity and a better mutual understanding between peoples and nations.

The Linguistic Digital Divide

If the Global Digital Divide is to be bridged, as envisioned in the WSIS process, an obvious problem is that the majority of the world's people still do not enjoy access to the information and communication technology (ICT) itself. Many people are working on this technical aspect of the Digital Divide, and a number of low-cost solutions are being developed. The Divide is not simply a technical one, however. The technology brings with it a whole range of cultural and linguistic biases that would need to be overcome (Paolillo, 2005; Pimienta, 2005).

The **Linguistic Digital Divide** refers to the relative advantage of certain languages and language communities over others with respect to ICTs. There are two sides to the Divide: the availability of information on the Internet and the availability of digital tools of various sorts. There is some disagreement concerning the actual proportion of documents in different languages on the World-Wide Web (FUNREDES/Unión Latina, 2006; O'Neill et al., 2003), but the general picture is clear. Roughly half of the pages are in English, and ten to fifteen other languages account for all but about 1-5% of the remainder. Missing from this list are numerically very important languages such as Hindi, Bengali, Indonesian, and Arabic. Clearly number of speakers (even readers and writers) for a language is only weakly associated with availability of material on the Internet. When it comes to the software that Internet users and software developers are accustomed to, the imbalance is even greater (Paolillo, 2005). For most of this software, both the interfaces that users encounter and the programming and markup languages that developers use to implement the software are strongly biased toward English. Many applications such as spell-checking, grammar-checking, and information retrieval and extraction rely on basic computational linguistic work. For many languages this work has not been done, so development of the applications is inhibited, especially when it is obviously more profitable to focus on software for languages of the Global North.

In one sense, this skewed distribution in the availability of information is not so surprising; it roughly mirrors the distribution of material in the world's languages that is present in physical libraries (O'Neill et al., 2003). But it means that the Digital Revolution has changed nothing in what was already a situation that was to the great advantage of users of the favored languages.

Furthermore, the overwhelming domination of a small number of languages, especially English, when it comes to both information and software means that native speakers of other languages with some competence in one of these favored languages are discouraged from using their own languages even when they are interacting with other native speakers (Paolillo, 2005). For some language communities the result is a digital elite who interact with one another in English, French, Spanish, or Portuguese; who fail to develop a technical vocabulary in their native languages; and who isolate themselves (even more than they already have) from the masses of people within these same language communities who don't have a command of the relevant language of privilege.

For national or regional languages with significant numbers of speakers, such as Thai, Tagalog, Tamil, Gujarati, Amharic, Yoruba, and Hausa, the result may be the relegation of these languages to narrower roles within their own societies, as English or another language becomes the medium of computer-mediated written communication, which is likely to take on ever greater significance within these societies. For minority languages such as the indigenous languages of the Americas, which may already suffer from limitations on the roles they play within their societies, the consequences are even more serious.

In summary, the Digital Revolution has made available vast amounts of information in English and a small number of other languages as well as a range of tools for finding and making use of this information, has stimulated the production of new documents in these languages, and in general has furthered the democratization of knowledge within the portions of the world where these languages are spoken. However, it has so far had relatively little effect on the half of the world who either do not know one of the languages or do not have access to the technology that is required. In fact it appears that language is standing in the way of much of the "shared knowledge and ... better mutual understanding between peoples and nations" that the WSIS Principles foresee. There are three aspects to the Linguistic Digital Divide: (1) the relative lack of digital (or in many cases analogue) material in disadvantaged languages (the Knowledge Gap), (2) the relative lack of input from the disadvantaged linguistic communities in the global decision-making (the Participation Gap), and (3) the lack of computational tools that facilitate the integration of the disadvantaged languages (and their speakers) into the digital world (the Software Gap).

A role for machine translation?

What can be done to address these inequities? Let's consider first the relative amount of information published in the different languages and the relative contribution of different linguistic communities to economic and social progress on a global scale, that is, the Knowledge Gap and the Participation Gap. Both of these problems are related to the inability of native speakers (writers, readers, hearers) of one language to understand another language: native speakers of Malagasy have difficulty reading a document in English, and native speakers of

Eng.lish can make no sense at all out of a document in Malagasy (which is most likely not available to them anyway). Such problems are most obviously alleviated by translation: a translation of the English document lets the Malagasy speaker know the gist of what the English-speaking writer intended, and an English translation of the Malagasy document gives the English speaker an idea of what the Malagasy-speaking writer intended.

Of course translation is a very labor-intensive process. There are, however, reasons to be hopeful that the Digital Revolution itself can facilitate this process. First, advances in machine translation now mean that the process can sometimes be partially automated. Second, the Internet has encouraged the formation of **electronic networks of practice** (Teigland, 2003), virtual communities of people collaborating on the solving of particular tasks. Probably the most impressive of these is the community of volunteers who edit the different editions of Wikipedia (Kolbitsch & Maurer, 2006; http://www.wikipedia.org/). The hope is that, given the appropriate tools and the larger task of translation into and out the language in question, such a community of volunteers committed to the enrichment (and survival) of the language could organize around the task. This community would serve multiple purposes: (1) polishing the imperfect results of machine translation; (2) developing norms and standards for translation between the language and other languages, related and unrelated; (3) providing further input for training machine translation systems; and (4) overseeing the development of the language as a medium of communication for and about ICTs.

