Establishing the Syntactic Rules of the Kankanaey Dialect using RNN

Laurie Lynne F. Aspiras¹, Ibrahim F. Hanbal², Jeffrey S. Ingosan³, Kim Benedict P. Panlilio⁴
College of Information Technology and Computer Science, University of the Cordilleras, Baguio City 2600, Philippines.

Email: ¹ laurielaspiras@gmail.com, ² ibrahimhanbal@yahoo.com, ³ jsingosan@uc-bcf.edu.ph, ⁴ kimbpanlilio@gmail.com

Abstract. The diverse culture and ethnic groups in the Philippines creates a beautiful mixture of ideas, traditions, and practices but also makes it hard for researchers to keep track of them all. One integral part of any culture is language, with one of the most spoken languages in the Cordillera Administrative Region (CAR) being Kankanaey. Unfortunately, it has very little resources and documentation for it. This paper presents a corpus created for Kankanaey that contains 3412 words and was trained with a dataset containing 400 Kankanaey sentences in order to establish its syntactic rules. Data for the collected texts for Kankanaey were taken from public sources online and were organized into various categories based on the type of content. Training and testing was done to establish the syntactic rules using the Keras API. The rules were derived by having each word in the training sentences tagged with the corresponding POS tag. After tagging, the number of POS tags were then expanded to all possible combinations of the POS which resulted in the documenting of 1,722 syntactic rules for Kankanaey with the model having an accuracy of 64% when it was tested to identify the syntactic rules in 50 test sentences.

1. Introduction

Language is the core of every culture which an individual is able to create the meaning of human experience, thought, feeling, appearance, and behavior. In context, language is integrally intertwined with the concept of both the personal and the broader, societal levels as it is used to develop, strengthen, or even create a cultural, national, or other form of group identity [1]. One study found that as many as 6,000 languages worldwide are in danger of becoming extinct as the number of speakers for these languages are slowly dwindling with no new speakers being taught the language [2].

One way to address this threat is through the use of Natural Language Processing (NLP). NLP is a field concerned with the study of the syntax and semantics of a language [3, 4, 5, 6]. NLP is very important in today’s globalized world as people from different backgrounds, professions, cultures and countries communicate with one another. A better understanding of all these languages will allow people to better understand one another and master these languages and dialects for various reasons [7, 8]. The first step in understanding a language is learning its grammar, or in more technical terms, its syntactic rules. Having knowledge of syntactic rules is very important especially for children as it enables them to better understand and use more complex versions of these rules as they grow up [9]. Once one is familiar with the syntax of a language, the next step will be understanding the semantics of the language.
Studies have also been conducted in using NLP tools to learn the Filipino language. As it is considered the national language of the Philippines [1] and is spoken by more than 55% of the population and a large portion of written, visual and audio content is created every day using Filipino [10]. Due to the large number of languages and dialects in the Philippines, researchers have turned to using NLP tools to help them study and synthesize the various languages and dialects in the Philippines. One example of this is a study conducted in 2003 for English-Filipino machine translation system. The system took input for a statement in the source language and produced the corresponding target language translation of the statement using a combination of rule based and corpus-based approaches [11].

A similar study was conducted in 2015 where a group of researchers created a Filipino-Maranao translator using machine translation technology. The translator was able to convert text-to-text and text-to-speech for Filipino-Maranao and vice versa with a high degree of accuracy [12]. Documenting the syntax of dialects in the Philippines is important now more than ever especially since there are many languages in the Philippines that endangered due to the shrinking number of people speaking the language [13].

This paper focused on the first step of understanding a language. The first objective was to collect and categorize texts for the Kankanae dialect, one of the common dialects in the Cordilleras. The next objective is to tag several sentences in Kankanae to establish the rules of the dialect. The last objective was the review the recorded rules and to determine which rules are correct and which are outliers in the data.

2. Methods

The Keras API was used to collect, tag and process the text data for training. The Keras API was utilized for the study due to its implementations of many machine-learning algorithms, its ease of use and its fast prototyping tools [14, 15]. The Keras API also has many tools from other deep learning libraries that make it very useful for building NLP applications with one of these tools being Recurrent Neural Networks (RNN).

