# Computer Systems for Analysis of Nahuatl

Carmen C. Martínez-Gil[1], Alejandro Zempoalteca-Pérez[1], Venustiano Soancatl-Aguilar[2], María de Jesús Estudillo-Ayala[3], José Edgar Lara-Ramírez[3], and Sayde Alcántara-Santiago[4]

[1]Universidad de la Cañada, Carr. Teotitlán-San Antonio Nanahuatipan Km. 1.7 s/n, Paraje Titlacuatitla, Teotitlán de Flores Magón, Oax., C.P. 68540, Mexico

[2]Universidad del Istmo, Carr. Chihuitan Ixtepec s/n, Ixtepec, Oax., C.P. 70110, Mexico

[3]Escuela de Ciencias, Universidad Autónoma Benito Juárez de Oaxaca. Av. Universidad s/n, Ex-Hacienda de 5 Señores, Oax., C.P. 68120, Mexico

[4]NovaUniversitas, Carretera a Puerto Ángel Km. 34.5, Ocotlán de Morelos, Oax., C.P. 71513, Mexico

{cmartinez, alejandro}@unca.edu.mx, venus@bianni.unistmo.edu.mx, salcantara@jacinto.novauniversitas.edu.mx

**Abstract.** This article describes two computer systems that allow us to analyze words written in the Nahuatl language. The main goal is the diffusion and preservation of indigenous language with great historical, linguistic, literary and nationalistic relevance by developing language resources for Nahuatl. One system automatically gets prefixes or suffixes of words from a text written in Nahuatl. This system was developed because Nahuatl writing contains agglutination, i.e. prefixes and/or suffixes are added to the root of a word to give it specific meaning. The other system is a Nahuatl to Spanish translator and vice versa, which also shows semantic information related to the terms in Nahuatl. This information includes the root, or roots of words as well as it´s grammatical category, which can be: a noun, adjective, pronoun, preposition, conjunction, article, adverb, verb or interjection. The system currently contains 1,514 terms.

**Keywords:** Computer systems, Nahuatl language, language resources, semantic information.

## 1 Introduction

Nahuatl is an indigenous language which is currently spoken in countries such as Mexico, El Salvador, United States, Guatemala and Nicaragua.

Currently, Nahuatl is the indigenous language most widely spoken in the Mexican territory, with approximately a million and a half fluent individuals, as reported by National Institute of Statistics and Geography and Computer. This language is valuable because it has great historical, linguistic, literary and nationalistic

significance. The states in Mexico where Nahuatl is still spoken include: State of Mexico, Puebla, Guerrero, Hidalgo, Veracruz, Oaxaca, Durango, Morelos, Mexico City, Tlaxcala, San Luis Potosi, Michoacán, Jalisco, among others.

Nahuatl is one of the American languages most studied and documented [1], [4-7], [9-11], there are several documents written in Nahuatl from which we can extract important information valuable to present and future generations.

Our interest in developing language resources for Nahuatl is primarily based on the following:

- Due to ignorance and mismanagement of the language terms, we are losing much of our culture. It is therefore important to accurately recognize and extract the information contained in documents written in Nahuatl;
- Preserving a language with historical roots, the loss of the Nahuatl language would represent the loss of part of the Mexican essence and identity. To better understand the cultures that still speak this language, as well as to communicate with people who only speak Nahuatl.

Moreover, the research field of Natural Language Processing (NLP) is a sub-discipline of Computer Science and Linguistics [2], [3], which is responsible for producing computer systems that facilitate the communication between man and man or man-machine using natural language. The purpose of NLP is to study the problems of automatic generation of natural language understanding. Some relevant applications of NLP are:

- Automatic Translation
- Speech Recognition
- Voice synthesis
- Extraction of information
- Information Retrieval
- Automatic generation of summaries
- Handwriting recognition
- Text Mining
- Question Answering

To build these applications, the NLP is assisted by linguistic resources.

Linguistic resources [8] are a set 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 or systems aimed to separate, collect, manage and use other resources. In this paper we present two software tools that allow us to analyze words in Nahuatl.