Beyond the issues of information access, there are the other pressures that favor the dominant languages over languages such as Kurdish, Telugu, and Inuktitut, the biases that are built into the software people use as well as the whole culture of the Internet. Part of this problem is cultural; it can only be overcome if the members of a linguistic community are comfortable with the new technology and with extending their language in the new domain in the same way this has happened with the favored languages. But the problem also has a technical side; the solution depends on software that is "friendly" to different languages. In the case of language-intensive applications such as information retrieval, some of this software development depends in turn on prior work implementing computational grammars and lexicons of the languages.

This paper proposes a large-scale project, **L3** (Learning Lots of Languages), centered on collaborative translation between selected "disadvantaged" languages of the Global South and "privileged" languages of the Global North (as well as among the languages of the South themselves). At the heart of the project is the development of a set of general-purpose tools for building machine translation systems for languages with few computational resources and the implementation of such systems for as many specific languages as possible. For each language, the subproject would require a committed team of bilingual participants, including a small number with some linguistic sophistication; most of these would work as volunteers. The translation system for a given language (or language pair) would be an ongoing effort; the system would continually evolve as new monolingual and bilingual data became available and as the language team interacted with it. Besides making available more texts in the language in question (thus addressing the Knowledge Gap) and more texts in other languages translated from the language (thus addressing the Participation Gap), the development of a machine translation system, as discussed below, would have as a side benefit basic computational linguistic work on the language that is needed for many applications (thus addressing the Software Gap).

In the rest of this paper, I discuss what sorts of hurdles would have to be surmounted to take on such an ambitious project and go into some detail about my initial experiences with our initial experience with the Mayan language K'iche'.

Machine translation

Approaches

Contemporary approaches to machine translation and machine-assisted translation, like other areas of research in computational linguistics, fall into two broad categories: **symbolic** (Nirenburg et al., 1994) and **statistical** (Brown et al, 1990). Symbolic approaches rely on explicit grammars and lexicons, usually related to one or another linguistic theory, whereas statistical approaches rely on co-occurrences discovered through exposure to large amounts of data. Statistical approaches have gained ground in recent years as it has been recognized that much of language can only be captured in the form of statistical tendencies and as large digital corpora for some languages have become available. However, these approaches are still hampered by their inability to handle grammatical phenomena such as long-distance dependencies, anaphora, and scoping and their awkwardness when it comes to phenomena captured well by linguistic rules. For these reasons, much current research is focused on ways to bridge the gap between symbolic and statistical approaches. This is not a trivial goal since the two kinds of approaches are incompatible in many ways.

For machine translation, there is another dimension along which approaches, especially symbolic approaches, differ. This concerns the degree of abstraction of the source and target language forms that are related in the system. On the one extreme are **transfer** approaches, in which the relationships are between surface forms in the two languages. On the other extreme are **interlingua** approaches, in which there is an abstract, shared level of representation between the two languages. In these approaches, source language input is analyzed into an interlingua representation, which is then converted to a corresponding target-language surface form.

Despite impressive progress in the last 15 years, particularly with the advent of statistical methods, machine translation is still far from what may have been the original goal for most researchers: quality translations requiring no human intervention. The fact is that, because of the extensive background knowledge that is behind much of human translation and that may never be something that is built into software, this goal may never be achievable. Instead, as Kay (1997) has argued, researchers should be thinking in terms of appropriate and efficient forms of human-computer collaboration in translation.

Translation from and into disadvantaged languages

Not surprisingly, most machine translation research has focused on favored languages such as English, French, German, Spanish, Russian, Dutch, Portuguese, Italian, Chinese, Japanese, and Korean. Among some of the languages of the Global South spoken over very large regions, such as Arabic, or in countries with significant research facilities, such as Hindi, Persian, and Malay, there is already a good deal of computational work, including some attempts at machine translation, and for some others, such as Amharic, there are the beginnings of computational research. However, for most of the languages of interest, there is little to go on. For the indigenous languages of the Americas, all we have are some dictionaries that have been converted to digital form, if that.

The goal of translation to and from large numbers of languages for which there are limited resources presents a number of challenges. In particular the project will have to have the following features.

1. Because statistical methods require large corpora, bilingual corpora in the case of statistical machine translation, and because such corpora do not exist for the languages of interest, these methods are ruled out, at least in the beginning of the project. For a given language, the specific project responsible for it will have to rely on symbolic methods to get rudimentary translation off the ground. That is, grammatical and lexical resources will have to be developed for the languages. To simplify this process for a large number of languages, general-purpose tools that are easily used by relative novices are called for.

