COMPUTATIONAL ENGINEERING

Part-of-speech tagging with maximum entropy and distributional similarity features in a subregional corpus of Spanish

Etiquetado gramatical por entropía máxima y rasgos de similitud distribucional en un corpus subregional del español

Antonio Rico-Sulayes[1§], Rafael Saldívar-Arreola[2], Álvaro Rábago-Tánori[3]

[1]*Grupo de Investigación en Lingüística Aplicada, Universidad de las Américas Puebla. Puebla, México.*
[2]*Cuerpo Académico Lengua, Tecnología e Innovación, Universidad Autónoma de Baja California. Mexicali, México.*
[3]*Cuerpo Académico Lengua, Tecnología e Innovación, Universidad Autónoma de Baja California. Ensenada, México.*
antonio.rico@udlap.mx, rafaelsaldivar@uabc.edu.mx, rabago@uabc.edu.mx.

(Recibido: Junio 15 de 2016 – Aceptado: Diciembre 01 de 2016)

Abstract
The present research study has used two state-of-the-art Spanish taggers with the primary goal of automatically tagging for POS a strictly assembled collection of unstructured text aimed at assisting a number of linguistic tasks, the subregional Mexican *Corpus del Habla de Baja California* (*CHBC*). These taggers, a Maximum-Entropy-based one and another one that adds to this statistical construct distributional similarity features, have recently been released but were missing an accuracy rate. Therefore, the second goal of this article is to evaluate and provide attested accuracy figures for the language models behind these taggers. In order to achieve these two goals, this article has proposed a novel, reduced tag set, which has also been proven useful for the goals here pursued. On a sample of almost 11,000 words and more than 12,500 tags for two genres (written text and transcribed oral speech), the Maximum Entropy tagger and the tagger with Maximum Entropy plus distributional similarity features have achieved results of 97.2% and 97.4%, respectively. By comparing these figures to a human ceiling or gold standard of 97.1%, also attested here, it is clear that the results of both taggers are competitive even when applied to an external data collection for which they have not been previously trained or tuned for. This is particularly important because under these kinds of experimental conditions taggers performance has been shown to deteriorate.

Keywords: *Mexican Spanish, stochastic POS tagging, tagged corpus.*

Resumen
Con el objetivo primario de etiquetar automáticamente las categorías gramaticales en una colección de texto no estructurado, la cual fue diseñada para asistir en una serie de tareas lingüísticas, esta investigación ha utilizado dos etiquetadores automáticos de primera generación para el español. Estos etiquetadores han sido aplicados al *Corpus del Habla de Baja California* (*CHBC*) que cubre una subregión de México. Los dos etiquetadores, uno basado en el principio de Máxima Entropía y el otro que le suma a este modelo estadístico rasgos de similitud distribucional, son de reciente introducción y no se ha ofrecido un rango de precisión para los mismos. Por tanto, este artículo ha tenido como segundo objetivo el evaluar y proveer una cifra de precisión comprobada para los modelos de lenguaje que subyacen a los etiquetadores en cuestión. Con la finalidad de lograr estos dos objetivos, este artículo ha propuesto un etiquetario reducido, el cual también ha resultado de utilidad en la búsqueda de estos objetivos. Aplicados a una muestra de alrededor de 11,000 palabras y más de 12,500 etiquetas gramaticales para dos géneros (texto escrito y discurso oral transcrito), los dos etiquetadores, el de Máxima Entropía y el que suma a ésta los rasgos de similitud distribucional, han obtenido resultados de 97.2% y 97.4%, respectivamente. Al comparar estas cifras con el criterio estándar de 97.1% obtenido entre anotadores humanos, los resultados de ambos etiquetadores se muestran competitivos, incluso al aplicarlos a una colección de datos externa para la cual no han sido previamente entrenados o calibrados. Esto es particularmente importante porque en este tipo de condiciones experimentales se ha encontrado que el desempeño de los etiquetadores puede deteriorarse.

