THE IMPACT OF WRITTEN TELECOMMUNICATIONS TECHNOLOGY ON THE WORLD'S LINGUISTIC DIVERSITY

by Gabriela Pérez Báez

(Editor's Note: Gabriela Perez-Baez is Curator of Linguistics in the Anthropology Department, National Museum of Natural History, Smithsonian Institution, and a member of the Recovering Voices Program. Here she highlights the impact and challenges of written telecommunications on the world's linguistic diversity through a case study of one indigenous language from Mexico.)

Introduction

Of the world's several thousand distinct languages, only a small fraction are used in written telecommunication. For example, even in content that contains different language translations, Wikipedia articles are written in only about two dozen languages. Despite long-held goals of software localization, developers' efforts only cover a few dozen languages. For instance, Apple's Mac OS X offers user preference settings in only 18 languages, 14 of which are European. Despite the challenges, there is growing interest in opening written telecommunications media for larger numbers of the world's languages, 90% of which see their long-term survival at serious risk.

Linguistic Diversity and Language Endangerment

Humans speak somewhere between 4,500 and 7,000 languages. Thirty percent of these languages are spoken in Africa; another 35% are spoken in Australasia and the islands of SE Asia, and some 1,200 languages in New Guinea alone. The Americas are home to approximately 1,000 languages, including about 200 Native American languages in the United States. In contrast, the over 30 languages of

Europe amount to less than 1% of the world's languages, yet many of them are widely represented in written telecommunications. Only about 80 languages are spoken by as much as 80% of the world's population, while 20% of the population is distributed across the remaining thousands of languages. Languages such as English, Spanish, or Mandarin Chinese are spoken by hundreds of millions of people each, while other language communities consist of thousands, hundreds, and sometimes only a few speakers.

Every single language embodies knowledge about human adaptation to natural and social environments, as well as insights into each culture's development through time. Every language contributes to our understanding of the human language faculty and our brain's inner workings. However, as many as 90% of them risk extinction as spoken languages by the end of this century. For many social and political reasons, languages become endangered as their domains of use – the social spaces such as the family, the community, the worksphere, and education, in which a language is spoken – become fewer and more restricted, with dominant languages taking over those domains.

The fact that written telecommunications are not available to the majority of languages excludes most languages from yet another domain of use. In other words, written communications have a linguistically homogenizing force that contributes to language endangerment. As awareness of this issue increases among language communities, technology developers are responding. To develop localized versions of software, especially for data input and display, requires an in-depth understanding of the linguistic, social, and cultural systems of the community to be served. The following case study of one of the largest indigenous languages of Mexico illustrates some of the necessary linguistic considerations needed in order to open the domain of written telecommunications to a new language community.

Map of Mexico

Meet a Zapotec language

"Zapotec" is not a single language form, but rather a family of languages comparable to the family of Romance languages. This family exhibits significant common traits but with enough differences to make it difficult if not impossible for speakers of different Zapotec languages to understand each other. Zapotec languages are pre-Columbian languages whose ancestral speakers can be traced back to over 2,000 years to the ancient Zapotec civilization. The actual number of distinct Zapotec languages is a matter of ongoing debate among linguists given the complexity of this language family but the number may be in the dozens.

The complexity of the family of Zapotec languages comes in great part from the complexity of their phonology – that is their sounds systems. Let us consider the way in which vowels are pronounced. Zapotec languages can have anywhere between four and six vowels and interestingly each vowel can have a different phonation type – that is, a different way of articulating the vowel. For instance, Juchitán Zapotec spoken in the eastern end of Oaxaca in the Isthmus of Tehuantepec, has five vowels and each can be pronounced in one of three different ways. A modal

vowel is a regular vowel in which the vocal chords vibrate unimpeded as in [e] in 'wet'. A rearticulated vowel is one where the glottis closes to interrupt the flow of air and the sonority of the vowel and then releases to allow both to continue. The English expression 'oh, oh!' is close to what rearticulated vowels in Juchitán Zapotec sound like, where the [o] is interrupted and then allowed to continue. A checked or glottalized vowel is one in which the vocal chords close up but do not release, ending the sonority of the vowel sharply. The phonation differences determine the meanings of words; in linguistic terms, these differences are contrastive. For instance, in Juchitán Zapotec, the word *gye* means 'stone' and features a modal [e] while *gye'* where the vowel is glottalized and marked by an apostrophe, means 'flower'. Similarly *gi* with a modal vowel means 'fire' while *gi'* with a glottalized vowel means 'excrement'. The examples in (1) show contrastive vowel phonation:

(1) gela 'horseshoe' ge'la 'depth' geela 'night'

Zapotec languages are also tonal, meaning that a change in the pitch of a vowel can be contrastive. In Juchitán Zapotec there are three tones. There is a contour rising tone where the pitch changes throughout the articulation of a vowel going from low to high. Then, there are two register tones, a high and a low tone in which the pitch of the vowel stays high in one case, low in the other, throughout the articulation of the vowel. In the word *nanda* 'cold', both vowels are low tone while in *nanda** 'hanging' the second vowel has a rising tone marked by an asterisk. Tone and vowel phonation interact with yet another feature that is prosodic stress – a system where one syllable is given more prominence than all others in a word. The interaction of vowel phonation, tone, and stress creates a complex system of contrast between words. Let us compare the last two examples with two more words that have the same consonants and vowels but have different tone, vowel phonation, and stress properties. The stressed syllable is marked in bold in the last column in Table 1.

Gabriela Pérez Báez's collaborators Rosaura López Cartas and Reyna Guadalupe López López from La Ventosa, Juchitán, Oaxaca.