2. Relatively sophisticated machine translation will almost certainly require statistical methods, however.

    a. Thus it will be desirable to elicit corpora, both monolingual and bilingual, on which the translation system can be trained. For this purpose, we will need a community of bilinguals contributors, and wiki software (Leuf & Cunningham, 2001), which makes material on the World-Wide Web editable by users, can help to facilitate this process.

    b. We will also need ways to integrate the symbolic and statistically acquired knowledge.

3. As the number of languages that are covered increases, the number of source-target language pairs increases exponentially, and it will hardly be feasible to build explicit representations for translation between all possible pairs. For this reason, the system will need to capitalize on known (and inferred) relationships between languages, so that knowledge

about translation between, say, Spanish and Guaraní, could be used to translate between Portuguese and Guaraní.

This paper focuses on some of the initial problems faced in achieving 1 and offers some tentative thoughts about 2.b and 3.

Translation in L3

Morphology

Practical computational linguistic systems, unlike typical linguistic theories, are not necessarily committed to the most parsimonious account of phenomena. Because the bottleneck in a modern system is more likely to be processing time rather than space (memory), it may be more efficient under some circumstances to build in some redundancy. Thus in a language, such as English, with relatively impoverished morphology, it may be possible to ignore morphology altogether and assign each separate word form (singular and plural of nouns; stem, *-ed*, *-ing* forms of verbs; stem, comparative, superlative of adjectives) its own lexical entry. For a language such as Spanish this approach becomes somewhat unwieldy, especially if the system is designed to translate between the language and a number of others, and for a language with more complex morphology such as is typical of native American languages, still all word forms is effectively impossible. Thus an important part of a translation project such as this is taking morphology, both inflectional and derivational, seriously.

Because each source language is also a potential target language, all morphological information must be bidirectional. That is, a surface word form will need to be parsed into its constituent morphemes and an underlying (parsed) representation of a word will need to be converted to a grammatical surface form. **Two-level morphology** (Koskenniemi, 1983) is computationally efficient approach to morphology (and phonology) that makes use of bidirectional rules and no intermediate levels of representation. However, because two-level rules are not always intuitive and because they handle non-concatenative morphology only awkwardly, a hybrid approach was used instead: wherever it is relatively simple to do so, knowledge is shared between parsing and generation, but the two processes are kept separate to some extent.

Consider the Spanish verb *almuerces*, the present subjunctive, second person singular form of *almorzar* 'to have lunch'. This word features the regular second person singular suffix for *-ar* verbs, *-es*, and in addition two predictable orthographic (one of them also phonological) changes in the stem. For parsing, we need to convert *almuerces* to the parsed representation:

(1) *almuerces* → *almorz-a-* + -SUBJ_PRES + -2S

For generation, we need to perform the reverse the process.

(2) *almorz-a-* + -SUBJPRES + -2S → *almuerces*

For parsing, we need to access the appropriate lexical entry, starting from the raw form. The lexicon stories the canonical lexical form (*almorz-a-*) as well as all alternative forms that the stem may take (*almorz-*, *almuerz-*, *almorc-*, *almuerc-*). These alternative forms are used during parsing to identify the lexical entry. Beyond this point, the orthographic changes become irrelevant for parsing. Generation, on the other hand, needs to consult a set of rules for how the canonical form may get rewritten as one of the three alternative. Since some of these processes are not completely regular and predictable, a complete list of verbs is maintained for each each of the change types. The verb *almorz-a-* appears on the lists for the *o* → *ue* and the *z* → *c* changes, and for each of the rules these is a separate list of conditions for the change. In this case the conditions for both changes are met, and the stem become *almuerc*. What remains for both parsing and generation is the relationship between the underlying sequence of inflections and the surface suffix. For this purpose there is a set of rules for each of the three Spanish conjugation classes that is used in both parsing and generation. One rule specifies the vowel for the subjunctive suffixes for this conjugation class:

(3) *-a-* verbs
-SUBJ_PRES ↔ *-e-*

Another rule specifies the person suffix:

(4) -2S ↔ *-s*

Syntax

The primitive units of the system are morphemes, rather than words; thus there are two levels at which elements need to be sequenced, within and between words. In what follows, "syntax" will refer to both levels (though the two are distinguished formally in the system).

Within computational linguistics, as in linguistics proper, there are many competing approaches to syntax. Again one of the main challenges is developing theories that lend themselves to both parsing and generation. Two additional considerations for the purposes of L3 are that the representations be intuitive and that it be possible to integrate the symbolic syntactic representations with what is learned statistically from corpora. **Dependency grammar** (Mel'cuk, 1988; Sugayama & Hudson, 2006;), which is much better known within computational linguistics (Debusmann et., 2004; Oflazer, 2003) than within linguistics proper, has advantages of both sorts since it dispenses with phrases and phrasal categories. In a dependency grammar, all syntactic information is encoded in **dependencies**, ordered relations between morphemes. In L3 both syntactic and morphological relations take the form of labeled dependencies that specify a preferred order and distance for the elements at its ends (for morphology, "distance" is in morphemes; for syntax, "distance" is in words). Each dependency takes as its source and destination either a morpheme or a morpheme category (such as noun or tense).

Sentence boundaries may also be dependency ends; this permits the encoding of the preferred position of particular words within the sentence.