## 2 Development of the Computer Systems

In order to perform analysis of the Nahuatl language we have developed two software tools: one for prefixes and suffixes of words in Nahuatl text and another for the translation of the Nahuatl-Spanish terms, wich also provide semantic information of

words in Nahuatl. Because the Nahuatl is an agglutinative language, that is, prefixes and/or suffixes are added to the root of a word giving them complex meanings, it was necessary to develop a system that allows us to analyze the words in their most basic form. On the other hand, in order to develop future specialized applications, such as a stemmer, we require more information related to the word, since the mere corresponding meaning in Spanish will not be enough.

The steps followed to develop the system to obtain prefixes and suffixes are as follows:

1. System requirements. We searched and analyzed Nahuatl documents to form a collection of texts in txt format.
2. System design. The system design consists of three layer architecture: interface, application logic and storage.
3. Implementation and testing. The System was implemented in the programming language C #.

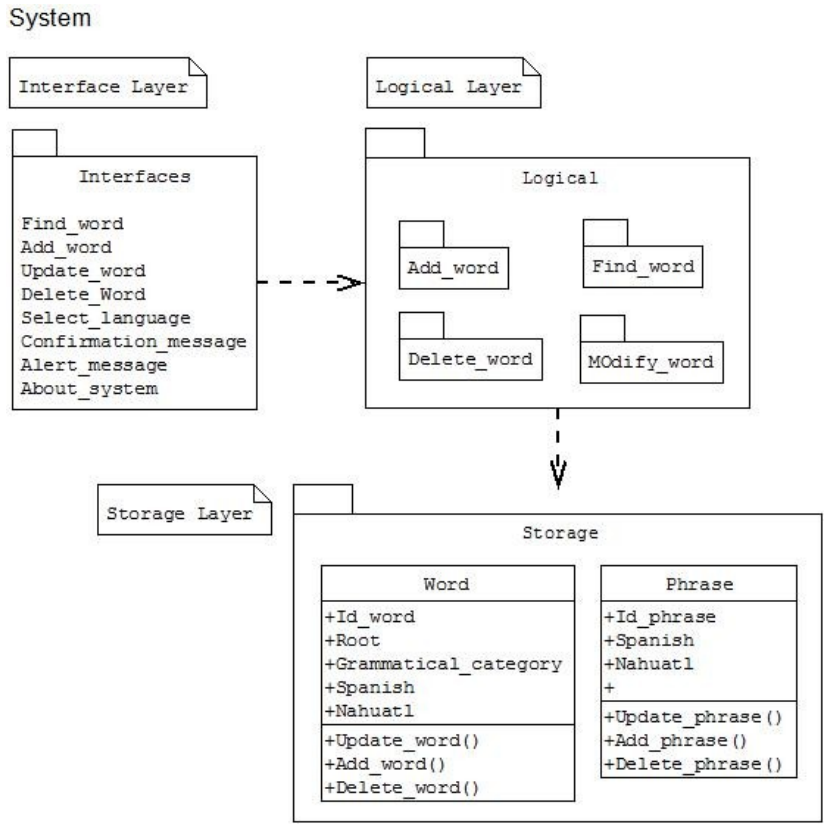


**Fig. 1.** Architecture of the translation and semantic information system.

Development of the Nahuatl-Spanish translator which includes semantic information was performed by the following process:

- System requirements. We identified the main functions of the system to keep the collection of terms, which are: search, add, modify and delete.
- System design. The system design consists of three layer architecture: interface, application logic and storage. Figure 1 shows the architecture interface of the system.
- Implementation and testing. The system was developed in C #, and the database was implemented in MS-Access 2000 because the database manager does not depend on a server to operate.

