Developing linguistic resources for the nahuatl indigenous language

Carmen C. Martínez-Gil[1], Alejandro Zempoalteca-Pérez[1],
Venustiano Soancatl-Aguilar[2], Rosa M. Ortega-Mendoza[3],
[1] Universidad de la Cañada, [2]Universidad del Istmo, [3] Instituto
Tecnológico Superior del Oriente del Estado de Hidalgo,
{cmartinez, alejandro}@naxoloxa.unca.edu.mx,venus@bianni.unistmo.edu.mx
rosy_om@hotmail.com

(Paper received on September 01, 2010, accepted on October 20, 2010)

Abstract. In this paper we present the development of linguistic resources with the purpose of studying and analyzing the Nahuatl indigenous language. Nahuatl is currently spoken in countries such as: Mexico, El Salvador, United States, Guatemala and Nicaragua. The linguistic resources developed are: a Nahuatl-Spanish corpus of unstructured text and dictionaries. The dictionaries are three, the first having terms in Nahuatl and Spanish, the second containing words in Nahuatl which are translated into Spanish phrases, and the latter having specific Nahuatl terms from the northern region of Oaxaca, Mexico

Keywords: Corpus, Linguistic resources, Nahuatl indigenous language.

1 Introduction

Nahuatl is an indigenous language currently spoken in many countries such as Mexico, El Salvador, United States, Guatemala and Nicaragua. Nahuatl is one of the most studied and documented American languages [1, 4, 6, 7, 10, 11, 14]. Nowadays we can still find manuscripts and documents written in Nahuatl and we can extract useful, valuable and important information from those documents which benefits current and future generations.

Our interest in developing linguistic resources for the Nahuatl language is mainly focused on: a) Due to ignorance and mismanagement of terms of this language, we are losing much of our culture. It is therefore important to recognize and extract the information contained in documents written in Nahuatl, b) To preserve and disseminate a language with historical roots, in order to keep the language alive, because if we lose it, we lose our essence and identity as Mexicans; c) To comprehend and understand cultures that use this language including speakers whose first and only language is Nahuatl.

On the other hand, the research area of Natural Language Processing (NLP) is a sub-discipline of Computer Science and Linguistics [2, 3], which is responsible for producing computer systems that enable man-man or man-machine communication using natural language. The aim of NLP is to study the problems derived from automatic generation and understanding of natural language. Some relevant applications of NLP are: Automatic translation, Speech recognition, speech synthesis, information extraction, information retrieval, automatic summarization, handwriting recognition, text mining, question answering, text classification [8], and Automatic Identification of Languages [9]. To strengthen these NLP applications, NLP uses linguistic resources. Linguistic resources are sets of language data in computer readable form and are used in the construction, improvement and evaluation of natural language systems, although the term also includes software tools aimed for separating, collecting, and managing other resources. Language resources are classified into three categories: body, tools and lexical resources. A corpus [12] is a collection of natural language texts, chosen to characterize a state or variety of a language. There are corpuses of written language and spoken language, in both cases the body acts as a repository of information which can be manipulated to extract knowledge. The tools help to analyze the texts, among the most common tools are: parts of speech labelers, morphological analyzers and parsers. The lexical resources [5] provide a set of valid words in the language, and may also contain language properties, the meaning of words and relationships between words or group of words. Some examples of lexical items are: word lists, dictionaries, thesauri, anthologies, glossaries, and others.

However, the texts that form the corpus can be classified [13] broadly into three categories: structured text, semi-structured text and unstructured text. Structured text is stored in a rigorous format that allows differentiation of relevant parts of the text. Semi-structured text contains some kind of structure, but not enough to be considered as structured text. And unstructured text that does not have a definite structure, it is free text format and extracting information from these texts is not easy and usually includes a pre-processing stage.

In the following section we describe the process to develop linguistic resources, the corpus and dictionaries. Then, in the section 3 we show the significant results obtained to the moment. Finally in the section 4 and 5, the conclusions and future word are present.

2 Developing Linguistic Resources

With the purpose of developing tools and applications for the Nahuatl language, we are currently developing the following linguistic resources:

a) a corpus of Nahuatl-Spanish texts
b) some lexical resources that consist of three Nahuatl-Spanish dictionaries

While it is true that nahuatl is one of the America's most studied and documented languages, it is also true that much of this documentation is in electronic PDF or hard copies. So, to build our corpus we need unstructured texts in a txt format in order to manipulate its contents to extract knowledge and in a later stage structure the texts according to our requirements. On the other hand, because there are different variants of Nahuatl, we have focused on Nahuatl from Northern Oaxaca mainly for three reasons:

a) Our geographic location
b) The viability of documents in this region
c) The participation of students whose native language is Nahuatl and their ability to collaborate on the development of the project

At the moment we are working with this variant of Nahuatl, although we have considered extending our linguistic resources to analyze any kind of text in Nahuatl. To form our corpus of texts, we have obtained Nahuatl-Spanish texts from various sources [4, 6, 7, 10, 11].