Palabras clave: *Corpus etiquetado, español mexicano, etiquetado gramatical estocástico.*

1. Introduction

Criterion-based, carefully assembled repositories of unstructured text, also called corpora, have been used for over 50 years in a number of computational language-related tasks. Among these tasks, a basic one that has a great impact on a number of computer applications is counting word n-grams, or word sequences. This action can be aimed at building statistical models in order to compute probabilities for a number of language elements, which include not only words and word n-grams themselves, but other elements such as characters, sounds, meanings, and their combinations. Computing language element probabilities has a wide range of applications, from speech recognition, to machine translation, to spell checking, among many other computational language-related tasks (1). In the context of language model construction, corpora can have three clearly different roles. First, corpora may be originally assembled to train and then test these kinds of statistical models to evaluate their efficiency in predicting language elements. In this case, the testing of such models has to follow strict protocols, such as cross-validation, which arranges corpus data so they are not simultaneously used in the training and testing processes (2). Through cross-validation, or other similar processes such as leave-one-out cross validation (3), the testing of these probabilistic models gains statistical significance and their performance with new data can be better estimated. Secondly, corpora may be produced with goals different from language model training. These corpora, however, may be used eventually as external resources to evaluate the performance of statistical language models. For this purpose, they represent an optimal test bed because they are usually assembled following strict criteria guided by specific research goals. Thirdly, corpora that are produced with goals external to language modelling may themselves become an end application of probabilistic, computational language applications, such as part-of-speech (POS) tagging, word sense disambiguation, authorship attribution, among several others (1).

In this research study, we have used a subregional corpus of Spanish from the northwestern region of Mexico that contains spoken and written samples of language. Denis & Sagot (4) state that a potential advantage of using this kind of data is to attain a better handling of unknown words. Along with this potential benefit, there are at least two purposes in using this corpus: exploiting it as an external resource to evaluate a language model application and improving it as an end application in itself through the use of this probabilistic language model. In order to accomplish the first objective, this corpus was used as a test bed to evaluate a pair of state-of-the-art taggers in Spanish. Added less than two years ago to a broader suite of taggers for five other languages (English, Chinese, Arabic, French, and German), the two Spanish language taggers here tested have not been provided with an accuracy rate by its developing team, the Stanford Natural Language Processing SNLP Group (5). This is true for both, the documentation provided with the very last version publicly available of the tagger suite, 3.6.0 (released just five months before the moment when this article is being written), and the specific information provided online for the two Spanish taggers (6). In addition to the lack of an evaluation measurement provided by the developing team, this study has conducted an external evaluation of the aforementioned taggers without prior training and tuning to the corpus here used, which is an important contribution for the comparison of taggers performance as it explores their portability (7). As for the second objective of improving the corpus itself, the corpus used here as a test bed was not assembled for the purpose of this evaluation, so tagging it for its POS had other additional goals. Among these, an immediate objective is to refine its retrieval capabilities for dictionary/glossary making, a common application of corpora (8). This is particularly important because the corpus, the tagging of which has obtained very high accuracy figures in this research study, should be the first subregional Mexican corpus (for Northwestern Mexican Spanish) fully tagged for POS. Only one developing team in México, the one behind *COCIEM* (*Corpus Básico Científico*

del Español de México), claims to have carried out POS tagging to develop their corpus, yet this corpus is not openly available for consultation in the team's website (9), and it is not a subregional but a genre-specific collection composed of 92 science textbooks. It is also worth mentioning that the organizing team of *PRESEEA*, a project that comprises several sub-regional, city-specific corpora of Spanish, does not specify in their methodology the need for POS tagging (10). Thus, the present article has two clear contributions: 1) filling the gap in the testing and evaluation of two state-of-the-art POS taggers recently released for Spanish, and 2) producing a highly accurate improvement and a valuable new feature for a useful language tool that, once fully tagged, will be unique in its kind.

2. Methodology

The need to identify POS classes for all the words in corpora derives from the information that these categories can provide regarding the words they classify and those surrounding these classified words. For example, POS classes, also known as lexical categories, enable a computer system to determine the meaning of a word with several senses. Knowing that the word *can* is a verb, or modal verb, allows a system to determine that its meaning is *having the ability, permission, or probability* rather than *a metal, sealed container*, or some other meaning for the noun function of this same word form. Similarly, the POS of a word can help a system identify the function of the surrounding words. Knowing that *can* is a modal verb allows a computer system to realize that: 1) the following word should be a verb, 2) this verb should be a lexical verb that contains the main meaning of the verb combination, 3) there may be some preceding word that is a noun – in some languages like English the presence of this word is mandatory, and 4) this noun could be the subject of the aforementioned verb combination. All this information is useful in improving a number of computational applications, such as a speech recognition engine that has to decide what a poorly recorded word is in some context where the surrounding words have been accurately detected, or an ontological database interface that has to identify the subject and predicate roles in a question to retrieve the correct answer from the knowledge resources it contains.