Table 1. Phonological features in Juchitán Zapotec

Juchitán Zapotec	English	Word-final Vowel Phonation	Word-final Vowel Tone	Stress
nanda	'cold'	modal	low	***na**nda*
*nanda**	'hanging'	modal	rising	***na**nda**
na-nda'*	'sour'	glottalized	rising	*na-**nda*'***
na-nda!'	'hot'	glottalized	high	*na-**nda!'***

The challenge is to represent orthographically the various phonological features of words and to convey, in writing, all of the information that a reader of Juchitán Zapotec may need to understand a written message unambiguously. In the words on Table 1, vowel phonation, tone, and stress are represented orthographically so anyone who knows how to read this orthography will know exactly how to pronounce each word to convey its meaning. Low tone is not marked because it is the most frequent; any vowel without a tone mark is then a low tone vowel. As mentioned earlier, an asterisk marks a rising tone; the exclamation mark indicates a high tone. Modal vowels do not take any special marking, but glottalized vowels take an apostrophe and rearticulated vowels as in *geela* 'night' are marked with a double vowel in analogy to the articulation, glottal constriction, and rearticulation of the vowel. Stress is generally in the first syllable of the written word but in some cases it is not. A hyphen is then used to precede the stressed syllable as in *na-nda** 'sour' and *na-nda!'* 'hot'.

Telecommunications and Linguistic Diversity

What would it be like for speakers of Juchitán Zapotec to type or text in their language using available interfaces? It is not easy even though, for instance, the orthography is designed to avoid the use of diacritics – symbols above or below a letter to indicate a change in pronunciation – that can be difficult to type in most PC-based computer keyboards. That is why tone is marked with a * and a ! which are readily available in any computer and even in typewriters.

So Juchitán Zapotec can be easily typewritten on any modern computer keyboard. However, texting is a very different matter. Even in sophisticated swiping systems such as those in Android phones, a message in Juchitán Zapotec requires toggling back and forth between the letter keyboard and the symbols keyboard. With possibly every word in a message requiring a symbol, toggling back and forth between keyboards could get annoying very fast. More modest cell phones may require multiple pulsations over a single key to get one character, an inconvenience that is minimized in certain phone models by predictive text – spelling suggestions made by the operating system. However, these systems are designed to assist the user with dominant languages such as English or Spanish, but no such system exists in Juchitán Zapotec. Clearly, technology can be relevant to issues of language endangerment.

Juchitán Zapotec is the most widely spoken Zapotec language with some 70,000 speakers and possibly more. Cell phones are ubiquitous even in the most modest communities. Texting by young people is quite common, but *alas!,* it is done *in Spanish,* in part because all the necessary exclamation marks, asterisks, and apostrophes are a hassle to type, yet are essential to convey meaning unambiguously. Also, and perhaps due to insufficient use of the Juchitán Zapotec for texting, no txtspk – whether for texting or for chatting — has developed in the language. Even though Juchitán Zapotec can be easily typed on any computer keyboard, the interface of most computers and of software, including the internet browsers used in Mexico, is in Spanish. This motivates the use of the dominant language (Spanish) over any other. For example, the internet site of *Zapotecos del Mundo*, an online Zapotec community mostly of Zapotecs from the Juchitán region, is very active with several postings a day, but these are overwhelmingly in Spanish. The domain of written telecommunications in Mexico is clearly Spanish and not yet available to Juchitán Zapotec nor to most other indigenous languages spoken in Mexico.

Technology and Its Role in Reversing Language Endangerment Trends

Languages become endangered as their domains of use shrink and become fewer in number. Numerous factors, interrelated in complex ways, lead speakers to use a dominant language such as Spanish over a language such as Juchitan Zapotec in the workplace, the school, public life, and so forth. As technology becomes more widely available to larger numbers of users, it becomes a new domain

Gabriela Pérez Báez at work with her collaborator Rosaura López Cartas with whom she has worked on a dictionary of Juchitán Zapotec for close to ten years.

of language use, yet one that at the moment is accessible only through a very reduced number of languages.

Significant progress has been made to design interfaces and keyboards around the particulars of languages such as Spanish, French, or Italian that require diacritics, and Chinese, Russian, Japanese, and Arabic that use non-Roman characters. Quite promising developments have opened written telecommunications to languages such as Hawaiian, Maori, Inuktitut, and more recently, Cherokee, for which an iPhone interface now exists.

Much more work is needed. Efforts are underway, for instance, at Mozilla México, to localize its browser for Mayan and for Southern Sierra Zapotec users. Similarly, Wikimedia México, a nonprofit educational organization affiliated with the Wikimedia Foundation, the parent of Wikipedia, has an emerging project to develop content in indigenous languages of Mexico. These new developments represent the opportunity to turn written telecommunications into a domain of use for more indigenous endangered languages around the world, thereby actively participating in the maintenance of the world's linguistic diversity.

Resources

UNESCO Atlas of the World's Languages in Danger http://www.unesco.org/culture/languages-atlas/index.php

Archive of the Indigenous Languages of Latin America Ethnologue www.ethnologue.com

Hans Rausing Endangered Language Program www.hrelp.org

Instituto Nacional de Lenguas Indígenas (INALI) www.inali.gob.mx

World's Atlas of Linguistic Structures www.wals.info/

Zapotecos del Mundo

http://zapotecosdelmundo.ning.com/

Gabriela Pérez Báez is a linguist in the Department of Anthropology, Smithsonian Instituiton.

MEET LINGUIST GABRIELA PÉREZ BÁEZ

Watch a Smithsonian podcast interview where Dr. Pérez Báez talks about her education, Smithsonian career, and reearch at http://anthropology.si.edu/video_interviews.html