Towards a part-of-speech tagger for Sranan Tongo

Nicolás Cortegoso Vissio, Viktor Zakharov

Abstract—**This paper is the continuation of a work submitted to the International Conference Corpus Linguistics 2021 [1]. On that occasion, a rule-based stochastic hybrid part-of-speech tagger (POS) was introduced for Sranan Tongo, a Creole language from South America with around half a million speakers. Since Sranan Tongo does not have a written corpus and text annotation is an expensive and time-consuming task, it was proposed to take a first step in training a POS tagger using only 550 hand-annotated sentences with part of speech tags.**

In this new contribution, the development of the POS tagger for Sranan Tongo goes a step further with the addition of more training data. For this matter, the tagger was used to annotate 2,406 sentences. The tagging results were hand-corrected and employed to retrain the model. A comparison is shown between the performance of the POS tagger on three texts before and after the inclusion of the new training data.

***Keywords*— part-of-speech tagger, Sranan Tongo, low-resource, Hidden Markov Model**

I. INTRODUCTION

Sranan Tongo (literally “language of Suriname”) is the most widespread Creole of the Republic of Suriname in South America. Like many other Atlantic creoles, it emerged among the slaves that were brought to America five centuries ago to work on the plantations. Nowadays Sranan Tongo is spoken by more than four hundred thousand people in urban areas along the coastline of the country and it is often used as lingua franca between the different ethnic groupManuscript received October 14, 2021.s. It also counts two hundred thousand speakers from the Surinamese diaspora living in The Netherlands. English is the main lexifier of Sranan Tongo, while Gwe and other languages from west-Africa are considered to be its substratum. Dutch became the superstratum when the colony passed from British hands to the Netherlands.

From an NLP perspective, Sranan Tongo is a low-resource language. There is no corpus available for Sranan Tongo in the public domain and, since written material is scarce, compiling one is not an easy task. After the independence from the Netherlands, Dutch remained as the official language of Suriname, and for this reason, the press and the government administration are carried out in the language of the former metropolis. Although literary works have been published in Sranan Tongo since 1960, this is still primarily a spoken language used in everyday communication and, despite attempts at standardization, written Sranan Tongon has significant variation in spelling.

Sranan Tongo has a somewhat small vocabulary. Speakers often fill in the lexical gaps with Dutch words and, in everyday communication, code-switching is almost the norm. In the example below [2], Dutch words are marked in italics:

Want we tan kree nomo fu den *prijs* ma un ap wan *president* nanga *regering*ete!

Consequently, the border between Sranan Tongo and Surinamese Dutch is blurry, and therefore it is difficult to decide whether a word can be considered a legitimate member of the Sranan Tongo vocabulary. Sranan Tongo also has many homonyms due to the loss of phonemic features from the lexifier language.

With very few exceptions, the words in Sranan Tongo do not change. However, reduplication and compounding are important aspects of Sranan Tongo’s morphology. Serial verb constructions are an extended syntactic resource for generating new meanings.

Another salient feature of Sranan Tongo is word multifunctionality: lexical elements can function as members of different grammatical categories without any change in their form [3]. As a result, the same word-form can have different grammatical functions depending on the place it occupies in the sentence. In most cases, the part of speech for a given word-form cannot be determined without the context. In the example below from the dictionary entry “dyadya” (qualified through training and experience) [4], the first “feti” is a noun and the second one, a verb:

Solanga feti no e feti, yu no man si suma na den dyadya srudati.

(As long as fight no is fought, you no can see who are the real soldiers)

(As long as there is no war, you cannot tell who the real soldiers are.)

Even though Sranan Tongo has a fairly strict word order, under the influence of Dutch, some constructions developed a closer alternative to Dutch syntax. The reader is referred to K Yapko, A Bruyn [5] for examples of how the Dutch language affected locative constructions.

After this brief overview of Sranan Tongo, two main challenges in building a POS tagger for this language can be mentioned. The first is related to the non-standardized spelling and the presence of foreign words (mainly Dutch). The second revolves around the lack of morphological features to identify parts of speech, extended word multifunctionality and homonymy.

III. THE RULE-BASED STOCHASTIC APPROACH

Compiling a lexicon for Sranan Tongo is a challenging task. The variation in spelling and the widespread use of Dutch words make it very likely to find many words in texts that are not included in the lexicon, regardless of its size. The principle that was followed when developing a hybrid tagger

Manuscript received October 14, 2021.

