Development and Verification of a Verbal Corpus Based on Natural Language for Ecuadorian Dialect

Wilbert G. Aguilar[1, 2, 4], Darwin Alulema[3], Alex Limaico[1,3] and David Sandoval[1,3]

[1] Centro de Investigación Científica y Tecnológica del Ejército, CICTE
[2] Departamento Seguridad y Defensa, Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
[3] Departamento de Eléctrica y Electrónica, Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
[4]Grup de Recerca GREC, Universitat Politècnica de Catalunya, Barcelona, Spain
{wgaguilar, doalulema, adlimaico, dssandoval }@espe.edu.ec

***Abstract*— The use of the corpus becomes essential in the development of applications based on natural language processing (NLP). In Ecuador, these applications are incompatible because in each region use words outside the context of Spanish. This article presents the development of a corpus compatible with Ecuadorian natural language words. We applied a identification algorithm to take advantage of local literature and power a new data base. The corpus mounted is verified by a quantitative and qualitative comparison with an open access corpus. The result is the first corpus in this country with high scalability and great versatility.**

Keywords-component; Computational linguistics, Natural language processing, Text processing, Database

I. INTRODUCTION

Communication of computers and humans is essential for developing applications such as artificial intelligence, Human Computer Interaction (HCI), computer science and linguistics. These applications must be understandable and be able to keep interaction with people. One of the main problems to be solved in Natural Language Processing (*NLP*) is the access to databases. A word database is defined as Linguistic Corpus, it has specific characteristics that allow to be used in processing text algorithms. In the future, this will depend of general advances in natural language processing, and the expanding the capabilities of traditional corpus [1].

Most of the corpus are developed by linguistic academies, basing their collection on structures attached to words spelling rules. A local issue is that for example in Ecuador has its own Natural Language and uses different words than Spain. Each word has his own function in the sentence and we need to identify it. The most used techniques are the labeling of words according to their grammatical characteristics.

The natural language of Ecuador added linguistic expressions such as cultural expressions, spelling errors, words from Quichuan Language (quichuismos), and other changes that should be considered [2]. To use *NLP* based applications in Ecuador or another Spanish-Speaking Country must consider all expressions within the linguistic corpus used.

To develop an Ecuadorian corpus is necessary include several linguistic cultures then we use Ecuadorian literature examples. Our proposed corpus was verified and implemented in Python, using Natural Language Tool Kit (NLTK) [3].

In the state off the art it has not been incorporated own expressions of a Spanish-Speaking country in a Spanish corpus [1]. Existing development tools are based on a set of language models and classifiers. Acoustic-Phonetic Decoding modules [4] assumes that each language has its own phonotactic characteristics categorizing sequences of phonemes.

We have created a corpus with only Ecuadorian verbs, which is based on a Spanish corpus, and has been modified to adapt to our natural language. A method for extracting words has been used, which takes advantage of a literary collection. The extracted words are subjected to an analysis to obtain verbs used only in the country, comparing with a database in Spanish. The incorporation of labels also is important. Artificial intelligence applications need a timely and effective access to the data, where the computer learns objectively [5, 6]. For future work, it can be improved by optimizing algorithms and creating corpus with all types of words and more complex labels.

This paper is organized as follows: In section two we present the previous work. In Section three we define our proposal. Section four describes our implementation and finally in section five we make a check.

II. PREVIOUS WORK

Corpus is a labeled database that contain words, expressions, conjugations, etc. The scope for developing a corpus is to attach a fundamental tool in the *NLP*. The objective of *NLP* is to design and develop software able to analyze, understand, and generate languages that humans use naturally. This will lead into increment the HCI. [7].

Part of the development of a corpus throughout the world is the use of various techniques for processing. Including statistical methods, connection methods or direct thresholds. The corpus will be described by a set of keywords called index words for this case, we setting the words with more weight determined from the base corpus. The definition of statistical natural language processing given by [8]. The processing of a new corpus can be divided into preprocessing and parameterization. The first is used to remove superfluous for

978-1-5090-4284-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ICSC.2017.82

example in this corpus are the stop words. The parameterization is to set weights to verbs more relevant.

For the elaboration of corpus there are the constructions concerning the verb in relation to the noun; sectioning, labeling and identifying of verb forms [9]. Inside the development of corpus, it is necessary to consider data formats, whether ID, LEMMA (root), HEAD, among others [4]. Once marked the data in the lemma format, they will undergo evaluation metrics as the thresholds described above.

One problem when you pick up a corpus is the copyright sued by certain sources. Copyright problems can be solved by choosing free corpus as possible. If republishing is not possible, at least it is possible to use the results as quantitative data and the texts will be shared with interested on private [10].