Consider the morphemes in a Spanish verb: verb stem + tense/aspect/mood + person/number/gender. These relationships take the form of two dependencies:

(5) VERB ⇒ TAM [d = 1]

(6) VERB ⇒ PERS_NUM [d = 2]

Translation relations

Explicit bidirectional associations relate units across languages. At the most primitive level, morphemes are associated with morphemes, for example, es:*habl-a-* ↔ en:*speak*, or dependencies with dependencies, for example, es:*subj* ↔ en:*subj*.

Morpheme-to-morpheme and dependency-to-dependency translation represents a relatively abstract point along the transfer-interlingua continuum discussed in the section on translation above because it presupposes a relatively deep analysis of the source input and correspondingly complex process of generation for the target output. But there are often straightforward direct translations of unanalyzed, morphologically complex words or of entire phrases. Storing these associations as translation "shortcuts" can save processing time required for analysis of source words/phrases and generation of target words/phrases. Therefore, in L3 there are also bilingual lexicons that store translation associations for larger, unanalyzed units for either or both languages, for example, es:*"por supuesto"* ↔ en:*"of course"*.

Unlike in some approaches to machine translation, the cross-language associations specify nothing about sequencing of units within the target language sentence. This information is encoded in the dependencies within each language.

Processing

As in other translation systems, processing can be divided into phases that make use of within-language knowledge (analysis of the source, generation of the target) and phases that make use between-language knowledge. In L3, the assumption is that both kinds of knowledge are graded rather than categorical, and in each case a solution is found on the basis of the satisfaction of multiple constraints operating in parallel.

The basic steps in the translation of a sentence are as follows (also see Figure 1).

1. A source language sentence enters the system.

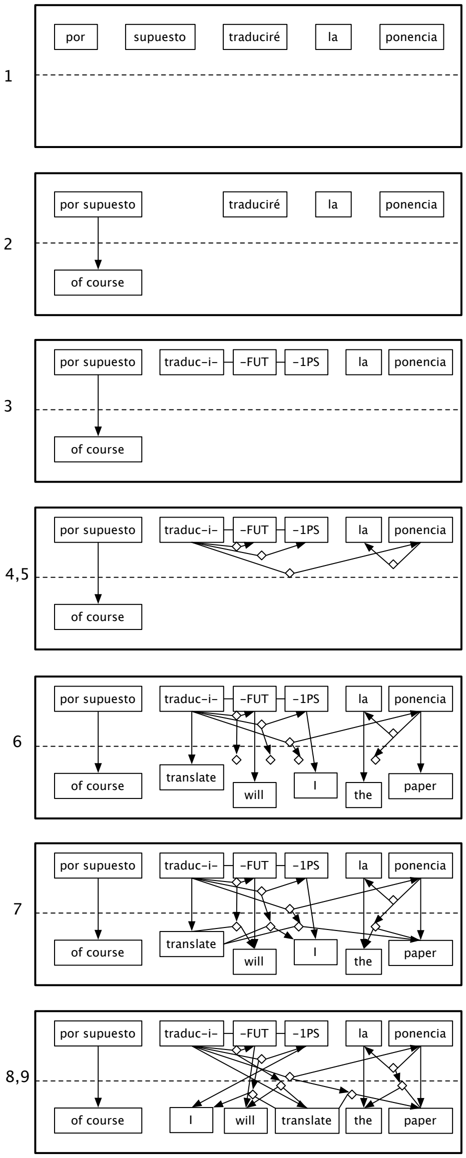

Figure 1: Steps in the translation of a sentence in L3 (see text)

2. A dictionary of phrase or unanalyzed translations is consulted. For each source language word or phrase that has a translation, the next step (morphological parsing) is bypassed, and the associated target-language units are added to the target "blackboard".

3. Source-language words are parsed into constituent morphemes.

4. Source-language morphemes are joined with all possible dependencies.

5. A constraint-satisfaction process prunes away dependencies that are not supported by the input, resulting in a complete parse of the input sentence.
6. The source-target association dictionary is consulted, and all possible target morphemes and dependencies are added to the target blackboard.
7. Target-language dependencies and morphemes are connected in all possible ways.
8. A constraint-satisfaction process prunes away morphemes and dependences that are not supported by the source input or the target-language grammar.
9. A constraint-satisfaction process orders the morphemes in the target sentence, using the information in the dependencies joining them.

Statistical model

L3 also runs in a separate mode in which it operates statistically. The only built-in grammatical knowledge for this part of the system is that required for morphological parsing or generation (the statistical mode takes as input and yields as its output strings of morphologically parsed words). There are no built-in dependencies, syntactic categories, or translation relations.

The statistical processor is presented with pairs of sentences that are translations of one another, and it learns to connect pairs of morphemes within languages by dependencies and pairs of morphemes or dependencies between languages by translation relations. Each dependency records statistics regarding relative position by keeping a histogram of the distances between the connected morphemes. Both dependencies and translation relations keep track of their frequency, and relations that occur relatively infrequently in comparison to the units they relate are deleted.