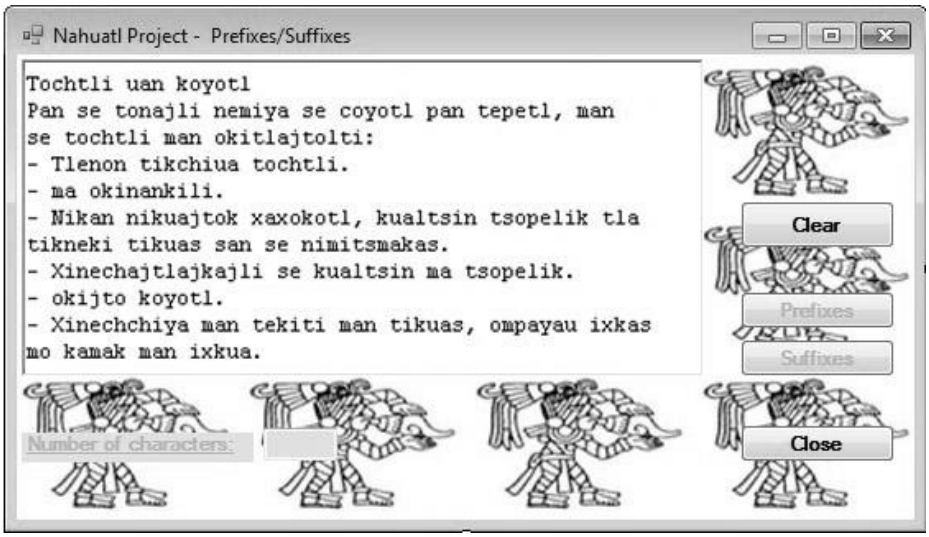


**Fig. 2.** Interface of the computer system to obtain prefixes and suffixes.

## 3 Results

Our results are two software systems, one for prefixes and suffixes of words in Nahuatl and another for the translation of the Nahuatl-Spanish terms and vice versa, which also shows semantic information related to the term in Nahuatl.

To test the first system we have a collection of 836 texts in Nahuatl classified into four categories: Poetry, Stories, Religion, and Miscellaneous. Figure 2 shows the interface of the system with an input text. Figure 3 shows the output of the system with prefixes of size 6 letters.

Figure 4 shows the computer system interface translation (using Spanish and Nahuatl languages) of terms and semantic information related to the term in Nahuatl. The semantic information is the root or roots of words and grammatical category,

which can be: a noun, adjective, pronoun, preposition, conjunction, article, adverb or verb. The system currently contains 1514 terms.