After the preparation and verification, certification is necessary. Certification involves setting appropriate parameters and verify that there are no errors. We will determine certain parameters and quality indicators, then a threshold is chosen for certification [11].

The corpus focuses on the extraction of verbs which implies that should set parameters and methods for their extraction. In [7] a method of extraction of verbs based on statistical methods is proposed. The nature of verbs is not regular and even if these verbs considered are based on languages such as Quechua which differ in their syntactic structure. You cannot use directly dictionaries which define the verb and its synonyms. On the contrary, what is done is to analyze the semantic context, discarding supplements, nouns and own words through a stemmer developing with NLTK.

Some corpus as SENSEM [12] are focused on the direct extraction of verbs and their location in the sentence, it refers in particular syntactic and semantic patterns. It is analyzed: reciprocal endings as: ‘ar’, ‘er’, and ‘ir’. Finally, periphrastic constructions determined from complementary words as: 'hacer' plus an infinitive.

The development of an Ecuadorian corpus involves accessing digital documents that are full of words used in the country. Digital documents belonging to Ecuadorian literature are analyzed. In the processing of Ecuadorian literature as "Huasipungo" it is common to find natural expressions or sentences that are repeated constantly. This expression called placement [13] serves to optimize processing. This involves making trees consecutive words that are repeated within each paragraph, then assign a unique identifier like: template base, nominal sentence or predictive relationships.

III. Development

To develop Ecuadorian Corpus, we use digital books written by local authors and based in everyday events. An example used is a traditional book called “Huasipungo” where is normal to find Ecuadorian expressions repeatedly.

The goal is to retrieve this information and natural language of the Ecuadorian literature using NLP recovery techniques. An open consultation is effective, as suggested in [8]. In “Fig.1”, it is described how the system retrieves textual information with a simple comparison.

Developed the corpus is necessary to detect errors. Errors can be of two types according to [11], spelling and syntax. Spelling errors can be detected in the following ways:

- Making a list of n-grams in which the words are verified. This verification is enhanced by use of a threshold that limits the frequency of words. This checklist must be completed with the words of the end user, in this case Ecuadorians.
- Reference to dictionaries using LEMMAS in dictionaries and usage rules that allow verifying the words.

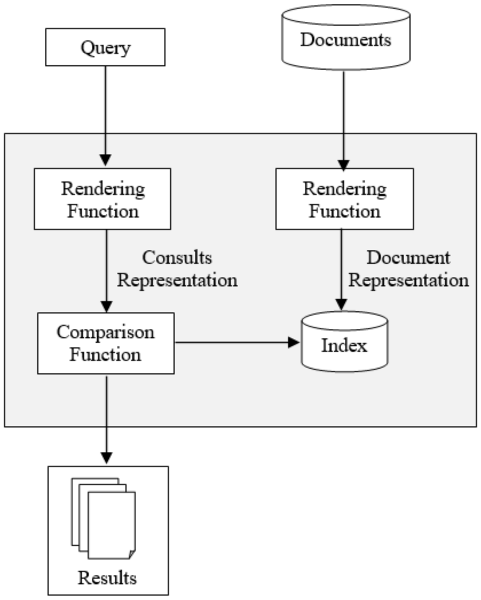

Figure 1. Architecture of an information retrieval system.

Syntax errors require consideration of the relationship between words in a sentence, and you can also check in two ways:

- Using a corpus trademarks or brands which are used to determine reference syntax errors.
- A set of syntactic rules, this depends on the language and requires that the rule system be very flexible and supports certain peculiarities.

This type of detections do not transcend in the corpus because are not flexible. It is necessary that the words are written correctly according with the popularly established. The appropriate approach is to use n-grams of words or own labels established in literary references of Ecuadorian writers.

The development of linguistic studies commonly used for applications in the computer, it is essential when working or solve a problem. In some cases, the computer is limited to do the user requests. If the computer needs to understand what the user types literally, the computer will not understand. There are various methods and applications that working in the performance and needs of the user. One method is through the machine learning, where asking questions to databases in natural language, improves the use of data stored in databases [14].

In the article, the processing of sentences entered by the user is presented written in Spanish. the respective analysis from the verb takes place, since it is the fundamental part of a grammatical sentence. Sentence is the smallest grammatical unit complete sense, and an act of coherent communication [9]. So this is the main point of study, and where you should begin the study. Processing a sentence, through software, it is the cornerstone for understanding a human language by a computer.

