ISSN 2007-9737

Building a Nasa Yuwe Language Corpus and Tagging with a Metaheuristic Approach

Luz Marina Sierra Martínez[1], Carlos Alberto Cobos[1], Juan Carlos Corrales Muñoz[1], Tulio Rojas Curieux[1], Enrique Herrera-Viedma[2], Diego Hernán Peluffo-Ordóñez[3,4]

[1] University of Cauca, Computer, Departments of Computer Science, Telematics and Anthropology, Popayan, Colombia

[2] University of Granada, Department of Computer Sciences and Artificial Intelligence, Granada, España

[3] University Yachay Tech, School of Mathematics in Imbabura, Ibarra, Ecuador

[4] Corporación Universitaria Autónoma de Nariño, Electronic, Pasto, Colombia

{lsierra, ccobos, jcorral, trojas}@unicauca.edu.co, viedma@decsai.ugr.es, dpeluffo@yachaytech.edu.ec

Abstract. Nasa Yuwe is the language of the Nasa indigenous community in Colombia. It is currently threatened with extinction. In this regard, a range of computer science solutions have been developed to the teaching and revitalization of the language. One of the most suitable approaches is the construction of a Part-Of-Speech Tagging (POST), which encourages the analysis and advanced processing of the language. Nevertheless, for Nasa Yuwe no tagged corpus exists, neither is there a POS Tagger and no related works have been reported. This paper therefore concentrates on building a linguistic corpus tagged for the Nasa Yuwe language and generating the first tagging application for Nasa Yuwe. The main results and findings are 1) the process of building the Nasa Yuwe corpus, 2) the tagsets and tagged sentences, as well as the statistics associated with the corpus, 3) results of two experiments to evaluate several POS Taggers (a Random tagger, three versions of HSTAGger, a tagger based on the harmony search metaheuristic, and three versions of a memetic algorithm GBHS Tagger, based on Global-Best Harmony Search (GBHS), Hill Climbing and an explicit Tabu memory, which obtained the best results in contrast with the other methods considered over the Nasa Yuwe language corpus.

Keywords. Part of speech tagger, Nasa Yuwe language, tagged corpus, harmony search, global-best harmony search, hill climbing, tabu memory.

1 Introduction

This research has been motivated by the need to support the revitalization and technological visibility of Nasa Yuwe (Páez), an official language in the Republic of Colombia spoken by 75% of the Nasa indigenous community, since its sociolinguistic situation places it in danger of extinction due to cultural, social, geographical, and even historical factors [1].

Information Technology (IT) has been involved through the development of various initiatives that include educational materials such as games, educational resources, and methodologies for its construction. Further strategies have sought to support the teaching, and use, of the language by visibilizing and sensitizing its use through computational tools [2, 3].

The inclusion of IT in the teaching and revitalization activities of Nasa Yuwe seeks to take advantage of the options available to a teacher in a combined learning environment (classroom + activities supported by computer resources), which is addressed in the same direction of the current dynamics of the Nasa community. The

development of these types of technological strategies has forced both Nasa speakers and those interested in the revitalization of this language to think about crucial aspects of this, such as: is it possible to access written documents in Nasa Yuwe, from any-where?; do the available documents go beyond just being an electronic document or can they be used for the different revitalization activities?; how well is the language known in its written form?; is it possible to create technological tools that allow the development of more complex activities in teaching Nasa Yuwe?

As a result, to continue working on technological solutions applicable to the teaching and revitalization of the language, which allow the analysis and advanced processing of the language, the construction of a Part-of-Speech Tagger (POS Tagger) for Nasa Yuwe in computer learning environments is crucial. This will allow the introduction of complex reading and writing activities in which the learner has to must create and identify correct sentences, considering grammatical elements of the Nasa Yuwe language. This is considered novel and valuable as an original contribution at each of the linguistic, anthropological, and computational levels, since there are no works in this sense relating to Nasa Yuwe or for languages with similar characteristics. A POS Tagger [4] would be a great resource that would provide many possibilities for the Nasa language, since it would be the basis for the development of several additional applications such as voice recognition systems, text-to-speech, text classification, automatic information retrieval systems, multimedia information retrieval systems, sentiment analysis, and resolution of ambiguities in the meaning of words in a context, among others [5].

However, for the building and quality evaluation of a POS Tagger for Nasa Yuwe, it is necessary to have in place such linguistic resources as a tagged corpus for this language, which is not a trivial task, since it is time consuming and expensive, especially for the development of applications in new domains such as languages either poor in linguistic resources or where none exist at all, as is the case of Nasa Yuwe. This work therefore focuses on presenting linguistic manual tagged corpus for the Nasa Yuwe (Páez), language and its process of building and the uses of this corpus with existing taggers such as Random tagger, three versions of a tagger based on the Harmony Search (HS), metaheuristic and three versions of a memetic tagger based on Global-Best Harmony Search (GBHS).

The rest of the paper is organized as follows: Section 2 provides a background on the Nasa Yuwe language and related works on building a corpus for traditional and non-traditional languages and the most relevant techniques for build POS Taggers; in Section 3, the methodology used for the process of building the Nasa Yuwe corpus and some details about the experiments carried out using the Nasa tagged corpus built; Section 4 presents details of the Nasa Yuwe corpus; Section 5 explains in detail the experiments conducted; and finally, Section 6 presents conclusions and intentions for future work.

2 Brief Background and Related Works

2.1 Brief Description of Nasa Yuwe Language

Nasa Yuwe is the language spoken by the Nasa people, who are located across seven different regions (departments) of the Republic of Colombia: Cauca, Huila, Tolima, Valle del Cauca, Meta, Caquetá, and Putumayo, with Cauca having the largest population [6]. Interaction with other communities, the market, state entities, private entities, and the Church was carried out in Spanish, making Nasa Yuwe a minority spoken language [1]. Currently, Nasa Yuwe is spoken more by adults rather than by young people or children and what is more, for some, Spanish has arisen as their primary language. Despite efforts made to maintain their culture, the language of the Nasa has suffered a series of processes that have threatened its conservation [1].