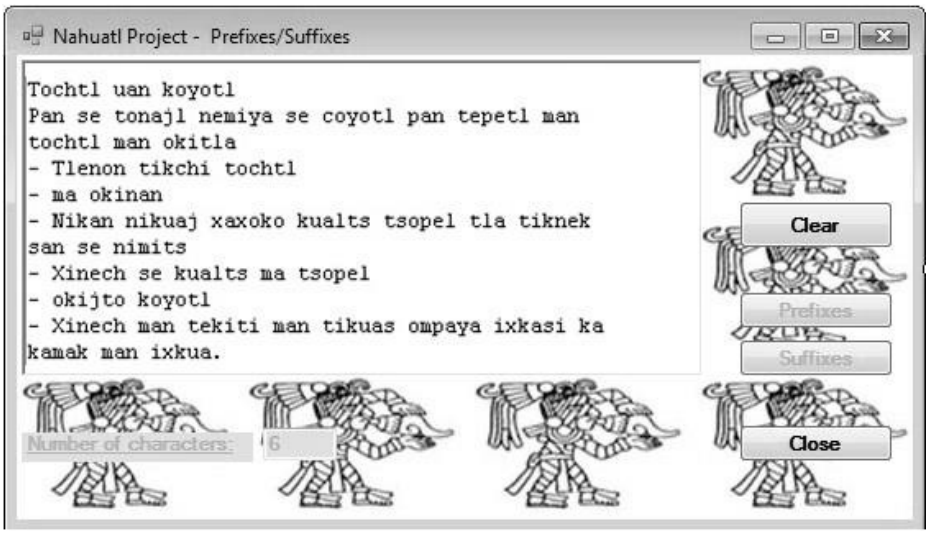


**Fig. 3.** Output of the computer system with prefixes of size 6.

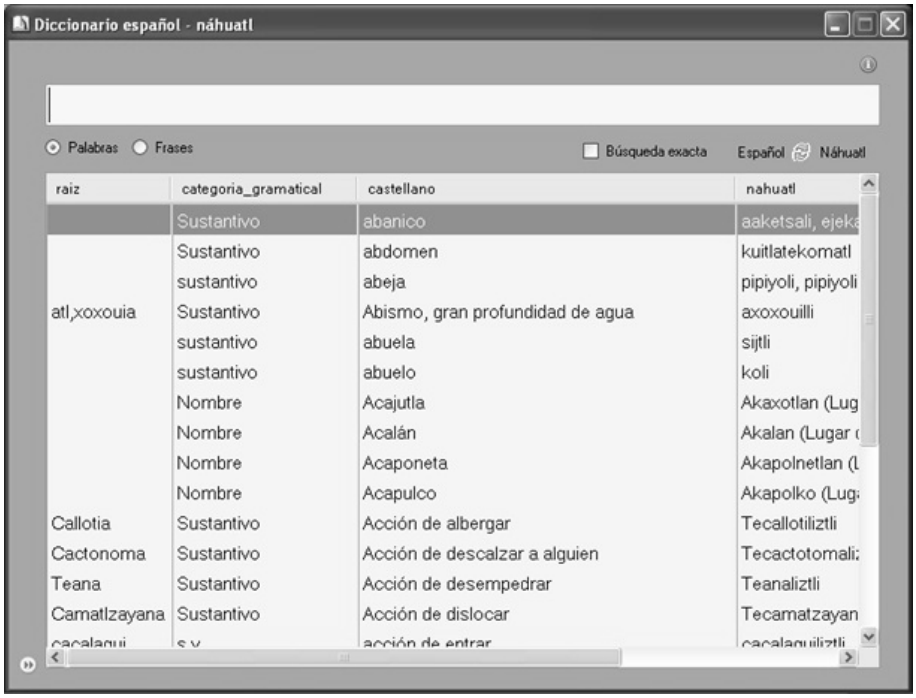


**Fig. 4.** Computer system interface for translation of terms and semantic information.

## 4 Conclusions

In an effort to preserve the use of pre-Hispanic Nahuatl language this article presents two computer systems that allow us to analyze words written in Nahuatl. One system automatically extracts the prefixes or suffixes of words from a text written in Nahuatl, and another system translates Nahuatl-Spanish terms and vice versa, and also shows semantic information related to the term in Nahuatl. Currently our computer system contains a database with 1,514 terms.

## 5 Future Work

As future work, we will develop a stemmer to continue the analysis and study of Nahuatl. To implement the stemmer we will use the two computer systems described in this work. As well as continue to develop linguistic resources for language Nahuatl such as part-of-speech tagging and statistical parsing.

## References

1. Andrews, J. R.: Introduction to Classical Nahuatl. University of Oklahoma Press (2003)
2. Brill, E., Mooney, R. J.: An Overview of Empirical Natural Language Processing. AI Magazine. vol. 18, No. 4 (1997)
3. Bolshakou, I., Gelbukh A.: Computational Linguistics. Ciencia de la Computación. IPN-UNAM-FCE, México (2004)
4. Garibay K. Á. M.: Panorama Literario de los pueblos nahuas. Editorial Porrúa, México (1997)
5. Garibay K. Á. M.: La llave del náhuatl. 9ª edición. Editorial Porrúa. México (2007)
6. Langacker, R. W.: Studies in Uto-Aztecan Grammar. Moder Aztec Grammatical Sketches. Vol. 2. Summer Institute of Linguistics (1979)
7. Launey, M.: Introducción a la lengua y a la literatura Náhuatl. México D.F., UNAM. (1992)
8. Ortega-Mendoza R. M.: Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado. Tesis de MAESTRÍA. Instituto Nacional de Astrofísica, Óptica y Electrónica (2007)
9. Saunders P., R.: A grammar of Tetelcingo (Morelos) Náhuatl. Journal of the Linguistic Society of America. Vol. 30, Num. 1 (1954)
10. Siméon, R.: Diccionario de la Lengua Náhuatl o Mexicana. [Paris 1885] Reprint: México (2001)
11. Wolgemuth, Carl: Gramática Náhuatl. Instituto Lingüístico de Verano. México (2002)