2.1. Building of the Dictionary

Dictionary refers to a large database of lexical and grammatical resources used for study and analysis in NLP [16]. The dictionary was built in using Microsoft Excel, a spreadsheet application, and texts were collected from various sources online such as blogs, personal websites, other research papers, and translated works. The collected text was then organized into nine different categories for easier sorting and analysis. The categories are: (1) books, (2) dictionaries, (3) news articles, (4) poems, (5) religious texts, (6) songs, (7) stories, (8) web sources and (9) others. The sentences and text that were used for training and testing the model were taken from these nine categories. A large number of the sentences that were used for training and testing came from the religious text category as this contained the largest amount of collected text. A database of words with their associated tags in Kankanae was created using the collected texts to be used for training and establishing the syntactic rules of the Kankanae language. The words were tagged using the Parts of Speech (POS) tags of the Penn TreeBank study [17, 18, 19, 20, 21] which was a study that has analysed over 13 million words in the English language and has been used as a foundation for creating language corpora for other languages such as Chinese [22], French [23] and Turkish [24]. The Penn TreeBank POS tags that were used consisted of 36 POS tags and 12 others for punctuations and other symbols. An additional tag labelled as “NF” was also added to the tagset for words that were in the training sentences but were not found in the database of tagged words.

The dictionary containing words with their associated tags in Kankanae was cleaned with duplicate, incorrect or misspelled words being removed [25, 26]. This resulted in a database containing 3412-tagged Kankanay words used for NLP analysis and training. The model was trained using 400 Kankanay sentences, which was used to establish the POS rules. After the POS patterns were compiled from these sentences, the patterns were then filtered so that only unique Kankanay POS patterns remained.
2.2. Training the Model

Text mining is the process of extracting information and patterns from text documents [27, 28]. Text mining involves making use of information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining in order to gain knowledge from a collection of text [29, 30]. Text mining was done to establish the syntactic rules of the Kankanaey dialect. The dictionary containing 3412-tagged words were used as basis to scan and analyse the 400 training sentences. The model only covered 400 sentences as the hardware used to run the model would run into GPU (Graphics Processing Unit) errors when a larger number of sentences were entered. Each word in each of the training sentences were tagged with the corresponding POS tag based on the database. The words were tagged like so:

```
PRP|MD PRP IN|DT JJ. -> A | denotes that the word is associated with multiple POS tags.
```

The number of POS tags were then expanded to all possible combinations of the POS. Using the above example; we would derive these possible sentence structures.

```
PRP PRP IN JJ.
PRP PRP DT JJ.
MD PRP IN JJ.
MD PRP DT JJ.
```

These POS patterns were then recorded, resulting in 6299 possible POS patterns and 3624 possible unique POS patterns. The Keras library's LSTM Sequential Model was used to turn the sentences into POS patterns. The LSTM Sequential Model is a RNN model that has LSTM cellblocks in place of standard neural network layers. These cells have various components such as the input gate, the forget gate and the output gate. The model was trained for 48 hours with 187 epochs, as further passes resulted in no change in the model. According to the Keras documentation, an epoch is an arbitrary cutoff, generally defined as "one pass over the entire dataset", used to separate training into distinct phases, which is useful for logging and periodic evaluation." The current rules in this state are not readable and usable for analysis as the POS tags was arranged in one continuous line of text. There was a need to further process the compiled patterns to turn them into readable and usable rules, a generator was created.

2.3. Establishing the Rules

A Keras generator was then created to write POS patterns based on the trained model. It was made to run for 40 hours and generated 10,182 lines of POS patterns. Of these 10,192 POS patterns, 1,722 were unique patterns. The model was then tested using 50 new sentences using the 1,722 generated POS patterns. The model attempted to identify if the POS pattern of the sentences are in the generated POS patterns. The model was able to identify the POS patterns of 32 sentences out of the 50 test sentences which resulted in an accuracy of 64%.

3. Findings

The results of training have provided 1,722 syntactic rules in the Kankanaey language. The recorded rules indicate that the syntactic rule “PRP SBAR RB UH VB MD” has the highest occurrence with a total frequency of 220. Twenty of the recorded rules during training can be found on Table 1. The model was then tested using 50 new sentences, which are different from the 400 training sentences, which resulted in 32 of the sentences having rules confirmed by the model and 18 having rules that were not recognized by the model. This resulted in the model having an accuracy of 64%.
4. Discussions
The results of the training indicate that 3412-tagged Kankanaey words was sufficient for training allowed for the generation of 1,722 unique POS patterns. The large number of POS patterns that came from the training was due to recording every possible combination of POS tags for the sentences. With a higher number of sentences, the model can be enhanced with a larger number of syntactic rules or higher frequencies for the recorded rules. The model was only able to read at most around 400 sentences due to limitations in the GPU used for training. A dedicated high end GPU can be utilized for future studies to allow for more sentences to be used for training and tagging. The training process can be further improved by adding a filter that will remove impossible or highly unlikely POS combinations when the POS patterns are being generated to better reflect real life sentences in Kankanaey.