2.1 Automatic POS tagging in spanish

In order to automatically assign POS classes to words, taggers have followed various building principles, but the two most common ones, according to Jurafsky & Martin (1), are rule-based and stochastic taggers. Rule-based taggers use long lists of hand-coded rule definitions that are aimed at disambiguating word classes in contexts where their lexical categories are hard to discriminate. An early example of rule-based architectures are taggers based on Constraint Grammars (11), with over three decades of their first, but still popular, proposal. Stochastic taggers, on the contrary, use previously POS-annotated corpora to learn probabilities for these tags. Early implementations of Spanish taggers show this general division with rule-based taggers like the proposed by Farwell et al. (12) and stochastic models such as Schmid's (13).

Regarding stochastic taggers, there have been different popular models for these probability-based taggers, with the two most popular being Hidden Markov Models (HMM) and Maximum Entropy (MaxEnt) models, as Jurafsky & Martin (1) point out. While HMMs were developed in the mid-sixties, the abstraction behind MaxEnt was first applied to language issues twenty years ago (14). Early stochastic taggers for Spanish, like Schmid's (13), are HMMs implementations. Schmid's tagger is a tree-based tagger that calculates tag probabilities using a Viterbi algorithm, which relies on HMMs (1).

Finally, there are also taggers that follow different architectures, and even share characteristics from the two aforementioned building principles, like the Brill tagger (15). In this sense, Farwell et al.'s tagger (12) for Spanish is a moderately hybrid model that uses some general probabilistic

information, such as the most common POS for frequent words. It should be noted that all the tagger architectures mentioned are still being applied and their novel application to different languages and data collections continues.

As previously noted, of the two most popular families of stochastic taggers, HMMs and MaxEnt, the former has a longer history in its application to natural language processing. This is also reflected in the abundant literature that uses Spanish HMMs-based taggers (e.g., 16-19), such as Schmid's (13) and Padró (20). In contrast, with a more recent history in language applications in general, MaxEnt models have been applied to Spanish more lately. The two Stanford NLP Group Spanish taggers that are evaluated in this study are both derived from recent MaxEnt implementations (5).

The difference between the two taggers here evaluated is that, while one is a plain MaxEnt tagger (21), the other one adds features derived from the computation of distributional similarities, exploited in language applications in recent years (22). Beyond its later exploitation in computational linguistics tasks, the relevance of exploring the performance of a distributional similarity-based POS tagger derives from the reported improvement that this information can bring into tagging results. In an early report of accuracy improvement between standard HMMs taggers and taggers augmented with distributional similarity information, Wang & Schuurmans (23) show an increase of accuracy from 81.32% for the former models to 90.03% for the latter. More recent reports of improvement from HMMs-based to distributional similarity-augmented taggers are Bienmann et al. (24), with a reported accuracy increase from 93.4% to 95.3%, Bienmann (25), with an increase from 93.05% to 97.33%, Bienmann & Riedl (26), from 95.28% to 96.07%, and Datla et al. (27), who report improvements in all 6 individual POS tags they explore. Adding to the relevance of the present article, none of the previously mentioned studies has reported any results for a distributional similarity-augmented tagger in Spanish.

Finally, it should also be mentioned that the suite of taggers that the two Spanish taggers here evaluated belong to was first released for English in 2004 and has been adding extensions since then. In more recent years it has added improved models for Arabic and German in 2011, a first model for French and an improved model for Chinese in 2012, and a first model for Spanish (the last language added) in August 2014 (5). The very last full version for the suite, 3.6.0, was released near the end of 2015. This last release is the one being used and tested in this study.

2.2 A simple POS tag set for Spanish

The elements that were actually compared and evaluated in this research study are the tags or labels assigned to words by the two Spanish taggers tested. In this respect, these taggers were trained using the grammatically annotated AnCora (28) corpus. This corpus uses the *Expert Advisory Group on Language Engineering Standards* (*EAGLES*) recommendations (29) for a tag set in Spanish, which results in a total of 227 different tags. From this large number of tags, the implementation of the two Stanford Spanish taggers uses a reduced version of 85 tags (6). For the evaluation presented in this article, a further reduction of the tag set was performed by using only the first or first two characters in each tag. This second reduction results in a list of 27 tags, shown in Table 1. As can be seen in this table, the final reduced set includes tags only for the most general lexical categories, with subcategories for only six of them (conjunctions, determiners, nouns, numerals, pronouns, and verbs).

