Applying Superstrate POS Taggers to their Derived Creole Languages: Chavacano

Raúl Camargo & Lily Kurek

LIN4770C Spring 2024

Dr. Zoey Liu

Table of Contents

I. Introduction

Under-researched creoles such as Chavacano, a Spanish-based creole spoken in the southern Philippines by about three hundred thousand people, may benefit from computational linguistic resources from their superstrates, often rich in computational developments. In the present paper, we present the first development of a Chavacano part-of-speech (POS) tagger based on an adapted Spanish POS tagger from the spaCy NLP python library. A POS tagger is a crucial computational linguistic tool with applications in constituency and dependency parsing and sentiment analysis, among other NLP tools. Developing a POS tagger could open the door for Chavacano-speaking communities to have language technologies available in their native language, such as machine translation or speech recognition software, and facilitate further research in Chavacano linguistic research.

We started by manually translating a sample of Chavacano-English parallel corpus to Chavacano-Spanish and evaluating the accuracy of using the Spanish POS tagger to raw Chavacano data. We then evaluated this approach's qualitative benefits and drawbacks and developed a modified version of the tagger to accommodate Chavacano's lexical and grammatical features. We finished by discussing the limitations of our results, the implications for developing POS taggers in other creoles languages and language technologies, and the possible improvements.

II. Previous Scholarship

A. Creole Languages

As long as people from different linguistic backgrounds have come into contact, there have been linguistic developments. Of the effects of linguistic contact, none has been more hotly debated than the emergence of creole languages. Early philologists and linguists in the centuries after colonization held the viewpoint that creole languages were inferior to their parent languages (DeGraff, 2005). It was in the mid- to late 20th century where "creology" – the study of creoles – took off.

Hall (1966) describes the formation of creoles as an "outgrowth" from their original pidgins (p. XI). Pidgins are languages created by the need to communicate between two or more groups/people that do not speak the same language. Though the sociohistorical context of each pidgin greatly affects its genesis, pidgins tend to take influence from two or more "substrate" and "superstrate" languages. Substrate languages tend to be indigenous languages that donate grammatical features, whereas superstrate languages tend to be the higher prestige language that donates its lexicon (Hall, 1966). These pidgins have no native speakers; they are used solely for communication that would otherwise be impossible. When pidgins are longstanding enough to flesh out their vocabulary and grammar more and they acquire first language status, they become creoles (Hall, 1966).

There have been attempts to discern whether creoles constitute a separate typological profile from other languages (Daval-Markussen, 2014). Creoles are not formed from continuous phonological and syntactic changes over time, rather they are sprung into existence as pidgins. Though their grammars are often analytic and rely little on agglutination (Daval-Markussen, 2014), creoles can vary widely in their morphology and syntax. Additionally, due to the effects

of colonization, many of the world's most studied creoles have influences from one or more Indo-European languages (Hall, 1966).

B. Creoles in Computational Linguistics

Of the 2,485 languages analyzed by Joshi et al. (2020), more than 88% were classified as "left behind", they "are still ignored in the aspect of language technologies." (p. 6284). As previously mentioned, the origin that many creoles have in colonization has led to them being understudied (DeGraff, 2005). Labeled and unlabeled data alike are hard to find for these low-resource languages, though there are creoles with millions of speakers, for example Nigerian Pidgin English with around 110 million (Affia, 2023), Jamaican Patois with around 3 million (Farquharson, 2013), and Haitian Creole with around 10 million (Fattier, 2013).

Language technologies have been rapidly progressing, and computational linguistics is a rapidly expanding field. European languages, which so often provide a basis for pidgins and creoles, have an abundance of research and resources (Joshi et al., 2020). Employing high-resource languages to benefit related low-resource languages is an exciting prospect for languages like creoles, which suffer from a lack of attention in the field.

Previous research has used English syntactic knowledge to improve accuracy on the English-based creole Singlish's POS taggers and dependency parsers (Wang, et al. 2019). Similar studies have been performed on French based creoles to mixed success (Mompelat et al., 2020; Lent et al., 2022). As creoles tend to be lower-resource but closely related to a high-resource language, the latter can be employed to propagate and improve resources for the former.