We propose to use Python 3.4. The development and promotion of Python is carried out through a non-profit organization, called Python Software Foundation [15]. This software has great advantages because of its variety of functions and libraries available on the web, also is cross-platform. We use Python because of its ease of interpretation of language. It also has a downloadable library that is able to work directly with the language known as Natural Language Tool Kit (NLTK).

A. Ecuadorian verbal corpus

In Ecuador, the official language is Spanish, therefore, we can use a corpus of Spain for the "PLN", however, it is discarded words used in the country, is necessary create a corpus with verbs that the country's inhabitants are commonly used the idea is to add new words to a general corpus of Spanish language, and create a new larger corpus

To create the new corpus and get Ecuadorians verbs have three cases where they can be obtained:

- From native languages, such as Quichua.
- Exclusive expression of a language (Idioms).
- New words by fads or trends.

Natural language in the Ecuadorian highlands is influenced by the Quichua and other expressions from the native languages of the country some verbs became popular and now commonly uses, these verbs have a different structural form verb in Spanish, but few of these words are used therefore an Ecuadorian Quichua book teaching language is used, in this book the most commonly used verbs are indicated [10].

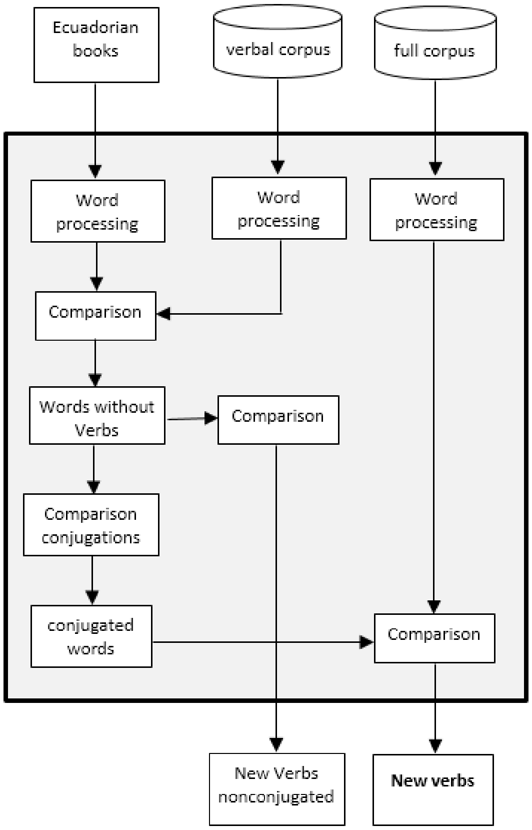

Figure 2. Architecture algorithm to obtain new verbs.

Idioms, have been used for several years in the country therefore use the linguistic resources of Ecuador where these words are used an algorithm is created to obtain the necessary words of Ecuadorian books for which is used as a resource for comparison existing corpus in Spanish

The Stopwords are removed from the Spanish corpus, verbs that match those in the corpus are extracted and it remains only words that do not appear in the database. Spanish verbs with conjugations are characterized by their endings these terminations are compared to have possible verbs to check the algorithm is compared to a database containing all the words not only verbs as a final filter, the words are not the same with these are considered as new words or verbs, also it keeps the words before passing them by comparing the conjugation also the comparison is made with the general database. This structure is represented by "Fig. 2".

IV. IMPLEMENTATION

Processing a digital book is performed by means of an application consists of four parts:

- Separation of plane text in sentences.
- Separation the subject and predicate of a sentence.

- Recognition of verbal conjugation used in sentence.
- Labeling and storage of words.

The user enters a text file, and is obtained as a result database processed according to the previously selected function. The application consists of the following parts illustrated in "Fig.3".

- Window cover.
- Main Window.

This application is focused on training with the Ecuadorian language therefore it can be used for learning. The windows also have the following aids to the user illustrated in "Fig.3.B":

- File
- Additional Information
- Tutorial videos
- About

The application starts with a cover and then a main window with two buttons that open Subject and Predicate separator or Verbs detector depending on the case. In the separator, the sentences are separated into subject and predicate also it has validated cases in which subjects are tacit or not a verb is entered. In the detector, after detecting a verb it is compared with the full corpus and storing if is new local word.

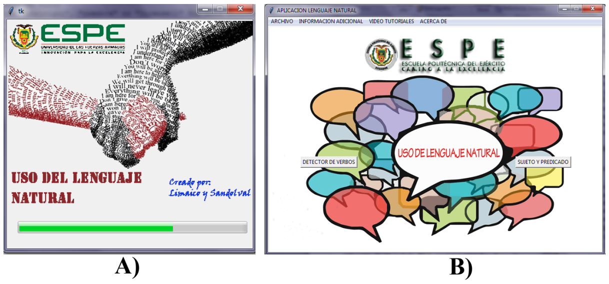

Figure 3. A) Window cover. B) Main Window.