The statistical model has the advantage over the symbolic model that it stores "translation" relations for elements that tend to co-occur whether or not they represent actual translations of one another. When translation takes place, each candidate morpheme or dependency in a target sentence is potentially the result of the whole combined effect of the elements in the source sentence. This features represents a straightforward way to handle some ambiguity. Associations link not only morphemes that can be viewed as translation equivalents but also those that tend to occur within the same sentential contexts (albeit with weaker strengths). For example, the association es:*flor* ↔ en:*rose* (along with more direct associations like es:*flor* ↔ en:*flower* and es:*rosa* ↔ en:*rose*) could help to disambiguate the Spanish noun *escaramujo* 'wild rose; barnacle'. Thus if the word *escaramujo* occurs in a sentence such as *cogí la flor de un escaramujo*, the combined weight of the two links (es:*escaramujo* ↔ en:"*wild rose"* and es:*flor* ↔ en:*rose*) would lead the translation (*wild*) *rose* to be preferred over *barnacle*.

Although the statistical model starts with no syntactic categories, it can learn categories on the basis of similar syntagmatic behavior, that is, similarities in the dependencies that leave or enter particular morphemes. For example, since *me*, *te*, *le*, *lo*, *la, nos, los, las,* and *les* usually occur right before verbs and often occur right after subject pronouns, the system could learn to group them together into one class (though it would of course be confused by the ambiguity of *la*, *los*, and *las*).

Although the symbolic and statistical modes do not yet operate together, as they will need to eventually, the fact that the basic elements of each (morphemes, morpheme categories, dependencies, and translation relations between morphemes and dependencies) are the same will simplify integrating the two modes.

A test case: K'iche'-Spanish

Mayan languages

Roughly 30 (depending on how we count "dialects" and "languages") Mayan languages are spoken today by perhaps 5,000,000 people in Mexico, Guatemala, and Belize. Urban migration, war (in Guatemala), and reactionary educational policies (especially in Mexico) threaten the languages, though most are still used by people of all ages and in a wide variety of functions (see the Ethnologue data on the languages: http://www.ethnologue.com/show_family.asp?subid=90711). In Guatemala, the Accord on the Identity and Rights of Indigenous People, signed March 31, 1995, has increased the status of the languages in that country. Among other acts, the Accord began the process of "officializing" 21 Mayan languages of Guatemala (Spence et al., 1998), although this process has since stalled somewhat (Sieder et al., 2002). Within Guatemala, standardization of the languages is governed by the Academia de Lenguas Mayas de Guatemala (ALMG, http://www.almg.org.gt/).

K'iche'

The most spoken Mayan language is K'iche' (or Quiché) (Mondloch, 1978; OKMA, 2000), with 1-3 million speakers in the central highlands of Guatemala. The language has at least seven dialects, some quite divergent, and no official standard, though the Central dialect, with the most speakers, may be taking on the function of a written standard, if material appearing recently is any indication. The current K'iche' orthography, regulated by AMLG, has converged on its current form in the last 20 years, but there is still some variation in recently published material, particularly with respect to epenthetic vowels, which some writers leave out. Most K'iche' speakers are not literate in their native language, but this situation may be changing as the language is now taught in some schools in K'iche' areas.

In the past ten years, digital materials for teaching K'iche' (and several other Guatemalan Mayan languages, especially Achi', Ixil, and Tz'utujil) and for training teachers for bilingual schools have been developed in a project overseen by the Guatemalan non-governmental organization Asociación Ajb'atz' Enlace Quiché (http://www.enlacequiche.org.gt/), which has won several awards for its innovative work. Many of these materials are available on the Internet at the Portal de Educación Bilingüe Intercultural (http://www.ebiguatemala.org/). Also available there are digital versions (in PDF format) of short poems and stories written by students in the bilingual schools, as well as several dictionaries for K'iche' and other Mayan languages.

Because of this encouraging beginning with the use of technology in language education and the availability of some digital materials, it was felt that K'iche', and eventually other Mayan languages, would be an appropriate initial testbed for the ideas behind the L3 project. Our first goal is translation between Spanish and K'iche' within the limited domain of folktales appropriate for children.

Implementing Spanish and K'iche' verb morphology

Because of the complex morphology of K'iche' and other Mayan languages, it was felt that morphology would need to be taken relatively seriously early on. To date, most work in the K'iche'-Spanish project has been concerned with morphology, and for both languages, verbs present most of the problems.

Although software already exists for Spanish morphological parsing, the rules for Spanish were implemented from scratch so that they would be consistent with the general-purpose tools being developed for L3. As discussed above, handling Spanish verb morphology is a matter of somewhat separate sets of rules for the three conjugation classes. The main complexity stems from the many very stems that are subject to orthographic changes. All of the changes have been implemented. This requires 126 general rules, most of them bidirectional (that is, used for both parsing and generation), and hundreds of rules applying to specific verbs which are either irregular or need to be assigned to one or another orthographic change class. Nouns required only five general rules and adjectives only three. Most category-changing derivational morphology (*hablar* → *hablador*, etc.) was not implemented.