Table 1. Twenty of the recorded syntactic rules and their frequency of the Kankanaey dialect

| #  | POS Tag Structure (Rules)                  | Frequency |
|----|------------------------------------------|-----------|
| 1  | PRP SBAR RB UH VB MD                    |           |
| 2  | DT PRP DT                               |           |
| 3  | PRP SBAR PRP MD VB MD                   |           |
| 4  | DT PRP MD PRP UH DT CC DT DT NN PRP MD |           |
| 5  | PRP MD DT PRP NN PRP UH DT CC DT SBAR NN VBD PRP UH DT PRP |           |
| 6  | IN DT NN RB UH VB MD PRP MD PRP UH DT IN DT NN IN PRP |           |
| 7  | DT NN PRP UH DT PRP MD CC PRP MD VBD DT NN IN PRP |           |
| 8  | PRP MD WRB DT NN IN PRP                 |           |
| 9  | PRP MD PRP CC DT SBAR PRP UH DT PRP MD RB UH VB MD PRP UH DT PRP UH DT NN |           |
| 10 | RB DT PRP PRP MD VB PRP IN DT PRP IN DT DT ADV |           |
| 11 | PRP MD VB CC DT DT SBAR IN DT NN       |           |
| 12 | CC DT SBAR                              |           |
| 13 | JJ NN CC DT DT NN PRP MD PRP UH DT     |           |
| 14 | JJ PRP MD CC PRP DT IN DT IN PRP IN IN PRP DT IN DT PRP MD IN DT |           |
| 15 | IN PRP DT IN PRP PRP UH DT PRP MD IN DT |           |
| 16 | CC DT SBAR DT PRP PRP CC PRP DT N DT   |           |
| 17 | PRP DT PRP PRP UH DT PRP MD IN DT CC VB IN PRP NN PRP UH DT CC DT NN PRP UH DT |           |
| 18 | PRP DT NN PRP MD IN DT                 |           |
| 19 | RB DT PRP PRP MD VB PRP IN DT PRP IN DT DT |           |
| 20 | PRP MD DT PRP DT IN DT                 |           |

The model was moderately successful in identifying the POS patterns of the sentences when it was tested, as its accuracy was 64%. As compared with the results of Gulordava et al., who obtained a prediction accuracy of 81% on English sentences from their test corpus. Considering that the English language is well documented while the Kankanaey dialect has a scarce of formally written documents - due to the speakers preferring oral tradition - the results were considered as moderately successful [31]. This suggests that the model is lacking in data and required more sentences for training in order to be effective. One possible reason for this is that the number of sentences used for training and testing were
not sufficient to establish a large enough collection of syntactic rules for the Kankanaey language. Another limitation is that many of the sentences used for training also came from religious texts since they were the most prevalent, which have a tendency to have long and unique sentences structures which led to the model not being able to identify the more common sentences structures. Having a larger and more diverse set of sentences for training along with more generated POS patterns can increase the overall accuracy of the model when presented with new sentences to identify.

5. Conclusions
There are not a lot of studies or formal documents regarding the syntactic rules of the Kankanaey dialect, the researchers therefore conclude that this study would be one of the initial attempts to fill the distinct lack of formal documents. With the result of the study presenting that Keras API can be used to generate commonly used syntactic rules, future researchers may perform an in-depth comparison between the results of this study with those that make use of widely-known Natural Language Processing Python libraries such as NLTK, CoreNLP, Gensim, spaCy, and polyGlot.

The model can be further improved by expanding the number of sentences used for training and testing. This will increase the number of recorded rules, as this will allow for a larger amount of combinations for the POS and provide a higher frequency for the already recorded rules. A more powerful system can be utilized to train the model with a larger number of words and sentences. The recorded POS patterns can be further expanded by increasing the number of sentences that were used for training the model. Other sources of words and sentences such as village elders, local songs, hymns and stories and written works (such as old books) can be considered as another source to further expand the data of the corpus and improve the diversity of the rules in the corpus.