N. Cortegoso Vissio is with the Mathematical Linguistics Department, Saint Petersburg State University, Saint Petersburg, Russia, ORCID 0000-0003-1683-7270 (e-mail: st082534@student.spbu.ru).

V. Zakharov is with the Mathematical Linguistics Department, Saint Petersburg State University, Saint Petersburg, Russia, ORCID 0000-0003-0522-7469 (e-mail: v.zakharov@spbu.ru).

was that even a small lexicon could include most of the closed-class words, even with their alternative spellings. The compiled lexicon was supposed to contain nearly all of the closed-class words like articles, pronouns, modals, etc. Open-class words such as nouns, verbs, interjections, etc were included only if they have homonyms in the closed-class words. For example, the lexicon contained “sa” as a noun (the saw) and as a verb (to saw), because it is a homonym of the modal “sa” (shall/should, a closed-class word). The compiled lexicon totaled 384 word-forms extracted from the online version of the “Wortubuku fu Sranantongo” with the parts of speech they can take. POS tags for words outside the vocabulary were expected to be predicted by the tagging algorithm, according to their place in the sequence of words within the sentence. For example, in Sranan Tongo, after an article, it can be expected to find a noun, an adjective but not a verb.

The valid sequence of part of speech for the POS tagger was not modeled by hand-written rules but by 3-gram probabilities. As mentioned above, despite having a strict word order, some constructions in Sranan Tongo show a degree of variability that may be very complex to capture with rules. On the other hand, a 3-gram model can easily account for alternative constructions and can be trained on a small representative set of sentences. For this purpose, 550 sentences were extracted and manually annotated with POS tags: 329 sentences (2853 tokens) from the APiCS database and 221 sentences (1660 tokens) from “Papers on Sranan Tongo”. Because both collections of sentences are part of language descriptions, they were trusted to represent the majority of valid POS tag sequences and therefore to be reliable sources for learning 3-gram probabilities.

The obtained tagger is a hybrid because it employs a lexicon and some rules to assign possible parts of speech tags to each of the words in the sentence as in a rule-based approach, but it relies on the probabilities of a 3-gram POS model to disambiguate them.

IV. THE TAGGING ALGORITHM

The tagging algorithm can be described in 5 steps. The first three steps consist of a set of rules to assign possible POS tags to each word in the sequence, while the last two apply probabilities to disambiguate them:

1. All words are pre-tagged with a list of open-class words, for instance, noun, verb, adverb, attributive and predicative adjectives. This is an attempt to model word multifunctionality in Sranan Tongo and to deal with words outside the vocabulary and the non-standard spelling.

2. The tagging algorithm looks up the words in the lexicon. If the word is found, then the pre-assigned POS tags are replaced with those from the lexicon. Closed-class words, especially functors are expected to be identified in this step. For example, the word-form “lobi” (love) can be a verb and a noun, but nothing else. This restricts the overgeneralization of the previous step when pre-assigning tags.

3. In case the word does not exist in the lexicon, then a simple rule is applied to identify proper names: if the first letter is capitalized and the word is in any position other than the start of the sentence, then it is considered to be a name and the pre-assigned tags are replaced by the proper name tag. However, if that word occurs at the beginning, the uppercase letter can no longer be taken as an indicator that the word is a name, and instead, the proper name tag is added to the already pre-assigned tags.

4. The POS tags assigned to the words are given a probability that is estimated from the distribution of the POS tags in the training set. These probabilities add up to 1.

5. The POS tag/word probabilities are then combined with a 3-gram POS tag model that finds the most likely POS-tag sequence.

For example, given the sentence “A umapikin lobi Kofi” (the girl loves Kofi), the tagging algorithm proceeds as follows:

Table I. The sequential steps of the the tagging algorithm

	A	UMAPIKIN	LOBI	KOFI
1	noun, verb, adverb, pred adj, attr adj	noun, verb, adverb, pred adj, attr adj	noun, verb, adverb, pred adj, attr adj	noun, verb, adverb, pred adj, attr adj
2	pronoun, article, copula, locational	noun, verb, adverb, pred adj, attr adj	noun, verb	noun, verb, adverb, pred adj, attr adj
3	pronoun, article, copula, locational	noun, verb, adverb, pred adj, attr adj	noun, verb	proper name
4	pronoun: 0.56, article: 0.35, copula: 0.02, locational: 0.06	noun: 0.40, verb: 0.39, adverb: 0.11, pred adj: 0.05, attrib adj: 0.04	noun: 0.50, verb: 0.49	proper name: 1
5	article	noun	verb	proper name

1. pre-assigns to each word the same list of POS tags: noun, verb, adverb, attributive and predicative adjective;

2. finds “a” and “lobi” in the lexicon, consequently, the pre-assigned POS tags for those words are replaced by those indicated in the lexicon;

3. identifies “Kofi” as a proper name;

4. estimates the probabilities of each POS tag;

5. disambiguates the tags finding the most probable POS tag sequence.