K'iche' verbs are much more complex than Spanish verbs because (1) they have both prefixes and suffixes, and (2) in addition tense-mood prefixes, there are are a range of voice and aspect suffixes, and (3) they code the person and number of both the subject and the object. As in other Mayan languages, there is a fundamental distinction between intransitive and transitive verbs, and separate lexicons were created for these two categories. Within the intransitive verbs, stative verbs represent a special subcategory, and within transitive verbs, there are two classes with some different voice and aspect suffixes, comparable to the conjugation classes of Spanish verbs.

Like most other Mayan languages, K'iche' is an ergative-absolutive language. The language has separate sets of absolutive and ergative prefixes. The absolutive prefixes code subjects on intransitive verbs and direct objects on transitive verbs. The ergative prefixes code subjects on transitive verbs and possession on nouns. A typical intransitive verb consists of a tense-mood prefix, an absolutive (subject) prefix, the verb stem, and optional aspect suffix, and an optional terminator suffix (used mainly in sentence- final position). A typical transitive verb in active voice includes a tense-mood prefix, an absolutive (object) prefix, an ergative (subject) prefix, the verb stem, and an optional terminator suffix. Transitive verbs also be intransitivized using one of three different voice suffixes, analogous to English passive (Mondloch, 1978).

Despite the overall complexity, there is little irregularity in the system, and K'iche' has nothing comparable to the orthographic changes the complicate Spanish verbs. However, the stative verbs have two different stem forms, used with different tense/aspect/moods, and for most verbs one of these is derived from the other through the reduplication of one stem vowel.

To implement K'iche' morphological rules, I relied mainly on the grammatical surveys of Mondloch (1978) and OKMA (2000). This was a confusing process since Mondloch's survey concerns a single dialect, and the OKMA book deals with dialectal variation. A native speaker or a linguist with more familiarity with the language could have gotten through this phase much more quickly. K'iche' intransitive verbs required 51 general rules, and transitive verbs required 85 rules. An example of a transitive verb that is correctly parsed, given the surface form, and correctly generated, given the parsed form, is *xujkitijoj* 'they taught us, nos enseñaron'.

(7) CMPL- + 1P_ABS- + 3P_ERG- + tijo + -TV_TERM ↔ *xujkitijoj* (*x-uj-ki-tijo-j*)

To handle the non-concatenative processes involved in generating the reduplicated stative forms, the morphological rule framework needed to be augmented so that it could handle variables. (Since reduplication and infixation will be aspects of languages that L3 will address in the future, this would have been necessary in any case.) An example of a stative verb form that the program can generate is *chixt'uyunoq* 'be seated (pl.)!, ¡siéntense!'. In this imperative verb the stem *t'uyun* is derived from the underlying stem *t'uyi'* through the reduplication of the first vowel and replacement of the ' with *n*.

(8) IMPV- + 2P_ABS- + t'uyi'+ -IMPV_TERM ↔ *chixt'uyunoq* (*ch-ix-t'uyun-oq*)

BUILDING SPANISH AND K'ICHE' LEXICONS

Separate lexicons of nouns, verbs, and adjectives were created for Spanish, mainly using the lists that are part of the Spanish EuroWordNet (Vossen, 1998).

For K'iche' I relied on two bilingual dictionaries that are available online in PDF format: Christenson's unpublished K'iche' dictionary (1993) and the K'iche'-Spanish, Spanish-K'iche' dictionary compiled by ALMG (2004). Both are incomplete and somewhat inconsistent in their grammatical notation. Classifying transitive verbs into one of the two basic categories is essential for morphological parsing and generation, and it was not always possible to identify the category from the form provided in the dictionaries (though Christenson sometimes indicates the category). To clear up the confusion, I had to rely on the much more consistent and reliable glossary at the end of Mondloch’s (1978) grammar, which is unfortunately not available in digital form.

Separate K’iche’ lexicons of nouns, adjectives, intransitive verbs, transitive verbs, and miscellaneous roots were compiled, about 5000 forms in all.

For Spanish-K’iche’ translation, the ALMG (2004) bidirectional dictionary has been the only basis for translation associations so far. This dictionary is very incomplete, however, and will need to be supplemented with other small glossaries that are available and perhaps with Christenson’s K’iche’-English dictionary.

An interesting problem has been the translation of personal pronouns and subject or object agreement inflections on verbs. Because K’iche’ is an ergative-absolutive language and Spanish an nominative-accusative language, there is no one-to-one to correspondence between morphemes such as *at* ‘you (sing.)’ and *tú*. The K’iche’ pronoun and absolutive verb prefix *at* translates as Spanish *tú* or second person subject agreement inflection only when the verb in the sentence is intransitive. With transitive verbs, *at* corresponds to the Spanish object, *te*. Our solution has been dependencies for transitive verbs that relate pairs absolutive and ergative morphemes in K’iche’ to the corresponding pairs of nominative and accusative morphemes in Spanish. For example, the following relates a first person plural absolutive and third person plural ergative in a K’iche’ sentence (as in (7) above) to a first person plural accusative and third person plural nominative in the corresponding Spanish sentence:

(9) (k’i:3P_ERG ⇒ 1P_ABS) ↔ (es:3P_NOM ⇒ 1P_ACC)

Intransitive sentences (and nouns, which use the ergative prefixes for possession in K’iche’) would be handled separately by morpheme-to-morpheme associations, which would apply when a rule such as (9) did not.