We improve the application based on feedback and errors. It was realized by distributing the tool to linguistic specialists that have text files recovered in their studies.

V. Check

Existing resources for Spanish within NLTK not contain what is necessary to perform this application, the library "cess_esp" which it contains several words that have their own labeling. They have a lot of words are not enough for the application this library has entered text with a total of 6030 sentences and 192,686 words. The labels used to represent morphological information are based on the morphosyntactic annotation lexicons proposed by the EAGLES group [11].

No good results from this library are obtained, and therefore the application does not work properly. The Spanish verbal corpus is then used to perform processing of sentences, but this corpus does not contain verbs native of Ecuador, therefore, it is necessary to expand the corpus, for better performance.

By the algorithm to obtain new verbs, whose architecture is illustrated in "Fig. 2", the database is increased, and then monitoring the results, and writing of existing conjugations for each new verb extends the initial corpus.

The new words entered in the first modification of the corpus it is added to a total of 494 new verbs of which 80 are most commonly used verbs Quichua and 414 obtained from idioms used by Ecuadorians "Fig. 4". That modification use three popular books and two specialists text files. Inputting more books, we can further increase this list of verbs and thus complement it.

Using the new verbal corpus, we proceed to verify the database, using manual asseveration and aleatorily verification analyzed with specialists and our knowledge.

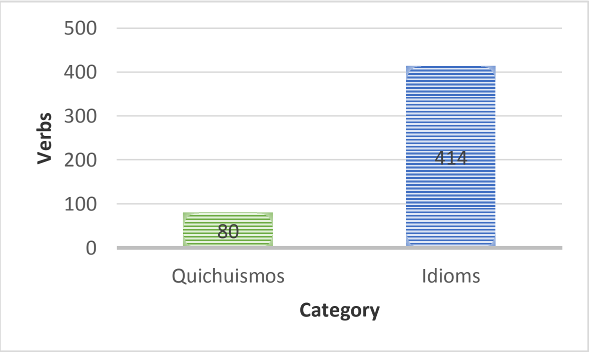

Figure 4. Number of verbs admitted to the corpus, according to their category.

The database generated in the sentence separator was used to make a comparison between sentences. Taking advantag of the library "cess_esp" the original corpus of Spanish verbs and verbs with Ecuadorian modified corpus.

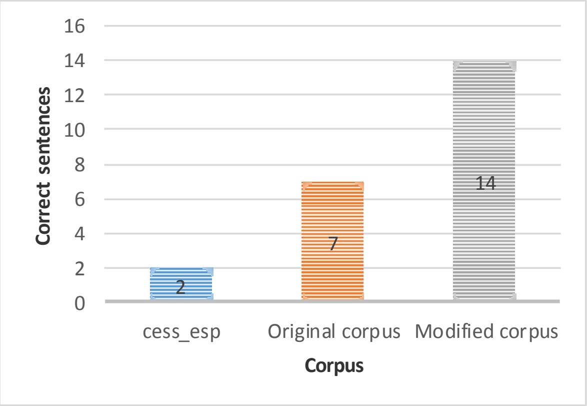

Figure 5. Comparison corpus, according correct sentences obtained.

In the separation function, 16 sentences used in Ecuador are entered, with different corpus loaded into the application to verify the correct results for each corpus. The modified corpus presents the best results, as illustrated in “Fig. 5”.

VI. CONCLUSIONS

The algorithm can be used to enter text as text messages, social networking states, documents, etc. to extract words frequently or new verbs that have been created by young Ecuadorians trends, and thus improve the corpus use.

The corpus made allows the implementation of the application for processing local books with better results. Then this corpus could be used to improve the HCI applied to Ecuador.

To obtain an optimal processing algorithm language is necessary to revise the parameters of the language, the corpus created is an improvement of an existing Spanish corpus, which adds to that corpus new books verbs extracted from native country.

The development of a corpus with the presented method implies the necessity of a monitoring system, because it can present orthographic errors or lack of words and meaning, in the obtaining of verbs from several sources.

The results shown, can be improved as they are pureed words to the corpus, because the more verbs are added, this will have a better performance.

REFERENCES

[1] A. C. a. K. S. Jones, "Natural language interfaces to databases.," *The Knowledge Engineering Review,* pp. 225-249, 1990.

[2] G. S.G., "The Vowel Systems of Quichua-Spanish Bilinguals," *Phonetica,* vol. 60, 2003.

[3] D. Colton, Text Mining and Visualization: Case Studies Using Open-Source Tools., 2016.