Table 1. *Reduced tag set used for evaluation.*

POS with subcategories			POS with single tags	
POS	Subcategory	Tag	POS	Tag
conjunctions	Coordinating	CC	abbreviations	Y
	Subordinating	CS	adjectives	AQ
determiners	Demonstrative	DD	adverbs	RG
	Possessive	DP	interjections	I

	Interrogative	DT	prepositions	SP
	Exclamation	DE	punctuation	F
	Article	DA	date	W
nouns	Common	NC	numeral	Z
	Personal	NP		
numerals	Cardinal	MC		
	Ordinal	MO		
pronouns	Personal	PP		
	Demonstrative	PD		
	Possessive	PX		
	Interrogative	PT		
	Relative	PR		
	Indefinite	PI		
verbs	Main	VM		
	Auxiliary	VA		

The reduction of tags presented in Table 1 was guided by three main goals. Firstly, for the immediate purpose of supporting lexicographic work, the grammatical nuances of much larger and more detailed tag sets were considered not essential in many cases. For example, in order to distinguish the preposition meaning of the Spanish word form *como*, translated as *like* in English, from its verbal meaning *I eat*, it is not necessary to know that the latter is a first person, singular, present indicative conjugated form. This kind of morphological and syntactical information is captured by the *EAGLES* tag set. Second, the reduction proposed here substantially simplified an independent manual tagging performed by humans, the product of which became the gold standard, i.e. the basis for the evaluation of the automatic taggers (1). Finally, after a simple conversion of the 85 tags used by Stanford NLP Group, the reduction also simplified the partially manual comparison of the human-generated tags against the automatically assigned tags.

2.3 Data

The two evaluated taggers were applied to data from the *Corpus del Habla de Baja California* (*CHBC*), which is currently under construction. This corpus is aimed at being representative of the Mexican state of Baja California, in the northwestern extreme of the country, covering its five municipalities or counties: Mexicali, Tijuana, Tecate, Rosarito and Ensenada. In terms of the corpus size, the developing team's final goal is to collect between five and six million words for two registers: written text and oral speech. Currently, over three million words have already been collected for six text genres, which are subdivided into 77 subgenres. As to the spoken part, 108 interviews have been recorded so far and are being transcribed. These interviews, which have around ten thousand words each, have been distributed among the five municipalities of the state following the population distributional patterns described by the Instituto Nacional de Estadística y Geografía (INEGI) (30). Therefore, the larger concentration of the oral speech samples comes from the two main urban areas in the state, Tijuana and Mexicali, with over one million inhabitants each.

As a subregional corpus, the *CHBC* is not the only one being constructed in Mexico. There are two teams that have made the most consolidated and advanced subregional corpus projects and have made their results publicly available in two corpora: one for central Mexico, *Corpus Sociolingüístico de la Ciudad de México* (31, 32), and the other for the northeastern part of the country, *Corpus del Habla de Monterrey* (33). It is worth noticing, however, that none of the Mexican subregional corpus projects aforementioned have released any grammatically annotated results (10, 31-33), neither have done other important regional projects undertaken in Mexico (e.g., 34, 35). Taking into account the advantages of interacting with a tagged corpus, which were mentioned in the previous section, this is an important missing characteristic. As noted above, the team that developed *COCIEM* is the only that claims to have carried out POS tagging to develop their corpus (2016). In this sense, this study contribution is clearly relevant to the various subregional corpus building projects not only in Mexico, but in several other Latin American countries that have similar ventures under

the *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (*PRESEEA*) (36).

Since the *CHBC* main division is the two registers that have to do with the language channel, written or oral, a sample was selected from both registers. The sample consisted of eleven, approximately 500-word excerpts for each register. The total size of the sample for the written register was 5,079 words and 5,665 words for the oral register, as shown in Table 2.

Table 2. *Data sample for taggers evaluation.*

Sample	Written genre	Spoken genre	Total
# of words	5079	5665	10744
# of tags	5548	6999	12547

Table 2 also shows the resulting total number of tags that were compared and evaluated for each genre. Since there are linguistic elements that are split for their annotation, such as punctuation marks that are separated from the words they are usually attached to, and are given their own tags, the final number of tags is slightly greater than the number of words. In the written genre the final number of tags was 5,548, compared to 5,079 words. The difference was even more significant in the spoken genre where there were 6,999 tags, and 5,665 words. This larger difference was due to the nature of spoken language and the extra prosodic information that is usually encoded in its transcription.