Current status

Most progress has been made in the area of morphology in both languages, which called for significantly more work than expected. The system successfully parses most nouns, verbs, and adjectives in both Spanish and K’iche’, including most inflected forms and many derived forms of thousands of lexical roots in each language.

Sentence-level parsing, however, is still very primitive and will remain a significant challenge, given the complexity of the parsing task, which has in fact been the primary focus of computational linguistics since its inception. In part this can be a matter of incorporating more and more relatively specific dependencies, but accomplishing this will probably not be feasible, given the effort required. Instead the hope is that parsing performance can be improved as the statistical model is exposed to corpora.

Translation is also very primitive at this point. Only the first translation for each morpheme is considered, and for two languages as different as Spanish and K'iche', this is clearly inadequate.

Despite these somewhat disappointing results, the project represents the first-ever computational linguistic work in a Mayan language, and what has been achieved so far already has potential applications in an educational context. We are currently considering ways to use the morphological parser/generator in courses for students who are training to teach in bilingual K'iche'-Spanish schools in Guatemala.

Next steps

K'iche'-Spanish

The K'iche'-Spanish project clearly has a long way to go before it is translating sentences of any complexity. The major goals of the next phase in this project will be to:

1. increase the coverage of Spanish and K'iche' syntax so that more than rudimentary syntactic (as well as morphological) parsing is possible
2. develop mechanisms to handle within-language and between-language ambiguity
3. develop a discourse component for the system (implementing between-sentence dependencies and a memory for discourse referents)
4. integrate the symbolic and statistical modes within the system so that the built-in grammatical and lexical knowledge can be augmented on the basis of statistics in corpora (as these become available)
5. develop a user-friendly interface for bilingual users to interact with the system and with each other on issues related to K'iche'-Spanish translation
6. involve native speakers in as many aspects of the development as possible.

A key question during this phase will be the person-hours required to implement all of these components. If the larger L3 project is to succeed, the effort expended to bring a new language on board must be something that can be accomplished in months rather than years. As yet we do not know whether this will be possible.

Beyond K'iche' in Mayan

We are interested in K'iche' in part because of its close relationship to other languages in the Quichean branch of Mayan such as Kaqchikel and its more distant relationship to other Mayan languages such as Tzeltal and Yukateko. We would like to be able to use the work on K'iche' to facilitate projects on these other languages. There are three ways in which we hope to transfer knowledge of one language to related languages.

1. Orthographically (phonologically) similar words in two closely related languages are likely to have the same translation in Spanish (or another distantly related language). Thus with little knowledge of Tz'utujil, a K'iche' speaker can readily recognize the (written) form *chi nuuwach* 'in front of me' as the Tz'utujil corresponding to K'iche' *chnuwach*. But "similarity" is a complex concept, and in this case it has to depend on what is known about sound correspondences among the languages. A very useful tool in this regard will be the Mayan etymological dictionary of Kaufman (2003).

2. The grammars of the languages will coincide in many ways that can simplify translation between them or permit using knowledge of one to translate to or from another. For example, most Mayan languages are ergative-absolutive, and it is therefore possible to translate person-number categories between them in a straightforward manner. Thus Tzeltal, a language only distantly related to K'iche', has its own set of ergative prefixes and absolutive *suffixes*, and most do not resemble the corresponding K'iche' morphemes, but their functions are very similar. In a larger Mayan translation system, the person-number categories will become Mayan-specific categories.

3. The lexicons of the languages will also coincide in many ways. For example, Mayan languages are well-known for their relatively specific verbs, verbs of eating for particular kinds of food, verbs of carrying for particular kinds of burdens, etc. Thus both K'iche' and Tzeltal have verbs meaning 'eat tortillas'. As with grammar, it will be useful to exploit these similarities to build a Mayan-specific set of "concepts" where this is possible. In fact the statistical version of the program could be seen as a way of discovering these lexical/conceptual commonalities, supplementing the cross-linguistic work of people like Kaufman.

References

Academia de Lenguas Mayas de Guatemala, Dirección Lingüística y Cultural. (2004). *K'iche' choltzij. Vocabulario K'iche'*. Guatemala City: ALMG.

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., and Roossin, P.S. (1990). A statistical approach to machine translation. *Computational Linguistics*, *16*, 79-85.

Bryant, S, Forte, A and Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of GROUP International Conference on Supporting Group Work*.