[4] M. G. a. L.-F. H. Emilio Sanchis, "Language identification with limited resources," *Jornadas TIMM,* pp. 7-10, 2014.

[5] A. M. B. M. C. Q. E. R. Anthony Aue, "Statistical Machine Translation Using Labeled Semantic Dependency Graphs," *Microsoft Research - 1 Microsoft Way,* 2016.

[6] N. D. J. R. G. Stephen H. Shum, "Limited Labels for Unlimited Data: Active Learning for Speaker Recognition," *INTERSPEECH,* 2014.

[7] B. Preeti, "NATURAL LANGUAGE PROCESSING," *International Journal Computer,* 2013.

[8] M. V. y. R. Pedraza-Jimenez., "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines".

[9] A. Todirașcu, A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions, Barcelona, 2008.

[10] M. Lounela, "A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions," *Universidad de las fuerzasarmadas ESPE,* 2008.

[11] C. Grouin, "• Certification and cleaning up of a text corpus: towards an evaluation of the “grammatical” quality of a corpus".

[12] G. V. I. C. Ana Fernández, «SENSEM: base de datos verbal del español,» *Universitat Autònoma de Barcelona (EUIS), ,* p. 8, 2004.

[13] S. Torres-Ramos, "Extracción automática de un diccionario de colocaciones en español".

[14] H. M. M. &. K. L. Bais, "Querying database using a universal natural language interface based on machine learning," *International Conference on Information Technology for Organizations Development ,* pp. 1-6, 2016.

[15] A. Todirașcu, A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions, 2008.

[16] R. Orbegozo, "Todo lo que debo saber sobre Tecnológia, Python y Linux," 3 Enero 2015. [Online]. Available: https://ricardo705.wordpress.com/2015/01/03/python-2-x-y-python-3-x-diferencias-de-sintaxis-en-solo-4-paginas/. [Accessed 10 Septiembre 2015].

[17] M. Taule and M. R. Antónia Martíı, *AnCora: Multilevel Annotated Corpora for Catalan and Spanish,* Morocco: LREC, 2008.

[18] A. Martí, Tecnologías del Lenguaje, Barcelona: UOC, 2003.

[19] A. López, "Programa Universal de Estudios," in *Lengua y Literatura*, Madrid, CULTURAL,S.A, 2002, pp. 7 - 103.

[20] F. Escolano, M. Cazorla, I. Alfonso, O. Colomina and M. Lozano, Inteligencia Artificial: Modelos, Técnicas y Áreas de Aplicación, Madrid: THOMSON, 2003.

[21] A. Castro, "AB INTRA," 24 Junio 2011. [Online]. Available: http://alejo-ab-intra.blogspot.com/2011/06/diccionario-de-verbos-ecuatorianos.html. [Accessed 11 Septiembre 2015].

[22] Ahmed, "La vida en un Punto," 18 Febrero 2009. [Online]. Available: http://lavidaenunpunto.blogspot.com/2009/02/ecuatorianismos.html. [Accessed 11 Septiembre 2015].

[23] E. Bahit, Curso: Python para Principiantes, Buenos Aires: Creative Commons Atribución, 2012.

[24] N. A. C.-S. y. G. Sidorov, "Extracción automática de los patrones de rección de verbos de los diccionarios explicativos".

[25] NLTK Project, "NLTK 3.0 documentation," 5 Septiembre 2015. [Online]. Available: http://www.nltk.org/#. [Accessed 10 Septiembre 2015].

[26] C. Aguilar, *Curso de Procesamiento de Lenguaje Natural,* Chile: Pontificia Universidad Católica de Chile, 2012.

[27] A. Bello, "Conjugación del verbo Regular," 1 Julio 1998. [Online]. Available: http://www.verbolog.com/0amar.htm. [Accessed 14 Septiembre 2015].

[28] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, O'Reilly Media Inc, 2009.

[29] A. Gelbukh and G. Sidorov, PROCESAMIENTO AUTOMÁTICO DEL ESPAÑOL, Mexico: INSTITUTO POLITÉCNICO NACIONAL , 2006.

[30] A. F. Montoro, Python al Descubierto, Madrid: RC Libros, 2012.

[31] I. Olea, "GitHub," 7 Octubre 2015. [Online]. Available: https://github.com/olea/lemarios/blob/master/verbos-espanol-conjugaciones.txt. [Accessed 11 Septiembre 2015].

[32] R. Pino, A. Gómez and N. d. Abajo, "Introducción a la Inteligencia Artificial," in *Sistemas Expertos, Redes Neuronales Artificiales y Computación Evolutiva*, Oviedo, UNIVERSIDAD DE OVIEDO, 2001.