Regarding lexical resources we have developed a dictionary of terms in Nahuatl and its translation into Spanish. Because the Nahuatl is an agglutinative language, ie, adds prefixes and / or suffixes to roots to form very long words. The first version of the dictionary is divided into two parts, the first one containing terms Nahuatl-Spanish and the second containing Nahuatl words whose meanings are sentences. Finally we are working on a dictionary of terms and Nahuatl words from the northern state of Oaxaca. This latter resource will help us to analyze the native texts in the region.

3 Results

The corpus that we are building is made up of texts in Nahuatl and their correspondent text in Spanish, so we have two files in txt format for each text. The texts are classified into four categories: Poetry, Stories, Religion, and Miscellaneous. Table 1 shows the number of texts that form the corpus.

Table 1. Number of texts by category.

Category	Náhuatl	Spanish
Poetry	125	125
Stories	68	68
Religion	550	550
Miscellaneous	93	93

Poetry texts were primarily obtained from the poems written by Nezahualcoyotl, Nahuatl stories were written in the state of Guerrero (state of Mexico) and translated into Northern Oaxaca Nahuatl, religious texts were obtained from the Bible and some others from prayers, finally the miscellaneous contains several texts such as legends, tales, stories, thoughts, wills, among others. Table 2 shows on average how many text words are contained in each category.

Table 2. Word average contained in each text by category.

Category	Average
Poetry	143
Stories	345
Religion	258
Miscellaneous	933

Three dictionaries have been built, one of Nahuatl words and their correspondent word in Spanish, the second provides he words in Nahuatl and their translation in sentence form and finally the third dictionary contains specific Nahuatl words from the upstate region of Oaxaca, Mexico. Table 3 shows the number of words for each dictionary.

Table 3. Number of words for each dictionary.

Dictionary	Number of words
Nahuatl-Spanish word-word.	9765
Nahuatl-Spanish word-phrase.	3288
Nahuatl words from the northern region of Oaxaca, Mexico.	560

4 Conclusions

In this paper we have introduced the development of language resources for studying and analyzing the native Nahuatl language with great historical, linguistic, literary and nationalist significance.

We have presented the development of a corpus of unstructured texts in Nahuatl and their correspondent text in Spanish, as well as the construction of three dictionaries: one with Nahuatl-Spanish terms, the other with Nahuatl words whose translation into Spanish is a phrase, and a final dictionary of Nahuatl specific terms from the northern region of Oaxaca, Mexico.

The results obtained so far are very significant because we have a high quality corpus that is, varied and large enough for getting significant results. The dictionaries contain information representative for the Nahuatl language and are useful for future applications. For the moment to obtain and use the resources you can contact the authors, soon we open a URL to obtain them.

5 Future Work

As immediate future work we will add semantic information to the terms of the dictionary, we will implement a system for suffixes or prefixes of a word in Nahuatl and we will continue increasing the amount of texts of the corpus and the terms of the dictionaries. As long-term future work we will consider building a Nahuatl speech tagger of parts of sentences as well as a stemmer.

References

1. Andrews, J. R. 2003. Introduction to Classical Nahuatl. University of Oklahoma Press.
2. Brill, E., Mooney, R. J. 1997. An Overview of Empirical Natural Language Processing. Artificial Intelligence. Vol. 18. No. 4.
3. Bolshakou, I., Gelbukh, A. 2004. Computational Linguistics. Computer Science. México.
4. Bible League, The. 2006. The New Testament. Northern Oaxaca Nahuatl. Mexico. (Spanish).
5. Gellerstam, M. 1995. Lexical resources and their application. Proceedings of the 1st Trans-European Language Resources Infrastructure (TELRI) Seminar on anguage Resources for Language Technology. Tihany, Hungary. pp 57-64.
6. Launey, M. Introduction to the Nahuatl language and literature. Mexico D.F.: UNAM. 1992. (Spanish).
7. Montemayor, C. 2001. Current literature in indigenous languages of Mexico. Mexico City: Universidad Iberoamericana. (Spanish).
8. Ramírez-de-la-Rosa G., Montes-y-Gómez M., Villaseñor Pineda L., Pinto Avendaño D., Solorio T. Using Information from the target Language to Improve Crosslingual Text Classification. IceTAL 2010: 305-313.
9. Reyes-Herrera A.L., Villaseñor Pineda L., Montes-y-Gómez M. A Straightforward Method for Automatic Identification of Marginalized Languages. FinTAL 2006: 68-75.
10. Róman, R. (Compiler). 2007. Anthology of Native stories Guerrero. National Council for Culture and the Arts. General Directorate of Popular Culture. Mexico. (Spanish).
11. Siméon, R. 2001. Dictionary of Nahuatl or Mexican language. [Paris 1885] Reprint: Mexico.
12. Sinclair, J. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.
13. Tustison, C.A. 2004. Logical form identification for medical clinical trials, Master's thesis. Department of Linguistics and English Language. Brigham Young University.
14. Wolgemuth, C. 2002. Nahuatl Grammar. México. (Spanish).