Nasa Yuwe had for many years been included within the Chibcha family [7, 8, 9], but in 1993 Constenla [10] determined that this classification was not correct. As a result, it was classified as an independent language [1, 11]. Nasa Yuwe has by tradition been an oral language. Only as recently as the year 2000 was it possible to unify the Nasa alphabet. Nasa Yuwe is still a language in the process of description. Some relevant studies on this language are: Jung in 1984 [12], CRIC in 2005,

and Rojas in 1998 [13] and 2012 [1]. To carry out the tagging of the corpus we used that presented by Rojas in 1998 [13] and in 2012 [1]. The formation of a word in Nasa Yuwe requires the presence of at least one simple radical per word, which should appear on its own or accompanied by flexional morphemes or derivative morphemes [1, 13]. The relationship between types of word and predication is important. The word classes defined by Tulio Rojas (linguist, expert in several indigenous languages and with more than 40 years of experience in the study of Nasa Yuwe) are [1, 13]:

- Predicative word: 1) Predicative base with lexical radical, for example: *tulyuth* (*I am Tulio*), *me-mi'kwe* (*you* (pl) *sing*), *walatha'w* (*we are great*). 2) Predicative base with grammatical radical. For example: personal pronouns (*idxgu*, *it is you*), demonstrative pronouns (*txa*', *it is that*), spatial deictics (*ayte*', *it is here*), interrogatives (*madzna'*, *how much is it*?), quantifiers (*weha'*, *it is not much*). 3) Negation. For example: *thegmeth* (*I did not see*), *walameg* (*you are not great*).
- Noun. This is the construction resulting from the application to a lexical base of a set of flexural marks, for example: *alku* (*dog*).
- Qualifying word, a qualifying radical can enter into the formation of a predicative word and into the formation of a qualifying word.
- Connector, these words do not have flexion, in addition they cannot be predictive bases. They are used as connectors in the sentence. Examples: *Sa'* (*and*), *atsa'* (*so*), *napa* (*but*).

It should be noted that, in Nasa Yuwe, articles are not found as a kind of word.

2.2 Related Works

A linguistic corpus is a vital part of NLP. Its content must be chosen to support its purpose, such as studying a language.

In general, terms, a corpus is made up of a collection of authentic texts readable by a machine (including spoken data transcriptions) which are representative of a natural language [14]. The aim in building the linguistic corpus for Nasa Yuwe is tagging the parts of speech. Therefore, to establish the main characteristics and elements that constitute the corpus and the different methods of tagging, several works have been reviewed for both traditional languages and non-traditional languages.

2.2.1. TagSet for Tagged Corpus

The tagset may vary for each language according to contexts and morphological structure, so that variations and unification trends are found, as well as different methods for carrying out tagging of the words that make up the texts. There follows a selection of related works: in 2014, Dinakaramani, et al [15] established a set of 23 POS Tags to tag 10,000 sentences from the IDENTIC corpus of the Indonesian language, containing 262,330 tokens. They defined three principles for the tagset (linguistically valuable, simplicity, automatically refined) and a methodology for manual tagging of the corpus with the proposed tagset (for the manual tagging, two human annotators were used).

In 2013, Ismael, et al, [16] presented an algorithm that compiles 320,443 Bangla words collected from newspapers, blogs, and other websites, and tags them as name, verb, and adjective, finding that the algorithm has more accuracy for verbs than for names and adjectives. In 2012, Petrov, et al, [17] presented a set of 12 unified tags from 25 tagsets for 25 languages from previous works. The proposal seeks to improve the accuracy of part-of-speech taggers across several languages.

The 12 POS tags defined by Petrov were: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation) and X (a catch-all for other categories such as abbreviations or foreign words). In 2008, Baskaran, et al, [18] presented IL-POSTS, a framework containing a tagset for most Indian Languages, taking the EAGLES guidelines into account [19], it is intended to be of general use; this paper describes the characteristics of the methodological design and the methodological strategies that give rise to the framework. Also in 2008, Rabbi, et al, [20] presented the procedure

followed for the design of a tagset for Pashto Language, taking into account the EAGLES guidelines for morphosyntactic annotation of corpora [19], obtaining 215 tags distributed as: 26 Tags for Noun, 77 for Verb, 60 for Pronouns, 19 for Adjectives, 15 for Punctuation, 7 for Adverb, 3 for Adposition (prepositions and postpositions), 6 for foreign words and 1 for each Interjection and Conjunction.

2.2.2 Building the Tagged Corpus

Building a tagged corpus as well as its corresponding set of tags is crucial for natural language processing, especially for parts of speech tagging. Some related works are presented below: in 2014, Scherrer, et al, [21] presented a large multilingual corpus for German, French, Italian, and English, which uses automatic processing and tagging of HTML files, uses the Universal tagging proposed in [17] for description of the words. The evaluation was done manually in small fragments of the corpus. The corpus has more than 6 million words for each language.

Also in 2014, Ariaratnam, et al, [22] described the tagging process of 500,000 words collected from Sri Lankan Tamil newspapers, since no corpus is available for Tamil; among the steps followed are, in the first instance, pre-processing, where the sentences were extracted with 20 or fewer words to facilitate the process and a pre-editing of the corpus was done to correct writing errors and eliminate unnecessary spaces. In the second instance, a set of 20 tags was proposed with the support of a linguist. In the third instance, manual tagging was done by creating a tagged corpus of 12,500 words, and due revision was done on the tagging.