Finally, it should be mentioned that the use of external data collections, non-related to the training data assembled by a tagger developing team, represents an important contribution in the evaluation of taggers algorithms. The contribution of taggers external evaluation becomes obvious with research studies like Parra-Escartín & Martínez-Alonso (7). Early Spanish taggers such as Farwell et al.'s (12) and Schmid's (13) report accuracy results of 95.44% and 96.36%, respectively. Lately developed taggers for this language report accuracies of up to 96%-98% (7). However, besides the fact that some of the recently developed taggers are not openly available, many which are such as Schmid (13), Martínez et al. (16), Padró & Stanilovsky (17) and Agerri et al. (37) have been reported to obtain results that are much lower than the previously mentioned accuracy range when they are applied to new data (7). After testing these four taggers on data for which they have not been previously trained and tuned for, Parra-Escartín & Martínez-Alonso report accuracy results of 86%, 85%, 88%, and 87%, respectively for these taggers. The differences in accuracy results observed between in-house and external evaluations represent a clear motivation to conduct research such as Parra-Escartín & Martínez-Alonso's and the present study.

3. Results and discussion

The eleven excerpts sampled for each of the two registers (written and spoken) were annotated by a group of eleven human taggers using the reduced 27-tag set formerly presented in Table 1. After a first annotation, a different human tagger evaluated the tags assigned by the first annotator in order to obtain an upper-bound or human ceiling (1). A human ceiling is simply an evaluation of how well humans perform in the task for which the statistical language model is being tested. Obtaining this ceiling also allowed the team to refine the first annotation and reach a gold standard to compare against the 22 tagged excerpts (eleven per genre) rendered by each of the two taggers. As mentioned above, these two taggers were a MaxEnt tagger and a MaxEnt tagger plus distributional similarities (MaxEnt + DS) (5). The human ceiling or agreement for the 12,547 total tags in all sampled text is shown in the first row of Table 3. As can be seen in this table, the human ceiling was 97% for written text and 97.1% for the transcribed oral speech. This renders an overall average of 97.1% human agreement for the whole sample, which is consistent with former research findings about a human ceiling obtained without human interaction and discussion to reach a final consensus, as explained in Jurafsky & Martin (1).

Table 3. *Upper-bound / human ceiling and accuracy for data sampled.*

Evaluation	Written genre	Spoken genre	Total
% human ceiling	97.0	97.1	97.1
% accuracy MaxEnt	97.7	96.8	97.2
% accuracy MaxEnt + DS	97.7	97.0	97.4

The evaluation of the two taggers is also shown in Table 3. As generally understood in the evaluation of POS tagging (1), this table expresses accuracy as the percentage of the coincidence between the 12,547 tags generated by the automatic taggers and the tags in the gold standard produced by human annotators. For both taggers the accuracy is marginally better for the written genre than it is for transcribed spoken language. The MaxEnt tagger improves its accuracy from 96.8% with the spoken genre to 97.7% with the written text sample, while the MaxEnt + DS tagger improves from 97% to 97.7%. Also, the overall performance of the MaxEnt + DS tagger, 97.4%, is slightly better than the overall performance of the other tagger, 97.2%. This improvement is mostly derived from the marginally better performance that MaxEnt + DS tagger achieved with oral speech data, 97%, compared to the 96.8% accuracy by the MaxEnt tagger with this same data.

From the above mentioned figures, it is possible to make some general interpretations about the performance of the recently released Spanish taggers in the Stanford NLP Group tagger suite. Firstly, both taggers, MaxEnt and MaxEnt + DS, perform with roughly the same level of accuracy. There is a difference of just 0.2% between the overall performances of both taggers. Secondly, despite the fact that the overall accuracy figure for both taggers is very similar, the MaxEnt + DS tagger did perform better that the non-augmented tagger, as expected from the literature (23- 27). Thirdly, the accuracy obtained for both taggers, 97.2%-97.4%, is at the upper end of the commonly reported 96%-97% accuracy range for taggers in general (1) and the 96%-98% accuracy range reported for taggers in Spanish (7). Fourthly, since the overall human ceiling, 97.1%, is slightly below the overall performance of both taggers, 0.1-0.3% lower, this means that these taggers performance is very close to optimal. This is true because reducing the current error rate, which is below 3%, would require a language model that computes probabilities for language elements that humans do not easily agree on either. Therefore, lowering this error rate to zero would imply modeling noise, as Berthelsen & Megyesi (38) and Jurafsky & Martin (1) suggest. Finally, the competitive results obtained by the two taggers have been achieved in a data collection for which they have not been previously trained or tuned for. This is a particularly important feature as it argues for the portability of the tagging models.