References
[1] Catama B et al 2018 Challenges in maintaining a heritage language Int. J. of Research Studies in Language Learning 2018 vol 7 no 1 pp 39-61
[2] Krauss M 1992 The world's languages in crisis Language 68(1) pp 4-10 https://doi.org/10.1353/lan.1992.0075
[3] Chowdhury G 2003 Natural language processing Annual Review of Information Science and Technology 37 pp 51-89 ISSN 0066-4200
[4] Zitouni I 2011 Introduction to Arabic Natural Language Processing Computational Linguistics 37 pp 623-5
[5] Lewis D and Sparck Jones K 1996 Natural Language Processing for Information Retrieval Communications of the ACM 39 DOI:10.1145/234173.234210
[6] Loper E and Bird S 2002 NLTK: the Natural Language Toolkit CoRR cs.CL/0205028 DOI: 10.3115/1118108.1118117
[7] David M K, Cavallaro F and Coluzzi P 2009 Language Policies - Impact on Language Maintenance and Teaching: Focus on Malaysia, Singapore, Brunei and the Philippines The Linguistics J. September
[8] Domingo D R 2016 Content Area Effectiveness: English Vs Filipino Medium of Instruction PEOPLE: Int. J. of Social Sciences 2(1) pp 1514-29
[9] Menyuk P 1964 Syntactic Rules Used by Children from Preschool Through First Grade Child Development 35(2) pp 533-46 doi:10.2307/1126717Fds
[10] Rau D V 1992 Language Contact in the Philippines Philippine J. Of Linguistics vol 23 no 1 and 2 June and December
[11] Roxas R E, Borra A, Cheng C, Lim N R, Ong E and Tan M W 2008 Building language resources for a multi-engine machine translation system in Language Resources and Evaluation vol 42 pp 183-95 (Berlin: Springer)
[12] Fallan J Q, Rasuman M A T, Sacil A V and Domingo I V R 2015 Filipino- maranao bi-directional language translator with text-to speech synthesizer Int. J. of Information Technology and Business Management 28th Feb 2015 vol 34 no 1
[13] Headland T N 2003 Thirty Endangered Languages in the Philippines Work Papers of the Summer Institute of Linguistics, University of North Dakota Session vol 47

[14] Chauhan K and Ram S 2018 Image Classification with Deep Learning and Comparison between Different Convolutional Neural Network Structures using Tensorflow and Keras Int. J. of Advance Engineering and Research Development vol 5 Issue 02 February 2018

[15] Vidnerova P and Neruda R 2017 Evolving Keras Architectures for Sensor Data Analysis Proc. of the Federated Conf. on Computer Science and Information Systems pp 109–12

[16] Van Der Eijk P, Bloksma L and Van Der Kraan M 2003 Towards Developing Reusable NLP Dictionaries COLING 1992 Proc. of the 14th Conf. on Computational linguistics vol 1 pp 53-9

[17] Taylor A, Marcus M and Santorini B 2003 The Penn Treebank: An Overview Treebanks: Text, Speech and Language Technology ed A Abeillé vol 20 (Dordrecht: Springer)

[18] Chen J and Shanker V K 2004 Automated Extraction of Tags from the Penn Treebank New Developments in Parsing Technology: Text, Speech and Language Technology vol 23 ed H Bunt et al (Dordrecht: Springer)

[19] Marcus M et al 1994 The Penn Treebank: Annotating predicate argument structure Proc. of the 1994 Human Language Technology Workshop pp 110–5

[20] Marcus M, Ann Marcinkiewicz M and Santorini B 2002 Building a Large Annotated Corpus of English: The Penn Treebank Computational Linguistics - Special issue on using large corpora: II vol 19 Issue 2 June pp 313-30

[21] Fleming A J, Kingsbury P T, Palmer M and Marcus M 1998 Adding Semantic Annotation to the Penn TreeBank Proc. of the second Int. Conf. on Human Language Technology Research 2002 pp 252-6

[22] Xue N, Xia F, Chiou F-D and Palmer M 2005 The Penn Chinese TreeBank: Phrase structure annotation of a large corpus The J. of Natural Language Engineering vol 11 Issue 2 June 2005 pp 207-38

[23] Abeillé A, Clément L and Toussenel F 2003 Building a Treebank for French Treebanks: Text, Speech and Language Technology vol 20 ed Abeillé A (Dordrecht: Springer)

[24] Oflazer K, Say B, Hakkani-Tür D Z and Tür G 2003 Building a Turkish Treebank Treebanks: Text, Speech and Language Technology vol 20 ed Abeillé A (Dordrecht: Springer)

[25] Batista G, Prati R and Monard M-C 2004 A Study of the Behavior of Several Methods for Balancing machine Learning Training Data SIGKDD Explorations 6 pp 20-9 DOI: 10.1145/1007730.1007735

[26] Guyon I, Matic N and Vapnik V 1994 Discovering Informative Patterns and Data Cleaning Advances In Knowledge Discovery And Data Mining pp 145-56

[27] Fayyad U, Piatesky-Shapiro G and Smyth P 1996 From data mining to knowledge discovery: An Overview Advances in Knowledge Discovery and Data Mining ed U Fayyad et al (Cambridge, Mass.: MIT Press) pp 1-36

[28] Simoudis E 1996 Reality check for data mining IEEE Expert 11(5)

[29] Tan A H et al 1999 Text mining: The state of the art and the challenges Proc. of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases vol 8 pp 65–70

[30] Jacobs P S 2002 Joining Statistics with NLP for Text Categorization ANLC ’92 Proc. of the third Conf. on Applied natural language processing pp 178-85 DOI: 10.3115/974499.974532

[31] Gulordava K, Bojanowski P, Grave E, Linzen T and Baroni M 2018 Colorless green recurrent networks dream hierarchically Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies vol 1 pp 1195–205