As well in 2014, Sing and Banergee [23] presented the tagging of a corpus for the Bhojpuri language (a North Indian language), which uses the BIS scheme, defined in 2010. The corpus data corresponds to approximately 5300 tagged words.

The data were collected from conversations and then transcribed. The tagset includes 33 categories, containing sublevels. In the work, the characteristics of the language are presented, observable in the light of the tagging.

In 2012, Spoustová and Spousta [24] presented the process of constructing a large corpus of Czech, which involved, in the first instance, a manual revision and cleaning of duplicate documents, in the second instance a near-duplicate algorithm to remove duplicate paragraphs from documents using a similarity measure based on an n-gram comparison, in the third instance, a language detection module was developed to remove words from Slovak, which consists of two unaccented words and general language filters.

The corpus contains 2.65 billion words from news and magazine articles, 1 billion words from blogs, diaries, and other non-reviewed literary units, 1.1 billion words from discussions, highlighting the high quality of the corpus words due to human intervention in the process of building the corpus.

In 2010, Ahmed and Qadir [25] described the analysis that was done to define the tagset for Shindi, its application in the tagging of the words, as well as the problems that appeared when applying it.

In 2005, Kohen [26] presented the Europarl corpus extracted from the Proceedings of the European Parliament, which includes versions in most European languages.

This corpus was initially constructed to be used in machine translation tasks. It indicates 5 steps for its compilation (Crawling by the European Parliament website, extract, and map parallel documents, divide text into sentences, prepare corpus for use, and map sentences in the languages).

In 1993, Marcus et al [27] presented the Penn Treebank corpus with a reduction in the tagset in comparison with the tagset of the Brown corpus (48 tags), and considering the syntactic context of the word to be tagged. The tagging process was automatic, with manual correction.

The corpus consists of about 4 million words of American English (World Street Journals) and is widely used for POS Tagging tasks. In 1979, Francis and Kucera [28] proposed the Brown corpus for American English, containing 1,014,312 words in categories of texts (such as reports, editorials, and reviews, among others).

This corpus has been expanded several times and currently has a total of 473 categories arising from the subdivisions of the 82 main tags and is widely used for tagging in English.

2.2.3 POS Taggers

There now follows some related work, grouped by the most important techniques for building taggers:

– Linguistic tagging approach, assigns the corresponding tag to a sequence of words using rules [29]. This approach is expensive and requires more knowledge of the language. Among the relevant works are: Brill (1992) [30] and 1995 [31], which are used today as the basis for new proposals such as: Alsuhaibani et al [32] and Mall & Jaiswal [33] in 2015, among others.
– Statistical tagging approach. These take the longest to run and obtain very competitive results. The purpose of this technique is to assign to each word in a sentence the most likely lexical tag according to the context of the word [34]. The most widely used techniques are: Hidden Markov Model (HMM), Trigram'sn'Tags (TnT) [35], Maximum Entropy Markov Models (MEMM) [36], Conditional Random Fields (CRF) [36]. Relevant works are: Keyaki & Miyazaki (2017) [37], Zhonglin et al (2016) [38], Albared et al (2016) [39], and Sun & Wan (2016) [40].
– Neural Network tagging approach, such as Schmid (1994) [41], Nakamura and Shikano (1989) [42], Hin et al (2017) [43], Kabir et al (2016) [44], Carneiro et al (2015) [45], among others.
– Metaheuristic algorithm tagging approach can use both statistical or rules approach such as: Lv et al (2017) [46], Forsati & Shamsfard (2012) [47] and (2015) [29], Silva et al (2014) [48], Forsati et al (2010) [49], among others proposals.
– Memetic algorithm tagging approach that use a statistical approach as: Sierra et al (2017) [50].

3 Methodology

The methodology used was Iterative Research Pattern [51], which consists of four basic steps: field observations, problem identification, technological development, and field tests. As a basis for carrying out this work, it is assumed that there is no tagged linguistic corpus for Nasa Yuwe, added to the fact that it is the first time that a task like this is carried out in this language.

3.1 Building a Nasa Yuwe Language Corpus

The process followed to obtain the tagging corpus for Nasa Yuwe and the alignment of the corpus with Universal tagging was manual and develop in two iterations:

– In the first iteration, two versions of the annotated corpus were obtained: the first version, corresponded to the tagging of the words in each sentence, using the tagset defined by Rojas [1, 13], (such as Predicative, Qualifying, Noun, Connector, Deictic, Pronoun, and additional label used for Punctuation). The second version of corpus was obtained from the results of applying Delphi technique (for expert judgment) on the first version of the corpus Nasa.
– In the second iteration, likewise, two additional versions of the Nasa annotated corpus were obtained: the third version corresponded to the manual tagging of the words in each sentence, considering the universal tagset [17], which was carried out based on the second version of the tagged Nasa corpus of the first iteration.

Other considerations to highlight in this work are:

– The process to build the annotated corpus of Nasa Yuwe was guided through analysis and review of similar works.
– The correction and adjustments to the corpus versions in both iterations were made manually.
– The learning curve for the task of manual tagging was high, as mentioned before it was the first time that the Nasa language was subjected to this task. It should be noted that Nasa Yuwe speaking teachers (who teach this language in the educational institutions of their community) had not gone into the detail of the problem of studying the role of a word in a sentence in this language. Therefore, several sessions were required for the understanding

of the products that were desired to be obtained with the development of this task, as well as to agree on the process to be followed.
– The task of tagging was worked in sessions of 6 hours per week for a period of approximately 6 months, that is, the tagging speed was very low at the beginning, which improved over time.
– The structure defined for the Nasa Yuwe tagged corpus had similarities with the Corpus Brown (one of the most used [28]), that is, for each sentence, each word was labeled with its respective label, to facilitate its subsequent processing and use.