One last piece of information to share is the accuracy results for each of the 27 POS tags in our proposed tag set. Table 4 shows the error rate for individual POS by the MaxEnt tagger when applied to the written genre and Table 5 shows the corresponding figures for the MaxEnt + DS tagger. Therefore, these tables show the results for the data set in which both taggers performed the best. In this respect, as it can be seen by comparing the rightmost bottom cell in the two tables, the MaxEnt + DS tagger obtained the best results, with a marginal improvement from a 0.0234 error rate to 0.0231, with respect to the non-augmented model. An aspect that can also be observed in the two tables is that there is very little variation in the assignment of individual tags in this data set. A few differences worth noticing are the presence of more mistakes by the MaxEnt tagger in the assignment of tags for conjunctions, numerals, pronouns and verbs, and the production of more mistakes by the MaxEnt + DS tagger with determiner and noun tags. Due to space constraints we have not presented individual tables for the application of the two taggers to the spoken genre data set. However, the data shown in the tables here discussed should be useful for developers interested in improving these two families of algorithms for their application to Spanish data.

Table 4. *Confusion matrix for POS tags assigned by MaxEnt tagger.*

Actual POS	Predicted POS																										POS error rate
	CC	CS	DD	DP	DT	DE	DA	NC	NP	MC	MO	PP	PD	PX	PT	PR	PI	VM	VA	Y	AQ	RG	I	SP	F	W	Z	
CC	**150**																		1			1						0.013
CS		**103**																										0.000
DD			**49**																									0.000
DP				**46**					1																			0.021
DT					**0**																							0.000
DE						**0**																						0.000
DA				1			**654**	1	2			1					2	2			5	1		2				0.025
NC								**1005**	11									5			23			1				0.038
NP								1	**534**				1					1			6					3		0.022
MC			3							**0**																	3	1.000
MO			4								**0**																	1.000
PP									1			**123**																0.008
PD													**9**															0.000
PX														**0**														0.000
PT															**1**													0.000
PR																**150**		1										0.007
PI																1	**15**											0.063
VM								10	4									**585**			8	1		2		1		0.043
VA																			**21**									0.000
Y																				**0**								0.000
AQ								2	9				1						1		**314**							0.040
RG									2													**123**		1				0.024
I																							**0**					0.000
SP	1								2															**960**				0.003
F																									**462**			0.000
W																										**21**		0.000
Z																											**93**	0.000
																										Total ER:		0.0234
																										Total tags:		5548

Table 5. *Confusion matrix for POS tags assigned by MaxEnt + DS tagger.*

Actual POS	Predicted POS																										POS error rate
	CC	CS	DD	DP	DT	DE	DA	NC	NP	MC	MO	PP	PD	PX	PT	PR	PI	VM	VA	Y	AQ	RG	I	SP	F	W	Z	
CC	151																											0.000
CS		103																										0.000
DD			54																									0.000
DP				46					1																			0.021
DT					0																							0.000
DE						0																						0.000
DA							651	6	6			2						3			6	2						0.037
NC							1	993	5				1					9			26	1						0.042
NP							1	2	547									1			5			1		6		0.028
MC			1							0																3		1.000
MO											0																	0.000
PP												128																0.000
PD													6															0.000
PX														0														0.000
PT															1													0.000
PR																148												0.000
PI																	20											0.000
VM								6	5									578			5	1		2				0.032
VA																			18									0.000
Y																				0								0.000
AQ								8	3				1					2			318							0.042
RG									2							1						113						0.026
I																							0					0.000
SP									1									1	1					969				0.003
F																									463			0.000
W																										17		0.000
Z																											96	0.000
																										Total ER:		0.0231
																										Total tags:		5548

4. Conclusions

This study has been aimed at responding to two research shortages identified in computer-based language applications in Spanish. First, the original goal of this study was to improve a linguistic corpus under construction for the immediate purpose of supporting dictionary-making projects through the added capability of retrieving grammatical annotations for the words stored in its database. Achieving this goal has the added benefit of producing the first grammatically annotated subregional corpus for Mexican Spanish, which therefore becomes a precedent for many other similar corpus building projects – only the *PRESEEA* (36) introduces or hosts 42 of these projects. In this sense, with the high accuracy results obtained here, which match current performance expectations and standards, this first goal has been fully accomplished.