C. Chavacano

Chavacano is a Spanish-based creole spoken primarily in Zamboanga, a city in the southern Philippines. According to the Philippines Statistics Authority, Chavacano is the main language in 106,375 households, the vast majority of which speak the Zamboangueño dialect. Chavacano is thought to have emerged in the beginning of the 17th century in the Manila-Cavite area from the interaction of Spanish soldiers and non-European groups in their service (Jacobs & Parkvall, 2018). Because Manila was Spain's gate to trade with China during the colonial era, Spanish soldiers would have interacted with merchants, slaves, and soldiers from different linguistic backgrounds, consolidating Chavacano's birthplace. Jacobs and Parkvall, in their analysis of Chavacano genesis, suggest twelve main substrate languages mainly from surrounding geographical areas (i.e. Halmaheran languages, Tagalog, Bisayan languages), Chinese and Portuguese trade (i.e. Min Nan, Portuguese Pidgin), and more than twenty substrate languages spoken by slaves and merchants from Europe and Asia.

As a result of its inheritance from multiple substrate languages, Chavacano has a drastically different grammar than Spanish. It allows both VSO and SVO word order in sentences, but the latter, inherited from Spanish, has a different connotation than the most common VSO: it is used to highlight the subject of the sentence. Unlike Spanish, Chavacano only has head-final noun phrases, it lacks grammatical gender, and it marks plural number preceding the noun with the particle *mga* and an optional reduplicant, both aspects inherited from Tagalog (Steinkrüger, 2013). Another important difference with Spanish is the tense-aspect system. While Spanish inflects verbs to connote tense, aspect, and mood, Chavacano precedes the infinitive form of the verb with one of three particles: the perfective marker *ya* for the past

tense, the imperfective marker *ta* for present tense, and the irrealis mood marker *ay* for future tense. All three particles are of Spanish origin (*ya* → *ya* 'now/already', *ta* → *está* 'it is', *ay* → *hay* 'there is'). Though these particles have been debated to work morphophonologically as prefixes (Steinkrüger, 2013), we refer to them in this paper as particles as they are orthographically separated from verbs. Similarly, Chavacano negates propositions using three particles: *no hay,* used for the past tense; *hindé,* a negator inherited from Cebuano used for the present and future, and *no*, used for imperatives. Other examples of particles used in Chavacano are: *ba*, an optional question marker of Tagalog origin; *si*, an Austronesian determiner for proper nouns, and *con*, particle of Spanish origin (*con* → *con* 'with') used as an object marker.

III. Methodology

As Spanish is a high resource language and the superstrate of Chavacano, its lexicon is similar to that of Chavacano with the exception of some minor orthographic variations. This makes Chavacano a good candidate for testing with Spanish word-level computational tasks. In order to analyze a Spanish POS tagger's effectiveness on Chavacano, the Spanish POS tagger imported from the spaCy python NLP library was chosen. The tagger was trained on a Spanish news corpus accessible via the `es_core_news_md` spaCy pipeline.

Data from a Chavacano-English parallel corpus from Taoteba, an open online collection of sentence-level translations, accessed via OPUS, an online collection of parallel corpora, was collected (Tiedemann, 2012). The Chavacano-English parallel corpus, having roughly 2500 sentences, was chosen. A Chavacano-Spanish parallel corpus exists, but it has significantly fewer (roughly 100) sentences.

Sixty sentences were randomly selected and manually translated to Spanish by a native Spanish speaker with a linguistic academic background. Translating from the larger Chavacano-English parallel corpus allowed for a larger and more representative sample of Chavacano sentences with more grammatical and lexical variety than the sentences present in the smaller Chavacano-Spanish corpus. Both the original sentences (Chavacano) and their translations were manually tagged to serve as gold standards for testing.

The Chavacano sentences and their translations were tagged using the spaCy Spanish POS tagger without any adaptation. The results were then compared to the gold-standard tagged sets to estimate the accuracy of the tagger for Chavacano in relation to Spanish. Based on these results, we performed a qualitative analysis and identified possible improvements that could be used to modify the POS tagger. Finally, the differences between the adjusted and unadjusted taggers were statistically analyzed in order to reject the null hypothesis.