3.2 Using the Nasa Yuwe Language Tagged Corpus

An experiment was developed to evaluate and compare with different taggers over the Nasa Yuwe tagged corpus. These taggers are based on the following approaches:

3.2.1 Memetic Tagged Algorithm Approach

Three versions of a memetic algorithm called GBHS Tagger presented in [50] that uses the Global-Best Harmony Search metaheuristic [52] (which hybridizes Harmony Search with the swarm intelligence concept proposed in PSO [53]) and includes knowledge of the problem through the use of a local optimizer (based on Hill Climbing and an explicit Tabu memory) for the best harmony of the harmony memory, whose use is controlled by the ProbOpt parameter.

– First algorithm is called GBHS Tagger that involved the local optimizer and the random initialization of the harmony memory, which for effects of the experimentation, were defined 4 values to the ProbOpt parameter, as they were without optimization (0.0), with an optimization value of 0.3, 0.5 and 0.7, so it was named: GBHS Tagger with 0.0, GBHS Tagger with 0.3, GBHS Tagger with 0.5 and GBHS Tagger with 0.7.
– Second algorithm is called GBHS Tagger 2 that involved improved initialization (which fills the harmony memory considering the most likely labels of the word in each sentence) using the Alpha parameter, and the local optimizer with the same values for the optimization parameter. For experimental purposes, it was named: GBHS Tagger2 with 0.0, GBHS Tagger2 with 0.3, GBHS Tagger2 with 0.5 and GBHS Tagger2 with 0.7.
– The third version of the algorithm involved combining the random initialization and the improved initialization of the harmonic memory, plus the local optimizer with the same values for the optimization parameter. This version was called GBHS Tagger3, for the purposes of the experiment, it was named: GBHS Tagger3 0.0, GBHS Tagger3 0.3, GBHS Tagger3 0.5 and GBHS Tagger3 0.7.

3.2.2 Metaheuristic Tagged Algorithm Approach

Three versions of HSTagger, a proposal of Forsati & Shamsfard (2010) [49] and (2015) [29], based on Harmony Search (HS) algorithm and that shows good results in comparison with other recognized taggers (HMM, ME and Brill's model taggers, among others), and it was selected for that reason.

– First algorithm is called HSTagger has a random initialization for harmony memory.
– Second algorithm is called HSTagger 2 which has been included an improved in the initialization using the Alpha parameter.
– Third algorithm is called HSTagger 3, which also involves the use of improved initialization and has been added a modification at the time of creating the improvise with the HCMR parameter, which uses the highest occurrences of each word in the harmony memory, which have been previously calculated.

3.2.3. Random Approach (base line)

– Random tagger that generates new solutions randomly for the tagging of the words in each sentence.

Table 1. Description of the texts Nasa Yuwe

Texts Nasa	Texts English	# sentences	# words
Nasa vxanxi's pta'sxnxi	The origin of man	12	136
kutxh wala ũpxhnxi yuwe	Corn origin	28	332
Jũth upxhnxi yuwe	History of sweet potato	14	163
Eçxthẽ' vxuu naamu'	Story of the devil	11	134
Ũ' tasx tuthenxi	Origin of food	16	245
Yu' vxaanxi yuwe	Origin of water	40	272
Wejxa yuwe	Origin of the wind	35	501
Kus	The night	19	172
Total		175	1955

Table 2. Tagset for Nasa Yuwe

Tagset for Nasa Yuwe	Frequencies	Probabilities
Predicative	661	33%
Qualifying	225	11.20%
Noun	641	32%
Connector	200	10%
Deictic	79	4%
Pronoun	20	1%
Punctuation	176	8.8%

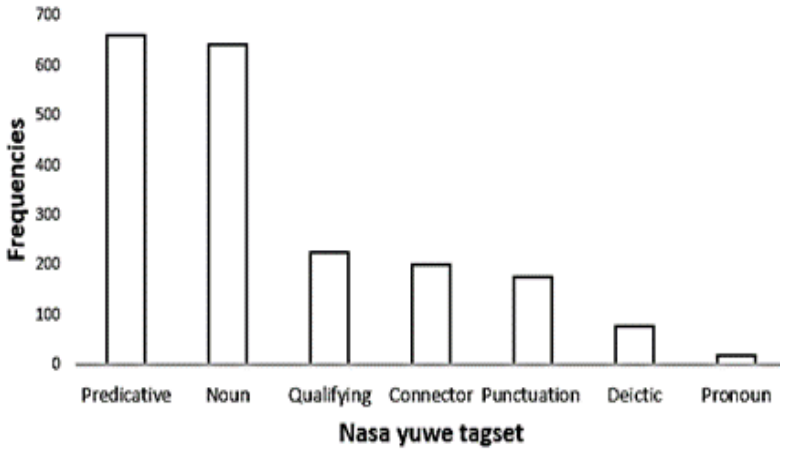

Fig. 1. Distribution of tags of Nasa Yuwe corpus

Table 3. Top ten most frequent words

Position	Word	Frequencies
1	Txãa	36
2	Wala	27
3	txã'w	24
4	sa'	23
5	teeçx	19
6	nawã	17
7	aça'	17
8	mẽh	15
9	aççxa	15
10	u'pu'	13

Table 4. Example of tagged phrases

# of sentence	Nasa words	Tag	Order
8	Naa	Deictic	1
8	seka'	Noun	2
8	nmẽh	Qualifying	3
8	Wala	Qualifying	4
8	açxasayũ'ne'	Qualifying	5
8	sa'	Connector	6
8	luuçxwe'sxyakh	Noun	7
8	wẽt	Qualifying	8
8	fxi'zeya'	Predicative	9
8	ãjamene' /ãhamene'	Qualifying	10

4 Nasa Yuwe Language Tagged Corpus

4.1 Data Set

As mentioned above, the sentences tagged in the Nasa Yuwe corpus were taken from 8 texts from the Nasa Yuwe test collection [3], the texts make references to popular stories of Nasa life and cosmovision, leaving the corpus conformed as presented in Table 1.