Secondly, in the search for state-of-the-art POS tagging algorithms applied to Spanish, two taggers were identified as recently released and lacking an attested accuracy. In order to respond to the first problem, the missing evaluation of these two taggers was carried out. This evaluation has a two-fold contribution. Firstly, it has shown that these taggers are highly accurate and consistent across samples of two very different genres (written text and transcribed oral speech) extracted from external data collections. Secondly, it has also made it possible to compare these taggers results with those obtained by eight other taggers in the SNLP Group suite, which have already been evaluated in four more languages (English, Chinese, Arabic, and German) (5). It should be noted here that although these other taggers have been derived from data in other languages and therefore are independent from the Spanish taggers here evaluated, they are still based on the same tagging algorithms (MaxEnt and MaxEnt + DS) and for linguistic purposes, it is relevant to explore how these algorithms behave across languages. In the context of similar tasks, the taggers in these four languages have achieved an accuracy of 96.97%-97.28% (for three English taggers), 93.46%-93.99% (for two Chinese taggers), 96.61-96.9% (for two German taggers), and 96.26% for the only Arabic tagger. Two languages in the suite, French and Spanish, have not been given accuracy rates by the developing team. With all these figures, it is clear that the 97.2%-97.4% accuracy range for the two Spanish taggers is at the upper end of the ranges reported for the languages tested by the Stanford NLP Group. This is of course encouraging not only for the several corpus building projects going on in this language but for other language applications that may benefit from using these kinds of computational resources.

5. Acknowledgements

We want to thank the students of the upper-level undergraduate course LIO 306 Language and Computers in Universidad de las Américas Puebla. Their participation in the human evaluation has been invaluable. Our special thanks to Silvia Fernanda Montaño García and Karen Adriana Rodríguez Gutíerrez for the many work-study hours they invested in the project. This research study has also benefited from a number of grants obtained by the members of the Cuerpo Académico Lengua, Tecnología e Innovación, at Universidad Autónoma de Baja California.

6. References

(1) Jurafsky D, Martin JH. Speech and language processing: an introduction to language natural processing, computational linguistics, and speech recognition. 2nd ed. Upper-Saddle River, NJ: Pearson-Prentice Hall; 2008. 1024 p.

(2) Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington, MA: Morgan Kaufmann; 2011. 664 p.

(3) Burns RB, Burns RA. Business research methods and statistics using SPSS. UK: Sage. 2008. 556 p.

(4) Denis P, Sagot B. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. Language Resources and Evaluation. 2012 jan;46(4):721-736.

(5) Stanford Natural Language Processing Group (SNLP Group) [On-line]. 2016a. Stanford Log-linear Part-Of-Speech Tagger; [Accessed 14 June 2016]. Available at: http://nlp.stanford.edu/software/tagger.shtml.

(6) Stanford Natural Language Processing Group (SNLP Group) [On-line]. 2016b. Spanish FAQ for Stanford CoreNLP, parser, POS tagger, and NER; [Accessed 14 June 2016]. Available at: http://nlp.stanford.edu/software/spanish-faq.shtml#tagset.

(7) Parra-Escartín C, Martínez-Alonso H. Choosing a Spanish part-of-speech tagger for a lexically sensitive task. Procesamiento del Lenguaje Natural. 2015 mar;54:29-36.

(8) Rico-Sulayes A. Towards a supervised rescoring system for unstructured data bases used to build specialized dictionaries. Revista Facultad de Ingeniería. 2015;24(38):97-106.

(9) Corpus Básico Científico del Español de México (COCIEM) [On-line]. 2016. Bienvenidos; [Accessed 14 June 2016]. Available at: http://www.corpus.unam.mx/cociem.

(10) Proyecto para el Estudio de Sociolingüístico del Español de España y de América (PRESEEA) [On-line]. 2016a. Metodología; [Accessed 14 June 2016]. Available at: http://preseea.linguas.net/Portals/0/Metodologia/Marcas_etiquetas_minimas_obligatorias_1_2.pdf.

(11) Karlsson F. Constraint grammar as a framework for parsing running text. In: Karlgren H, ed., Proceedings of 13th International Conference on Computational Linguistics, Volume 3; 1990; Helsinki, Finland. PA, USA: Association for Computational Linguistics Stroudsburg; 1990. p. 168-73.

(12) Farwell D, Helmrich S, Casper M. SPOT: a Spanish part-of-speech tagger. Procesamiento del Lenguaje Natural. 1995;17:42-53.

(13) Schmid, F. Probabilistic part-of-spe-ech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing; 1994. p. 44-9.

(14) Berger AL, Della-Pietra VJ, Della-Pietra SA. A maximum entropy approach to natural language processing. Computational Linguistics. 1996 mar;22(1):39-71.

(15) Brill E. Transformation-based errordriven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics. 1995 dec;21(4):543-65.

(16) Martínez H, Vivaldi J, Villegas M. Text handling as a Web Service for the IULA processing pipeline. In: Proceedings of the Language Resources and Evaluation Conference 2010, ELRA; 2010. p. 22-9.

(17) Padró L, Stanilovsky E. FreeLing 3.0: Towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference 2012, ELRA; 2012. p. 2473-9.