IV. Results

A. Baselines

For languages with few resources, taking "off the shelf" taggers could be a viable way to increase their standing in computational linguistic research. However, it can hardly be expected that a tagger for a language would work perfectly for a different (albeit related) language without adjustment.

Each iteration of the tagger had its accuracy calculated and compared to three baselines (Table 1). It is important to note that, due to the use of grammatical particles in Chavacano in

contrast to the fusional verbal morphology and pronoun dropping in Spanish, the Chavacano data contains more tokens (n=560) than the Spanish data (n=509) and a similar number of types (Chavacano: 244, Spanish: 251).

	Accuracy (%)		Number correctly tagged	
	SP	CH	SP	CH
Baseline 'PUNCT'	16.67	13.97	75	70
Baseline 'PRON'	11.56	12.77	52	64
Baseline Random	7.11	6.19	32	31

Table 1: Accuracy of tagger baselines for Spanish (SP) and Chavacano (CH) data

The most accurate baseline was the majority tag baseline. The punctuation tag was the most common in the translated Spanish sample, so this tag was applied to every token in the Chavacano sample, and was compared for accuracy against the gold standard. The pronoun tag was the most common tag in the Chavacano data. This was slightly less accurate than the punctuation baseline. The most inaccurate baseline was the random baseline, which used python's `random` module to select one of the thirteen tags that appeared in the Spanish gold standard data for each token.

B. Model Accuracy

The first iteration of the model involves no manipulation of the spaCy results. The Spanish trained tagger was applied to the Spanish and the Chavacano texts. These results were then compared to the gold standard for each language (Table 2).

	Accuracy (%)	Number correctly tagged
Spanish Gold Standard Data	88.67	399
Chavacano Gold Standard Data	56.69	284

Table 2: Accuracy of spaCy Spanish tagger for Spanish and Chavacano data

Though the model was far more accurate than any of the baselines, areas of possible improvement were noted after this iteration. Many words in the Chavacano text were falsely tagged as proper nouns. Of the tokens tagged as proper nouns in the Spanish data, only four were not proper nouns according to the gold standard. In the Chavacano data, tokens falsely tagged as proper nouns were mostly particles (n=26) and pronouns (n=22).

The second iteration, wherein particles *mga*, *ya*, etc. were automatically tagged as particles by including them in a stoplist after the spaCy tagger, yielded a significantly ($p=0.0194$) better accuracy rate of 63.07% for the Chavacano data.

The third iteration involved tagging Chavacano pronouns automatically after the application of the second iteration. This yielded a slightly better accuracy rate of 66.47%, but this was not significantly better than the previous iteration ($p=0.1303$). When only the pronouns are handled (particle tags left unchanged from the spaCy version), this iteration did still improve upon the first ($p=0.0265$), but less so than the second iteration.

V. Discussion

The Spanish POS tagger clearly outperformed the baselines when applied to the Chavacano corpus. Despite Spanish's head initial noun phrases and Chavacano's head final, the tagger did not have any issues correctly tagging adjectives. Though the accuracy hovered slightly above 50% (Table 2), this tagger offers a less resource-intensive way to contribute to the body of computational linguistic research on creole languages like Chavacano.

Chavacano currently lacks a large POS tagged corpus. POS tagging is necessary and helpful for other linguistic tasks such as constituency parsing, sentiment analysis, and even speech synthesis (Schlünz, 2010). However, when a language lacks the manually tagged data for use in training a POS tagger, all subsequent tasks will remain out of reach. As mentioned previously, cross-linguistic applications of POS taggers have been minimally tested, especially on creole languages. The improved accuracy of the Spanish POS tagger on Chavacano after minimal adjustments shows promise for this methodology on other creole languages.