4.2 Results of the Tagging Process of Nasa Yuwe Corpus

4.2.1 Tagset for Nasa Yuwe

The tagset for Nasa Yuwe language used was that described in Sections 2, adding a tag for punctuation marks and a pronoun tag that was included by the linguist at the time of the review of the tagged corpus. In Table 2, the frequencies of each label in the Nasa Yuwe corpus can be seen and Fig. 1 shows the distribution of the tags in the

Table 5. Alignment of the tagset for Nasa Yuwe

Universal Tagset	Tagset for Nasa Yuwe	Frequency
Verb	Predicative	661
Adj	Qualifying	152
Adv	Qualifying/ Connector	212
Noun	Nouns	642
Num	Nouns / Qualifying	5
Det	Deictic	80
Pron	Pronoun / Connector	27
Conj	Connector	47
Prt	Not Applicable	-
Adp	Not Applicable	-
Punctuation	Punctuation	176
X	Other words	-

Table 6. Training and evaluation datasets in first experiment

Test data folder	Sentences in test data	Words on test data	Training data folders	Words on training data	Common words	Unknown words
1	18	197	2,3,4,5,6,7,8,9,10	1805	109	88 (44.67 %)
2	18	153	1,3,4,5,6,7,8,9,10	1849	86	67 (43.79 %)
3	18	179	1,2,4,5,6,7,8,9,10	1823	93	86 (48.04 %)
4	18	233	1,2,3,5,6,7,8,9,10	1769	113	120 (51.50 %)
5	18	229	1,2,3,4,6,7,8,9,10	1773	117	112 (48.91 %)
6	17	198	1,2,3,4,5,7,8,9,10	1804	102	96 (48.48 %)
7	17	249	1,2,3,4,5,6,8,9,10	1753	136	113 (45.38 %)
8	17	179	1,2,3,4,5,6,7,9,10	1823	98	81 (45.25 %)
9	17	194	1,2,3,4,5,6,7,8,10	1808	93	101 (52.06 %)
10	17	191	1,2,3,4, 5,6,7,8,9	1811	110	81 (42.41 %)

corpus, showing a high presence of predicative and nouns words.

4.2.2 Tagged Corpus for Nasa Yuwe

The tagged corpus for Nasa Yuwe is made up as follows:

1. Words and size. 1176 words, with a maximum length of 14 unified Nasa alphabet characters and a minimum of 1, with an average of 6. Table 3 presents the top ten most frequent words in the corpus.
2. Tagged phrases. 175 tagged sentences, with maximum length of 34 words per phrase and minimum length of 1 word.
3. Table 4 shows an example of the tagged phrases within the corpus, detailing the corresponding tag for every as well as the word order in the sentence.

Table 7. Results of running algorithms in the both experiments. Best results are showed in bold

Algorithms	Parameters (ProbOpt)	First Experiment (10 folds cross validation)		Second experiment (leave one out cross validation)	
		Precision (%)	Standard deviation	Precision (%)	Standard deviation
Random Tagger	-	53.862	3.427	57.7022	17.1942
HSTagger	-	57.294	3.395	60.1914	17.0776
HSTagger2	-	57.957	3.468	60.8964	17,3815
HSTagger3	-	50.893	3.585	53.7983	16.6512
GBHS Tagger	0.0	63.536	2.842	**66.5787**	**16.9290**
GBHS Tagger	0.3	62.529	2.701	66.4297	17.6616
GBHS Tagger	0.5	62.529	2.701	66.4297	17.6616
GBHS Tagger	0.7	62.529	2.701	66.4297	17.6616
GBHS Tagger 2	0.0	**63.867**	2.884	65.9432	16.9991
GBHS Tagger 2	0.3	63.783	3.035	**66.2706**	17.4027
GBHS Tagger 2	0.5	63.783	3.035	66.2706	17.4027
GBHS Tagger 2	0.7	63.783	3.3035	66.2706	17.4027
GBHS Tagger 3	0.0	63.614	**2.701**	65.9909	16.8131
GBHS Tagger 3	0.3	63.333	2.955	66.0765	17.4176
GBHS Tagger 3	0.5	63.333	2.955	66.0765	17.4176
GBHS Tagger 3	0.7	63.333	2.955	66.0765	17.4176

Table 8. Friedman ranking for both experiments

Algorithm	Ranking first experiment	Ranking second experiment
GBHS Tagger 2 con 0.0	**3.45**	7.0457
GBHS Tagger 2 con 0.3	4.55	**6.5857**
GBHS Tagger 2 con 0.5	4.55	**6.5857**
GBHS Tagger 2 con 0.7	4.55	**6.5857**
GBHS Tagger 3 con 0.0	4.9	7.3571
GBHS Tagger con 0.0	5.3	7.0657
GBHS Tagger 3 con 0.3	7.3	6.7086
GBHS Tagger 3 con 0.5	7.3	6.7086
GBHS Tagger 3 con 0.7	7.3	6.7086
GBHS Tagger con 0.3	9.6	7.8371
GBHS Tagger con 0.5	9.6	7.8371
GBHS Tagger con 0.7	9.6	7.8371
HSTagger 2	13.25	11.3857
HSTagger	13.75	11.7171
Azar	15	13.3657
HSTagger 3	16	14.6686

4. Table 5 shows the tagset alignment of Nasa Yuwe in relation to the Universal tagset [17]. This was not a simple process since in most cases it was necessary to re-tag, for example:

– Some words that were tagged as Noun (Nasa tagset) had to change to Noun and Num in the Universal tagset.
– With the words tagged Qualifying (Nasa tag), it was necessary to review them thoroughly to define what the corresponding tag was in the Universal tagging (Adv or Adj).