The model was tested on 70 sentences extracted from the dictionary entries of the “Wortubuku fu Sranantongo”. The sentences were chosen *ad-hoc* to cover different grammatical constructions and include ambiguous words regarding parts of speech. The testing set as a whole contained at least twice each of the POS tags.

Regarding the POS tags employed in the tagger, in addition to the classical parts of speech such as noun, pronoun, proper name, verb, adverb, preposition, interjection, subordinating, and coordinating conjunctions, some others specific to the language were used as tense and aspect markers.

In the previous paper, three different metrics were proposed for assessing the probability tag/word from step four. The metrics translated the total POS tags counts from the testing set into a probability distribution for a given word-form. An experiment was designed to test their performance on different sizes of the training set. The best performing metric in combination with the 3-gram model trained on the larger training set achieved an average F-score of 79% for all the POS tags. However, this 79% can not be taken as the overall performance of the model, but simply as an indicator to choose the best working metric with the learning data.

V. ONE STEP AHEAD

In this new stage of development, the rest of the sentences of the dictionary “Wortubuku fu Sranantongo” were automatically tagged and manually corrected in order to obtain more annotated data to retrain the model. The example sentences from the dictionary entries show how a word is used in its various senses, and therefore, they are good candidates to provide lexical variation to the training set. After deleting the repeated examples, the entries of the “Wortubuku for Sranantongo” account for a total of 2406 sentences (plus 70 that were already tagged for testing purposes in the previous paper).

Table II. The size and word variation of the training data

TRAINING DATA	SENTEN-CES	TOTAL WORDS	UNIQUE WORDS	RATIO TOTAL / UNIQUE WORDS
Nickel, Wilner + APiCS	478	3495	549	0.1571
Wortobuku fu Sranantongo	2476	22398	1873	0.0836
Total	2954	25893	2008	0.0775

As stated above, the previous lexicon was compiled by hand. It included a list of 384 word-forms and the POS tags they can take. In this new phase, the hand-compiled lexicon is no longer used. Instead, a 3-gram Hidden Markov Model (HMM) is implemented, so that the vocabulary is learned directly from the training set. For an overview of how an HMM works, the reader can refer to D. Jurafsky, J.H. Martin [10]. The difference with a pure HMM resides in retaining the step of the hybrid tagger of pre-assigning open-class POS tags to handle the words that were not observed in the training set. Therefore, these words are given a probability based on their distribution in the training set.

VI. TESTING SET

Unlike the previous paper, where the experiment was carried out in a collection of independent sentences, here the performance of both the HMM and the hybrid model are evaluated on texts. Short stories and poems constitute the majority of texts published in Sranan Tongo. On the Internet written material in Sranan Tongo is really scarce. Even in the year 2021, Wikipedia still has very few articles that can be considered well-formed, with a length exceeding one paragraph and written with a consistent spelling and syntax. For this reason, Wikipedia is dismissed for the moment as a reliable source for testing the POS tagger.

The following three texts were selected and manually tagged with POS tags:

Text 1: “Skowtu hori yu na ini a tori fu wan ordru fu den bakrakondre, nanga den tyari yu na skowt’oso noso wan tra presi pe den o yere yu” [11]. English translation: “You have been detained under a European arrest warrant and taken to a (police) station or another interrogation location”. This is one of three texts in Sranan Tongo from the Ministry of Security and Government from The Netherlands found after a simple web search.

Text 2: “A gridi frow fu fisman Albert”. Grace MacBean. Institut voor Taalwetenschap (SIL), 1993 [12]. English: “The greedy wife of Albert the fisherman”. As mentioned above, the SIL website offers a selection of traditional folk stories in Sranan Tongo. The HMM is expected to provide good coverage for the words in any of the short stories on the SIL website, as the sentences in the training set come from the dictionary featured in the same place.