Chinn, M.D. and Fairlie, R.W. (2004). The determinants of the Global Digital Divide: a cross-country analysis of computer and Internet penetration. Center for Global, International and Regional Studies, University of California, Santa Cruz, California, USA. Available at the California Digital Libary eScholarship Repository: http://repositories.cdlib.org/cgirs/CGIRS-2004-3/

Christenson, A.J. (1993?). *K'iche'-English dictionary*. Foundation for the Advancement of Mesoamerican Studies. http://www.famsi.org/mayawriting/dictionary/christenson/quidic_complete.pdf

David, P.A. and Foray, D. (2003). Economy fundamentals of the knowledge society. *Policy futures in education – an e-journal*, *1*(1). http://www.wwwords.co.uk/pfie/content/pdfs/1/issue1_1.asp

Debusmann, R., Duchier, D. , Koller, A., Kuhlmann, M., Smolka, G., and Thater, S. (2004). A relational syntax-semantics interface based on dependency grammar. *International Conference on Computational Linguistics*, 20*,* Geneva.

Fisher, D.R. and Wright, L.M. (2001). On utopias and dystopias: toward an understanding of the discourse surrounding the Internet. *Journal of Computer-Mediated Communication*, *6* (3). http://jcmc.indiana.edu/vol6/issue2/fisher.html

FUNREDES/Unión Latina. (2006). En quelles langues parle Internet? Observatoire de la diversité linguistique et culturelle dans l'Internet. http://funredes.org/lc/documentos/dep_lenguas_fr_fevrier2006.pdf

International Telecommunication Union. (2003). World Summit on the Information Society: Declaration of Principles. Geneva, Switzerland: ITU. http://www.itu.int/wsis/docs/geneva/official/dop.html

Kay, M. (1997). The proper place of men and machines in language translation. *Machine Translation*, *12*, 3-23.

Kaufman, T. (2003). *A preliminary Mayan dictionary*. Foundation for the Advancement of Mesoamerican Studies. http://www.famsi.org/reports/01051/pmed.pdf#search=%22kaufman%20mayan%20dictionary%22

Kolbitsch J. and Maurer H. (2006). The transformation of the Web: how emerging communities shape the information we consume. *Journal of Universal Computer Science, 12*, 187-213.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki.

Leuf, B, and Cunningham, W. (2001). *The Wiki way: quick collaboration on the Web*. Boston, Massachusetts, USA: Addison-Wesley Longman Publishing.

Mel'cuk, I. (1988). *Dependency syntax: theory and practice.* Albany, New York, USA: State University of New York Press.

Mondloch, J.L. (1978). *Basic Quiche grammar*. Institute for Mesoamerican Studies, University at Albany, The State University of New York, United States, publication no. 2.

Nirenburg, S., Carbonell, J., and Tomita, M. (1994). *Machine translation: a knowledge-based approach.* San Francisco, California, USA: Morgan Kaufmann.

Oflazer, K. (2003). Dependency parsing with an extended finite-state approach. *Computational Linguistics*, *29*, 515-544.

O'Neill, E.T., Lavoie, B.F., and Bennett, B. (2003). Trends in the evolution of the public Web: 1998-2002. *D-Lib Magazine*, *9*(4). http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

Oxlajuuj Keej Maya' Ajtz'iib' (OKMA). (2000). *Ujunamaxiik ri K'ichee' ch'ab'al, Variación dialectal en K'ichee'.* Guatemala City: Cholsamaj.

Paolillo, J. (2005). Language diversity on the Internet: examining linguistic bias. In UNESCO Institute for Statistics (Ed.), *Measuring linguistic diversity on the Internet.* Montreal, Canada: UIS. http://www.uis.unesco.org/template/pdf/cscl/MeasuringLinguisticDiversity_En.pdf

Pimienta, D. (2005). Linguistic diversity in cyberspace; models for development and measurement. In UNESCO Institute for Statistics (Ed.), *Measuring linguistic diversity on the Internet.* Montreal, Canada: UIS. http://www.uis.unesco.org/template/pdf/cscl/MeasuringLinguisticDiversity_En.pdf

Rheingold, H. (2001). *The virtual community: homesteading on the electronic frontier* (revised edition). Cambridge, Massachusetts, USA: MIT Press. http://www.rheingold.com/vc/book/

Sieder, R., Thomas, M., Vickers, G., and Spence, J. (2002). Who governs? Guatemala five years after the Peace Accords. Cambridge, Massachusetts, USA: Hemisphere Initiatives.
http://www.hemisphereinitiatives.org/whogoverns.pdf

Spence, J., Dye, D.R., Worby, P., de Leon-Escribano, C. R., Vickers, G., and Lanchin, M. (1998). Promise and reality: implementation of the Guatemalan Peace Accords. Cambridge, Massachusetts, USA: Hemisphere Initiatives. http://www.hemisphereinitiatives.org/promise.htm

Sugayama, K. and Hudson, R.A. (Eds.) (2006). *Word Grammar: new perspectives on a theory of language structure*. London: Continuum International.

Teigland, R, (2003). Knowledge networking: structure and performance in networks of practice. Stockholm: Stockholm School of Economics.
http://www.hhs.se/NR/rdonlyres/4165BDC8-C42C-43CF-8EEF-57DCEB0939BC/0/TeiglandthesisKnowledgeNetworking.pdf

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, *32*, 73-89.