5. The tagging corpus for Nasa Yuwe was published online at link.

5. Experiments, Analyses and Comparisons

5.1 Experimental Setup

Two experiments were run. For the first experiment, the sentences of the Nasa Yuwe Corpus were divided into 10 folders, so that the tests could be performed using cross-validation, and the second experiment used the "leave one out" strategy. Table 6 shows the quantity of the sentences in each test and training data set, for the first experiment, that is, if the sentences of folder 1 are taken as test data, the training sentences are taken from folders 2 to 10 and so on for the other folder

The second experiment (leave one out) used one sentence as test data and the remaining sentences in the corpus as training data.

In all of the experiments, each algorithm was run 30 times over each sentence and its average precision values were calculated. For each algorithm, a maximum of 110 evaluations of the objective function was run for each sentence.

For the HSTagger and GBHS tagger algorithms the objective function was calculated as the probability of each word and its possible tags in the different sets of information, in the same manner as with the trigram probabilities [29, 50].

The measure used for the evaluation of the algorithms is presented in Eq. 1 [29]:

$$Precision = \frac{\#\ correctly\ tagged\ words}{\#\ words}. \quad (1)$$

The parameters used for HSTagger were defined according to its original paper: HMS = 20, HMCR = 0.65 and PAR = 0.25. The parameters used for GBHS tagger also were defined according to its original paper: HMS = 10, HMCR = 0.95, PARMin = 0.01, and PARMax = 0.99, Alpha =0.5.

5.2 Results

Table 7 shows the performance of the precision and standard deviation values for each of the algorithms evaluated for both experiments, where the best results are seen in the performance of the proposed GBHS tagger algorithm in all versions, especially GBHS tagger 2 without local optimizer for first experiment (k = 10 folds) and GBHS tagger 2 with local optimizer for second experiment.

The results presented in both experiments show significant improvements in the performance values of all tagger algorithms for experiment 2 in comparison with experiment 1. This increase indicates that the size of the corpus is relevant to the performance of the algorithms.

For both experiments, the Friedman non-parametric statistical test was applied for multiple comparison, to establish the differences between the algorithms. Table 8 shows the scores obtained (For first experiment P-value was: 5.7568E-11 and for second experiment P-value was: 2.6622E-10). It supports the conclusion that GBHS tagger outperforms the other algorithms.

Additionally, for both experiment the Wilcoxon test was performed, and the results showed that with 90% of confidence, GBHS Tagger in all its versions improves on the results of the other algorithms.

6. Conclusion and Future Work

The scope of this work can be expressed in two main outcomes. Firstly, a synthesis of the process of building a tagged corpus is carried out, through analysis and review of similar works. Such a process involves tagset definition. The analysis presented here highlights the characteristics of an independent language such as Nasa Yuwe, which is still in the process of description. This corpus therefore constitutes an important contribution for future work regarding both this particular language

as well as other languages that are in danger of extinction and have not been matter of study for natural language processing investigations.

Secondly, two experiments were conducted aimed at using the Nasa Yuwe language tagged corpus to select which POS Tagger is the best with this corpus. In these experiments, three tagger algorithms were used, namely: Random tagger, three versions of taggers that used a metaheuristic approach (HSTAGger proposed by Forsati, et al in previous work [49, 54, 29]) and three versions of a memetic tagger algorithm GBHS tagger [50]. The GBHS tagger is based on Global-Best Harmony Search algorithm, Hill Climbing, and an explicit Tabu memory, which outperforms the other methods considered. This fact can be attributed to the hybrid nature of GBHS since it uses Harmony Search with Particle Swarm Optimization, together with the use of explicit Tabu memory that prevents the algorithm from being trapped in local optima as well as avoids over-exploitation of areas of the solution space.

Future work will focus on two key aspects: 1) improve GBHS tagger for identifying parts of speech for the Nasa Yuwe language, aiming at increasing precision values. To do this, both analysis of the different methods used for building a tagger (e.g. statistical techniques, among others), and definition of a strategy to identify and assign the most likely tag for each word in a sentence must be carried out. 2) enrich the tagging corpus for Nasa Yuwe both in size and in the tagset used to increase the accuracy of the tagging process.

Acknowledgements

Sierra, Cobos, Corrales and Rojas are grateful to the University of Cauca and its research groups GTI, GIT and GELPS in the Computer Science, Telematics, and Anthropology departments. The vice-President of Research, Enrique Herrera-Viedma would like to acknowledge the University of Granada and its SECABA LAB research group and its Computer Science and Artificial Intelligence departments. Peluffo is grateful to the University Yachay Tech and Corporación Universitaria Autónoma de Nariño. We are especially grateful to Luis Dicue, Professor of the Nasa community for his collaboration in this research and to Colin McLachlan for suggestions relating to the English text.