A. Limitations

Through the course of our project, data and time limitations were encountered, which will have to be addressed to improve performance and applicability.

i. Data

The parallel corpus we used to create the gold-standards is small (n=60 sentences) and has a small type-token ratio (0.087), indicating a low degree of lexical complexity. Because most of the sentences in the corpus are everyday informal Chavacano, the POS tagger might not perform well in data from other genres, such as literature or academic works. Furthermore, the Taoteba sentences and translations were provided by five Chavacano native speakers, which may not accurately represent the language spoken by the wider community; a more comprehensive dataset would need more contributors. A final limitation related to the corpus is the orthographic conventions followed by the different contributors. Because Chavacano has no formal orthographic conventions, the corpus contains repeated words with different spellings. This might have impacted the performance of the Spanish POS tagger depending on how similar the spelling is to Spanish. For example, the word *ay* is also spelled *hay* in different instances in the corpus. Both words exist in Spanish but differ in meaning and POS tag: *ay* is an interjection, while *hay* is the present tense of the auxiliary verb *haber* 'to have.' Though this particular instance in the data was addressed in the methodology due to its frequency as a particle, other words may have created a similar issue for the tagger.

ii. Time

The time constraint for this project also had a negative impact on the results. First, the translations to Spanish provided are based on the limited knowledge gleaned from Chavacano grammars and the original translations to English, which might have affected the quality of the gold-standards. Because of the need for more time and human resources to translate more sentences, the data set is small (n=60 sentences) and contains only 13.7% (n=244) of the types present in the parallel corpus (n=1775). As sentences were selected randomly from the corpus, some lexical and grammatical information may not have been captured in the data.

B. Applications

For languages lacking a large, POS-tagged corpus, bootstrapping could be a viable way to annotate data for further use. Bootstrapping is the process of using a "seed" corpus – a small annotated corpus – to train a POS tagger. This tagger is then applied to a slightly larger, unannotated corpus, the errors corrected manually, and the output is used to repeat the process over again. This process has been shown to produce satisfactory results when applied to a corpus of unannotated Urdu tweets (Baig et al., 2020). As Chavacano has a large online presence on Facebook (Zamboanga de Antes, n.d.), unannotated online data could be applied in a similar fashion to the data from Baig et al. (2020). The creation of a large corpus of POS tagged data would open the door for more computational tasks to be performed on Chavacano, and for the applications of those tasks in language technology (such as machine translation).

Additionally, the methodology used in this paper could be applied to other creole languages, especially those which have a high-resource superstrate language, but lack a large corpus of manually annotated data (e.g. Tok Pisin, Hawaiian Pidgin English, and Kreyol).

C. Further research

In the future, an orthographic normalization method could be tested to see if it improves performance. Many words in Chavacano are spelled differently than their Spanish counterparts, and the preprocessing step of altering the data's orthography to more closely match that of Spanish may improve in-vocabulary accuracy. Another strategy to address this issue is to use a normalization method such as minimal edit distance, which may eliminate orthographic variation as a confounding factor for often unstandardized creole languages.

Application of this method to other Spanish-based creoles may give insight into the lexical similarity (or lack thereof) between creoles of the same superstrate. Creoles of unknown or mixed superstrate origin could use the accuracy of a superstrate POS tagger to determine to which language they are more lexically similar.

VI. Conclusion

Creole languages, though they are often widely spoken and have cultural significance around the world, lack the annotated linguistic data and systems needed to move on to the next stage in language technology. This is in contrast to their (mostly) European, high-resource superstrate languages, from which they derive their lexicon. One such example of a creole is Chavacano, a Spanish-based creole spoken in the Philippines. The development of a POStagger

for Chavacano using an off-the-shelf Spanish tagger with minimal adjustments is a promising outcome for other low-resource creoles.

Though the methodology was limited in previously-annotated data (a circular problem) and by time constraints, the tagger achieved a significantly better accuracy than three baselines, though it was less accurate than for its intended language, Spanish. Head finality proved to be a non-issue for the tagger, and minor adjustments for grammatical particles not found in the superstrate language significantly improved performance.