Text 3: “San pesa ini Kaneri” (Eddy Pinas) [13]. English: “What happened in Kaneri”. This is an excerpt from a text found in a bundle of short stories from Suriname (in Dutch or in Sranan Tongo) compiled by Michiel van Kempen. This collection is available in the section dedicated to Surinamese literature of the online library Digitale Bibliotheek voor de Nederlandse Letteren (Digital Library for Dutch Literature) ww.dbnl.org. The story was cut to match approximately the length of the two previously selected texts. This text is included to provide a short story from a source other than SIL.

VII. EXPERIMENT AND DISCUSSION

Before tagging the texts, it was verified the vocabulary coverage for both the hand-compiled lexicon from the hybrid tagger and the learned lexicon from the HMM model. For the sake of comparison, an attempt has been made to use texts with a similar quantity of words. Table III below shows the obtained values:

Table III. The vocabulary coverage of the pre-compiled lexicon of the hybrid tagger and the learnt lexicon from the HMM model

TEXT CONTENT	HYBRID MODEL COVERAGE	HMM MODEL COVERAGE

text	sent.	words	unique	ratio	out-of-voc.	coverage	out-of-voc.	coverage
1	91	1657	215	0.1298	135	0,3720	66	0,6930
2	161	1706	224	0.1313	105	0,5312	15	0,9330
3	111	1627	396	0.2434	281	0,2904	128	0,6768

The first five columns describe the general statistics for each of the selected texts. The columns “sent.” and “words” contain the number of sentences and words (excluding punctuation) in each text. The column “unique” indicates the number of unique words and “ratio” refers to its proportion in the total. The first two texts are similar regarding these values, while the third shows much more variety in terms of vocabulary.

The last four columns count how many words are not covered by the lexicons of the respective models. The hand-compiled lexicon from the hybrid model covers 53% of the words in text 2 and less than 30% in text 3, which contains the highest proportion of unique words. The vocabulary of the HMM has better coverage of words than the hand-compiled lexicon. This is not surprising, since the HMM gathers 1873 words while the hand-compiled lexicon has only 384. An important remark is that, despite the hand-compiled lexicon containing fewer words, they are also the most frequent words in the language and, therefore, they are likely to appear in any text, regardless of the topic. When the lexicon of closed-class words is expanded further to include open-class words, then the source of these new entries becomes more relevant. This can be observed in the case of text 2, where the coverage of the HMM model reached 93%. Text 2 was taken from the SIL website, the same place that hosts the online dictionary, whose entries were extracted to train the model. Moreover, the 7% of the words in text 2 that were not learned by the HMM were those that, despite being in the online dictionary, their entry does not have a sentence that exemplifies its use. The coverage percentage drops significantly when it comes to texts from other sources than SIL, as shown by the values for texts 1 and 3.

The POS tagger code (written in Python3) and the training and testing data are available in a public Github repository [14].

VIII. Tagging Results

The texts were tagged by the rule-based stochastic hybrid and the HMM tagger. The hybrid tagger consists of a hand-compiled lexicon of 384 words and a 3-gram POS model trained on 478 sentences from the APiCS dataset and the “Papers on Sranan Tongo”. The HMM was trained first on the 2476 sentences from the “Wortubuku fu Sranantongo” and then employing the totality of the annotated sentences, that is, the previous 2476 with the addition of the 478 used to train the 3-gram POS model of the hybrid tagger. The results are shown in Table IV below:

Table IV. The efficiency values of 3 tagging models

	Hybrid model			HMM (2476 sentences)			HMM (2476 + 550)		
	precision	recall	F-score	precision	recall	F-score	precision	recall	F-score
text 1	.61	.68	.60	.68	.60	.62	.73	.66	.67
text 2	.64	.67	.63	.76	.67	.69	.77	.70	.71
text 3	.55	.57	.53	.68	.59	.61	.69	.61	.63
average	.60	.64	.59	.71	.62	.64	.73	.66	.67

All three trials achieved better results when tagging text 2 and performed the worst with text 3. This might show a correlation between vocabulary coverage and overall performance. In the case of the HMM, the addition of 478 sentences improved a 3% on average the F-score for the three texts. According to the experiment, combining an HMM with the current amount of annotated data an F-score between 60-70% can be expected when tagging texts. Not a very satisfying result yet, but certainly an improvement over the hybrid POS tagger.