(18) Solorio T, Liu Y. Part-of-Speech tagging for english-spanish code-switched text. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, ACL; 2008; Honolulu, Hawaii. PA, USA: Association for Computational Linguistics Stroudsburg; 2008. p. 1051-60.

(19) Vivaldi J. Corpus and exploitation tool: IULACT and bwanaNet. In: Proceedings of CILC-2009; 2009; Murcia, Spain; 2009. p. 224-39.

(20) Padró L. A hybrid environment for syntax-semantic tagging [Doctoral Thesis]. Departament de Llenguatges i Sistemes

Informatics. Barcelona, Spain: Universitat Politecnica de Catalunya; 1998.

(21) Toutanova K, Klein D, Manning C, Singer Y. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL; 2003; Edmonton, Canada. Stroudsburg PA, USA: Association for Computational Linguistics; 2003. p. 252-9.

(22) Geffet M, Dagan I. Feature vector quality and distributional similarity. In: Proceedings of Coling; 2004; Geneva, Switzerland. Stroudsburg PA, USA: Association for Computational Linguistics; 2004. p. 247-53.

(23) Wang Q, Schuurmans D. Improved estimation for unsupervised part-of-speech tagging. In: Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering; 2005; Wuhan, China. IEEE; 2005. p. 1-6.

(24) Bienmann C, Giuliano C, Gliozzo A. Unsupervised Part-Of-Speech tagging supporting supervised methods. In: Proceedings of Recent Advances in Natural Language Processing; 2007; Borovets, Bulgaria; 2007.

(25) Bienmann C. Unsupervised part-of-speech tagging in the large. Research on Language and Computation. 2009 dec;7:101-35.

(26) Bienmann C, Riedl M. From Distributional to Contextual Similarity. Darmstadt (GE): Computer Science Department, Technische Universität Darmstadt; 2013. Technical report.

(27) Datla V, Lin K, Louwerse M. Part of Speech Induction from Distributional Features: Balancing Vocabulary and Context. In: Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference; 2014; Florida, USA; 2014.

(28) AnCora [On-line]. 2016. Corpus; [Accessed 14 June 2016]. Available at: http://clic.ub.edu/corpus/en/ancora.

(29) Expert Advisory Group on Language Engineering Standards (EAGLES) [On-line]. 1996. Recommendations for the Morphosyntactic Annotation of Corpora; [Accessed 14 June 2016]. Available at: http://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf.

(30) Instituto Nacional de Estadística y Geografía (INEGI). [On-line]. 2016. México en Cifras; [Accessed 14 June 2016] Available at: http://www3.inegi.org.mx/sistemas/mexicocifras/default.aspx?e=02.

(31) Butragueño PM, Lastra Y. (editors). Corpus sociolingüístico de la ciudad de México. Volumen 1: Nivel alto. Mexico: El Colegio de México; 2011.

(32) Butragueño PM, Lastra Y. (editors). Corpus sociolingüístico de la ciudad de México. Volumen II: Nivel medio. Mexico: El Colegio de México; 2012.

(33) Rodríguez-Alfano L. (editor) [On-line]. Monterrey : 2016. PRESEEA; [Accessed 14 June 2016]. Available at: http://preseea.linguas.net/Equipos/Monterrey.aspx.

(34) Corpus del Español Mexicano Contemporáneo (CEMC) [On-line]. 2016. Bienvenido al Corpus del Español Mexicano Contemporáneo (1921-1974); [Accessed 14 June 2016]. Available at: http://www.corpus.unam.mx/cemc.

(35) Corpus Diacrónico y Diatópico del Español de América (CORDIAM) [On-line]. 2016. Corpus Diacrónico y Diatópico del Español de América; [Accessed 13 June 2016]. Available at: http://www.academia.org.mx/Cordiam.

(36) Proyecto para el Estudio de Sociolingüístico del Español de España y de América (PRESEEA) [On-line]. 2016b. Equipos

de PRESEEA; [Accessed 14 June 2016]. Available at: http://preseea.linguas.net/Equipos.aspx.

(37) Agerri R, Bermudez J, Rigau G. IXA pipeline: Efficient and ready to use multilingual NLP tools. In: Proceedings of the 9th Language Resources and Evaluation Conference, ELRA; 2014. p. 3823-8.

(38) Berthelsen H, Megyesi B. Ensemble of classifiers for noise detection in PoS tagged corpora. In: Proceedings of the Third International Workshop on Text, Speech and Dialogue; 2000; Brno, Czech Republic: Springer-Verlag; 2000. p. 27-32.