References

1. **Rojas, T. E. (2012).** Esbozo Gramatical de la lengua nasa (lengua Paéz). In: UNICEF (Ed.), *El Lenguaje en Colombia, Tomo I: Realidad Lingüística de Colombia*. Bogotá: Academia Colombiana de la Lengua e Instituto Caro y Cuervo.
2. **Sierra-Martínez, L., Cobos, C., & Corrales, J. (2016).** Tokenizer adapted for Nasa Yuwe language. *Computación y Sistemas*, Vol. 20, No. 3, pp. 335–365. DOI: 10.13053/CyS-20-3-2455.
3. **Sierra-Martínez, L. M., Cobos-Lozada, C. A., Corrales, J. C., & Rojas-Curieux, T. (2015).** Building a Nasa Yuwe Test Collection. *Processing Computational Linguistics and Intelligent Text,* Vol. 9041, pp. 112–123. DOI: 10.1007/978-3-319-18111-0_9.
4. **Attia, M., Rashwan, M., & Al-Badrashiny, M. (2009).** Fassieh (R), a Semi-Automatic Visual Interactive Tool for Morphological, PoS-Tags, Phonetic, and Semantic Annotation of Arabic Text Corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, No. 5, pp. 916–925. DOI: 10.1109/TASL.2009.2019298.
5. **Baeza-Yates, R. & Ribeiro-Neto, B. (1999).** *Modern Information Retrieval.* Pearson-Addison Wesley.
6. **Instituto Colombiano de Cultura Hispánica. (2008).** Geografía Humana de Colombia. *Región Andina Central,* Vol. IV.
7. **Rivet, P. (1913).** Les familles linguistiques du Nord-Ouest de l'Amérique du Sud en Année Linguistique (Société Philologique). *Journal de la Société des Américanistes,* Vol. 10, No. 1, pp. 117–154.
8. **Greenberg, J. (1987).** Language in the Americas. *Stanford University Press.*
9. **Loukotka, C. (1968).** *Classification of South American Indian Languages*. Latin American Studies Center, University of California.
10. **Constenla, A. (1993).** *La Familia Chibcha en Estado Actual de la Clasificación de las Lenguas Indígenas de Colombia,* pp. 75–125.
11. **Landaburu, J. (2000).** Clasificación de las lenguas indígenas de Colombia*. Lenguas Indígenas de Colombia: una visión descriptiva*, Santafé de Bogotá, pp. 25–48.

12. **Jung, I. (1984).** Gramática del Páez o nasa yuwe. *Descripción de una Lengua Indígena de Colombia*, Published by LINOM GmbH 2008.
13. **Rojas, T. (1998).** *La Lengua páez.* Bogotá: Ministerio de Cultura.
14. **Xiao, R. (2010).** Creation Corpus. *Handbook of Natural Language Processing CRC Press,* pp. 147–166.
15. **Dinakaramani, A., Rashel, F., Luthfi, A., & Manurung, R. (2014).** Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus. *International Conference on Asian Language Processing (IALP´14),* pp. 66–69. DOI: 10.1109/IALP.2014.6973519.
16. **Ismail, S., Rahman, M., & Al-Mumin, M. (2014).** Developing an Automated Bangla Parts Of Speech. *16th International Conference on Computer and Information Technology (ICCIT)* Khulna: IEEE, pp. 355–359. DOI: 10.1109/ICCITechn.2014.6997347.
17. **Petrov, S., Das, D., & McDonald, R. (2012).** A Universal Part-of-Speech Tagset. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*.
18. **Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., Choudhury, M., Nath Jha, G., & KVS Subbarao. (2008).** *A Common Parts-of-Speech Tagset Framework for Indian Languages. Proceedings of (LREC´08)*, pp. 1331–1337.
19. **Expert Advisory Group on Language Engineering Standards. (1996).** EAGLES *Recommendations for the Morphosyntatic Annotation of Corpora*.
20. **Rabbi, I., Abid-Khan, M., & Ali, R. (2008).** Developing a Tagset for Pashto Part of Speech Tagging. *Second International Conference on Electrical Engineering*. DOI: 10.1109/ICEE.2008. 4553909.
21. **Scherrer, Y., Nerima, L., Russo, L., Ivanova, M., & Wehrli, E. (2014).** SwissAdmin: A multilingual tagged parallel corpus of press releases. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC´14).
22. **Ariaratnam, I., Weerasinghe, A., & Liyanage, C. (2014).** A shallow parser for Tamil. *International Conference on Advances in ICT for Emerging Regions (ICTer),* pp.197–203. DOI: 10.1109/ICTER. 2014.7083901.
23. **Singh, S. & Banerjee, E. (2014).** Annotating Bhojpuri Corpus using BIS Scheme. *Proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE-2), Ninth International Conference on Language Resources and Evaluation (*LREC´14).
24. **Spoustová, J. & Spousta, M. (2012).** A High-Quality Web Corpus of Czech. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
25. **Ahmed-Mahar, J. & Qadir-Memon, G. (2010).** Rule Based Part of Speech Tagging of Shindi Language. *International Conference on Signal Acquisition,* pp. 101–106. DOI: 10.1109/ICSAP. 2010.27.
26. **Koehn, P. (2005).** Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of the Tenth Machine Translation Summit (MT Summit XX)*, pp. 79–86.
27. **Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993).** Building a large annotated corpus of English: the penn treebank. *Journal Computational Linguistics - Special issue on using large corpora*: II, Vol. 19, No. 2, pp. 313–330.
28. **Francis, W. & Kucera, H. (1979).** *Brown Corpus.* http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM#bc8.
29. **Forsati, R. & Shamsfard, M. (2015).** Novel harmony search-based algorithms for part-of-speech tagging. *Knowledge and Information Systems,* Vol. 42, No. 3, pp. 709–736. DOI: 10.1007/s10115-013-0719-6.
30. **Brill, E. (1992).** A simple rule-based part of speech tagger. *Proceedings of the third conference on Applied natural language processing (ANLC'92)*, Association for Computational Linguistics, pp. 152–155. DOI:10.3115/974499.974526.
31. **Brill, E. (1995).** Transformation-based error-driven learning and natural language processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, Vol. 21, No. 4, pp. 543–565.
32. **AlSuhaibani, R., Newman, C., Collard, M., & Maletic, J. (2015).** Heuristic-Based Part-of-Speech Tagging of Source Code Identifiers and Comments. *IEEE 5th Workshop on Mining Unstructured Data (MUD*) pp. 1–6. DOI: 10.1109/MUD.2015.7327960.
33. **Mall, S. & Jaiswal, U. (2015).** Innovative Algorithms for Parts of Speech Tagging in Hindi-English Machine. *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference,* pp. 709 – 714. DOI: 10.1109/ICGCIoT.2015.7380555
34. **Alba, E., Luque, G., & Araujo, L. (2006).** Natural language tagging with genetic algorithms. *Information Processing Letters,* Vol. 100, No. 5, pp. 173–182. DOI: 10.1016/j.ipl.2006.07.002.
35. **Brants, T. (2000).** TnT - a statistical part-of-speech tagger. *Proceedings of the sixth conference on*

Applied natural language processing (ANLC'00), Association for Computational Linguistics, pp. 224–231. DOI: 10.3115/974147.974178.