IX. Conclusions

Although a correlation is assumed, an experiment should be conducted to substantiate the relationship between vocabulary coverage and model performance. In either case, heuristic methods should be explored to preprocess the input text and normalize the spelling when necessary. This will help to reduce the noise that spelling variation introduces into the model. Furthermore, a simple rule-based morphological analysis that detects compounds and reduplication could ease the POS classification task when observing words that are not in the vocabulary.

As for taking another step to get more annotated data, perhaps the shortest route would be to use the current HMM model to tag the rest of the text on the SIL website. If the results obtained for text 2 generalize to the rest of the collection, the automatic tagging on them will require fewer corrections.

REFERENCES

[1] Cortegoso Vissio N., Zakharov V. A rule-stochastic hybrid POS-tagger for Sranan Tongo with minimal lexicon and training dataset. In: Proceedings of the International Conference «Corpus Linguistics-2021». Saint-Petersburg, Sofia-Press. 2021 (in print).

[2] Radke H. Niederländisch und Sranantongo in Surinamischer Online-kommunikation // Taal en Tongval. University Press, Amsterdam, 2017. Vol. 69. P. 113-136.

[3] Sebba M. Contact languages: pidgins and creoles. Palgrave Macmillan, 1997.

[4] Wortubuku fu Sranan Tongo. SIL International. URL: https://www.sil.org/resources/archives/13426 (accessed: 10.10.2021).

[5] Yakpo K., Bruyn A. Transatlantic patterns: The relexification of locative constructions in Sranan // Surviving the Middle Passage: The West Africa-Surinam Sprachbund / Pieter Muysken, Norval Smith (Eds.). De Gruyter Mouton, Berlin, 2015. P. 135–175.

[6] Wilner J. Wortubuku fu Sranan Tongo. Sranan Tongo-English Dictionary / John Wilner (ed.), Ronald Pinas, Lucien Donk, Hertoch Linger Arnie Lo-Ning-Hing, Tieneke MacBean, Celita Zebeda-Bendt, Chiquita Pawironadi-Nunez, Dorothy Wong Loi Sing. SIL International, 2007.5th ed.

[7] Wilner J. Wortubuku fu Sranan Tongo. Sranan Tongo-Nederlands Woordenboek / John Wilner (ed.), Ronald Pinas, Lucien Donk, Hertoch Linger Arnie Lo-Ning-Hing, Tieneke MacBean, Celita Zebeda-Bendt, Chiquita Pawironadi-Nunez, Dorothy Wong Loi Sing. SIL International, 2007. 5th ed.

[8] Nickel M., Wilner J. Papers on Sranan Tongo. Summer Institute of Linguistics, 1984. URL: https://archive.org/details/rosettaproject_srn_morsyn-1 (accessed: 05.04.2021).

[9] Winford D., Plag I. Sranan structure dataset // Atlas of Pidgin and Creole Language Structures Online / Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, Huber Magnus (Eds.). Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. URL: http://apics-online.info/contributions/2 (accessed: 03.10.2021).

[10] Jurafsky D. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition / Daniel Jurafsky, James H. Martin. Prentice Hall, New Jersey, 2008. 2nd edition.

[11] Rijksoverheid. Skowtu hori yu na ini a tori fu wan ordru fu den bakrakondre, nanga den tyari yu na skowt'oso noso wan tra presi pe den o yere yu. URL: https://www.rijksoverheid.nl/documenten/brochures/2014/07/01/u-bent-aangehouden-in-verband-met-een-europees-aanhoudingsbevel-en-meegenomen-naar-het-politiebureau-of-andere-verhoorlocatie.-wat-zijn-uw-rechten-sranan-tongo (accessed: 10.10.2021).

[12] MacBean G. A gridi frow fu fisman Albert. Institut voor Taalwetenschap (SIL). 1993. URL: http://suriname-languages.sil.org/Sranan/English/SrananEngLLIndex.html (accessed: 10.10.2021).

[13] Pinas E. San pesa ini Kaneri. Nieuwe Surinaamse Verhalen. Nieuwe Surinaamse verhalen. M. van Kempen (comp.). Uitgeverij De Volksboekwinkel, Paramaribo. 1986.

[14] Cortegoso Vissio N. A part of speech tagger for Sranan Tongo based on a Trigram Hidden Markov Model // GitHub repository. URL: https://github.com/nicolascortegoso/HMM-for-sranantongo (accessed: 10.10.2021)