In the future, this methodology may create POS tagged corpora through bootstrapping, be used to compare lexicons between creoles with common superstrates, and be improved upon with the use of orthographic normalization. Furthermore, the development of POS taggers for under-researched languages (along with other language technologies) will make tasks such as speech synthesis, sentiment analysis, or machine translation available. Currently, such technology is only available at its highest performance capability for a handful of widely spoken and studied languages, like English and Spanish.

VII. Bibliography

Affia, P. (2023). Nigerian Pidgin: the identity of a Nigerian away from home. *Working papers in Applied Linguistics and Linguistics at York 3*, 69-84. doi.org/10.25071/2564-2855.23

Baig, A., Baloch, A., Kazi, H., Rahman, M.U. (2020). Developing a POS tagged corpus of Urdu tweets. *Computers 2020, 9*(4), 90. doi.org/10.3390/computers9040090

Bali, K., Budhiraja, A., Choudhury, M., Joshi, P., & Santy, S. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). doi.org/10.18653/v1/2020.acl-main.560

Bugliarello, E., Lent, H., & Søgaard, A. (2022). Ancestor-to-creole transfer is not a walk in the Park. *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. doi.org/10.18653/v1/2022.insights-1.9

Dakota, D., Kübler, S., & Mompelat, L. (2022). How to parse a creole: when Martinican Creole meets French. *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4397–4406). https://aclanthology.org/2022.coling-1.387

Daval-Markussen, A. (2013). First steps towards a typological profile of creoles. *Acta Linguistica Hafniensia, 45*(2), 274–295. doi.org/10.1080/03740463.2014.880606

DeGraff, M. (2005). Do Creole languages constitute an exceptional typological class? *Revue française de linguistique appliquée*, X, 11-24. doi.org/10.3917/rfla.101.24

Farquharson, J.T. (2013). Jamaican. In: Haspelmath, M., Huber, M., Maurer, P., & Michaelis, S.M. (eds.) *The survey of pidgin and creole languages. Volume 1: English-based and Dutch-based Languages*. Oxford: Oxford University Press.

Fattier, D. (2013). Haitian Creole. In: Haspelmath, M., Huber, M., Maurer, P., & Michaelis, S.M. (eds.) *The survey of pidgin and creole languages. Volume 2: Portuguese-based, Spanish-based, and French-based Languages.* Oxford: Oxford University Press.

Hall, R.A. (1966). *Pidgin and creole languages.* Cornell University Press.

Jacobs, B. & Parkvall, M. (2018). The genesis of Chavacano revisited and solved, *Lingua, 215*, 53-77. doi.org/10.1016/j.lingua.2018.09.006.

Philippine Statistics Authority. (2023). *Language/Dialect Generally Spoken at Home.* psa.gov.ph/content/tagalog-most-widely-spoken-language-home-2020-census-population-and-housing

Schlünz, G.I. (2010). *The effects of part-of-speech tagging on text-to-speech synthesis for resource-scarce languages*. Master's Thesis, North-West University. repository.nwu.ac.za/bitstream/handle/10394/4944/Schlunz_GI.pdf?sequence=2

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

Wang, H., Yang, J., & Zhang, Y. (2020). From genesis to creole language. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(1), 1–29. doi.org/10.1145/3321128

Zamboanga de Antes. (n.d.). In *Facebook* [Group page]. Retrieved April 28, 2024, from www.facebook.com/groups/ZamboangaDeAntes/about

VIII. Statement of Contribution

Both team members took an equal share of the responsibility. Raúl was responsible for tagging Chavacano data and translating it to Spanish. He also came up with the idea to do a project on creoles, and Chavacano data was found and agreed upon by both members. Lily was responsible for tagging the Spanish translations, though Raúl, as a native speaker, checked the tags. As for the coding, Lily wrote the functions for calculating accuracy of the baselines and the models, and Raúl implemented the particle and pronoun stoplists as well as the functions to count the number of tokens and types in the data. Raúl made the handout and Lily made most of the poster. Raúl wrote parts I, II C, III, and V A, while Lily wrote parts II A-B, IV, V B-C, and VI. Both group members were heavily involved in the editing process to ensure proper flow, correct citation, and style, reviewing each other's work throughout the creation of both the presentation and the report.