36. **Lafferty, J., McCallum, A., & Pereira, F. C. (2001).** Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning,* pp. 282–289.
37. **Keyaki, A. & Miyazaki, J. (2017).** Part-of-speech tagging for web search queries using a large-scale web corpus. *Proceedings of the Symposium on Applied Computing*, pp. 931–937. DOI: 10.1145/3019612.3019694.
38. **Zhonglin, Y., Zhen, J., Huang, J., & Hongfeng, Y. (2016).** Part-of-Speech Tagging based on Dictionary and Statistical Machine Learning. *Proceedings of the 35th Chinese Control Conference* (CCC). DOI:10.1109/ChiCC.2016.7554459.
39. **Albared, M., Al-Moslmi, T., Omar, N., Al-Shabi, A., & Ba-Alwi, F. M. (2016).** Probabilistic Arabic Part of Speech Tagger with Unknown Words Handling. *Journal of Theoretical and Applied Information Technology*, Vol. 90, No. 2, pp. 236–246.
40. **Sun, W. & Wan, X. (2016).** Towards Accurate and Efficient Chinese Part-of-Speech Tagging. *Computational Linguistics*, Vol. 42, No. 3, pp. 391–419. DOI: 10.1162/COLI a 00253.
41. **Schmid, H. (1994).** Part-of-speech tagging with neural networks. *Proceedings of the 15th conference on computational linguistics*. *Association for Computational Linguistics*. pp. 172–176. DOI: 10.3115/991886.991915.
42. **Nakamura, M., & Shikano, K. (1989).** A study of English word category prediction based on neutral networks, Acoustics, Speech, and Signal. *Processing International Conference on Acoustics, Speech, and Signal Processing IEEE,* Vol. 2, pp. 731–734.
43. **Hnin, H., Pa-Pa, W., & Thu, Y. (2017).** Back-Propagation Neural Network Approach to Myanmar Part-of-Speech Tagging. *In Advances in Intelligent Systems and Computing,* pp. 212–220. DOI: 10.1007/978-3-319-48490-7_25.
44. **Kabir, F., Abdullah-Al-Mamun, K., & Nurul Huda, M. (2016).** Deep learning based parts of speech tagger for Bengali. *5th International Conference on Informatics, Electronics and Vision* (ICIEV). DOI:10.1109/ICIEV.2016.7760098.
45. **Carneiro, H., França, F. M., & Lima, P. M. (2015).** Multilingual part-of-speech tagging with weightless neural networks. *Neural Networks*, Vol. 66, pp. 11–21. DOI: 1016/j.neunet.2015.02.012.
46. **Lv, C., Liu, H., Dong, Y., Li, F., & Liang, Y. (2017).** Using Uniform-Design GEP for Part-of-Speech Tagging. *Journal of Circuits, Systems and Computers*, Vol. 26, No. 4, pp. 1–14. DOI: 10.1142/S0218126617500608.
47. **Forsati, R. & Shamsfard, M. (2012).** Cooperation of Evolutionary and Statistical PoS-tagging. *16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP),* pp. 446–451. DOI: 10.1109/AISP.2012.6313789.
48. **Silva, A. P., Silva, A., & Rodríguez, I. (2014).** Part-of-Speech Tagging Using Evolutionary Computation. *Nature Inspired Cooperative Strategies for Optimization,* Vol. 512, pp. 167–178. DOI: 10.1007/978-3-319-01692-4_13.
49. **Forsati, R., Shamsfard, M., & Mojtahedpour, P. (2010).** An Efficient Meta Heuristic Algorithm for POS-Tagging. *Fifth International Multi-Conference on Computing in the Global Information Technology (ICCGI),* pp. 93–98. DOI: 10.1109/ICCGI.2010.42.
50. **Sierra-Martínez, L. M., Cobos, C., & Corrales, J. C. (2017).** Memetic Algorithm Based on Global-Best Harmony Search and Hill Climbing for Part of Speech Tagging. **In A. Ghosh, R. Pal, & R. Prasath (Ed.)**, The Fifth International Conference on Mining Intelligence and Knowledge Exploration. *Lecture Notes in Computer Science,* Vol. 10682, pp. 198–211. DOI: 10.1007/978-3-319-71928-3_20.
51. **Pratt, K. S. (2009).** *Design Patterns for Research Methods: Iterative Field Research.*
52. **Omran, M. & Mahdavi, M. (2008).** Global Best Harmony Search. *Applied Mathematics and Computation*, Vol. 198, No. 2, pp. 643–656. DOI: 1016/j.amc.2007.09.004.
53. **Eberhart, R. & Kennedy, J. (1995).** A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micromachine and Human Science*, pp. 39–43. DOI: 10.1109/MHS.1995.494215.
54. **Forsati, R. & Shamsfard, M. (2014).** Hybrid PoS-tagging: A cooperation of evolutionary and statistical approaches. *Applied Mathematical Modelling*, Vol. 38, No. 13, pp. 3193–3211. DOI: 10.1016/j.apm.2013.11.047.

Article received on 12/12/2017; accepted on 20/02/2018.
Corresponding author is Luz Marina Sierra Martínez.