Establishing the Syntactic Rules of the Kankana-ey Dialect using TensorFlow

Angeli A. Cabilitazan¹, Jeffrey S. Ingosan², Jason Irabagon³, Philip Irving G. Jacinto⁴, Jan Jefferson O. Lam⁵

College of Information Technology and Computing Science, University of the Cordilleras, Baguio City, Philippines.

Email: ¹g.anncabilitazan96@gmail.com, ²jsingosan@uc-bcf.edu.ph, ³trize8081@gmail.com, ⁴philsicom@gmail.com, ⁵janjeffersonlam@gmail.com,

Abstract. Kankana-ey is a widely used dialect in the northern region of the Philippines. Unfortunately, there are no documented studies on the syntactic rules of this dialect. This study explored the development of a corpus for the Kankana-ey dialect. Further, the corpus was then used to establish the syntactic rules of Kankana-ey. A Kankana-ey version of the bible, dictionaries, news articles, songs and various online resources were used to collect words for the corpus of the Kankana-ey dialect. These identified words were also tagged using the parts of speech tags of the Penn TreeBank. Using the corpus and TensorFlow, 320 Kankana-ey sentences were analysed to determine the syntactic rules. In addition, 80 sentences were used to test the accuracy of the identified rules. At the end of the study, the created corpus has 3,412 tagged Kankana-ey words, while the analysis of the syntactic rules resulted to 1,722 rules. Testing also showed a 60% accuracy of the syntactic rules. In conclusion, the high number of identified rules from the 320 sentences was due to multiple Kankana-ey words having different possible tags. This also resulted to the low accuracy of the syntactic rules.

1. Introduction
One particular dialect known in the Cordillera Administrative Region is the Kankana-ey. Kankana-ey is a South-Central Cordilleran dialect spoken on the northern part of Luzon in the Philippines. On the 1990 to 2003 census, there are around an approximate of 240,000 native speakers cited and is also closely related with the Ilocano dialect. Alongside Ilocano, it is widely used in the Mountain Province and northern part of Benguet Province. Visitors of these provinces, such as linguists, tourists, researchers, teachers, advertisers, journalists and reporters, find it hard to communicate with the locals especially when these locals only speak Kankana-ey.

The only means to understand each other is through translators. However, human translators may not be readily available and there are still no available portable and electronic translators. The study on syntactic rules of the Kankana-ey language is also needed to be able to identify the rules that can be used for electronic translators [3,7,10].

Natural Language Processing (NLP) during the last two decades has made tremendous leaps in terms of language processing such as syntax analysis. However, current technologies would still require further studies of languages and dialects used today [3].

Researches stated the need to focus on studies in NLP. Knowledge acquisition, interaction with multiple underlying systems and partial understanding has been cited as each addresses a critical challenge in using the techniques developed to address problems encountered in NLP [3,7,11,12].
It is important to note that in the study of language, such as English, grammatical or syntactically valid does not imply sensible, or semantically valid statements.

This study aimed to do a basic syntactic analysis of the Kankana-ey dialect. Specifically, this study sought the creation a corpus for the Kankana-ey dialect; use the corpus to identify the syntactic rules; and determine the accuracy of the identified syntactic rules.

2. Review of Related Literature

In learning a language, especially during the early years, data that are not structured in any way are studied. After amassing this massive information, the learner processes and gives meaning to these based on the implications as to where these can be applied. Machine in turn also learn the same as people do, they follow a similar path where machine processes that unstructured data into a somewhat meaningful information and determines relevant connections to comprehend the context [12, 14, 15]. In the most recent decade, a sensational move in the NLP research has prompted the pervasiveness of extremely enormous scale uses of factual techniques, for example, AI and information mining. Normally, this additionally opened the path to learning and streamlining techniques that comprise the center of current AI, most eminently hereditary calculations and neural systems [2, 7].

As mentioned by D.B. Jones and H. Somers, classifiers could be applied on the training and testing of data, as there are several classifiers available. In tagging parts of speech, several techniques could be used and applied. Existing programs pick the word from the corpus and based on the most probable grammatical category a grammatical pattern is then created giving sense to the sentence and allowing a deeper understanding on how the sentence is created and used [16, 17].

In the study of corpus-based translation, the first thing to note is what is the meaning of corpora and encoding [16]. Part of the concern of the study is the availability of the data whether printed or in electronic form. With the available data, annotations are made in order to create a stable, flexible and accessible corpus resource to be used in language/translation studies [18, 19].

Content generated today on the web has substantially reached a point where a need for sensible computation and information aggregation has reached a critical point as the trend today by big data experts and the growth of new fields in data science. As NLP research shifts from lexical semantics to compositional semantics, it offers various new insights in the development of new NLP technologies [13, 14, 17].

Using RNN in the study, tags where given to the words using the Penn Treebank Parts-of-Speech tags. With that the creation of grammar rules was done, and the model was trained predicting the next set of P.O.S. based on the grammar rules created from the number of occurrences it made during the training/learning phase [18, 19].

In training a model, recently in the areas of machine learning and NLP a new field of study called learning to rank has emerged. In some sense it refers on how to rank the techniques used in machine learning and NLP based on practicality of the techniques used and applied in practice. This applies to the study, where a need to create a ranking system is thought of. Using supervised learning training data were used as input and allows the system to generate a ranking model based on that training data. It would then help improve the performance of the algorithm used [3, 15, 16, 19, 20].

This will help non-expert clients effectively recover data without the need to become familiar with any extra coding languages or projects. The examination information incorporates regular language messages gathered from different sources, both on the web and disconnected, to cover on both formal and semi-formal language level. By utilizing normal language handling and procedures, for example, semantic examples, philosophy, and intuitive discussion framework, this model can investigate the culmination and significance of common language explanations just as it enables clients to alter the fragmented or flawed proclamations, and improve the model by including new words, sentence punctuation and semantic examples for increasingly exact outcomes [10].

C. Tapsai [9] mentioned in their research the proposed usage of several algorithms in determining the relativity of the word used in the sentence to the meaning it would convey when used in a sentence using Word Usage Frequency (WUF). This is to determine how words are used and how frequent they appear in a sentence allowing a more reliable dictionary which significantly raises the efficiency and
precision all the way reducing the dictionary size by a significant amount of margin. Similar to this study is the usage of WUF in determining the frequency of words used allowing one to determine how often this kind of sentence structure is used on the Kankana-ey dialect [5, 17, 20].

3. Methodology
The study was conducted with the use of TensorFlow. The researchers used Recurrent Neural Networks (RNN) to accomplish the needed task for identifying the syntactic rules or pattern in the Kankana-ey dialect.

3.1. Creating the Corpus
Various texts from the internet such as dictionaries, poems, and articles were collected. The texts that were collected were categorized as (1) books, (2) dictionaries, (3) news article, (4) poems, (5) religious text, (6) songs, (7) stories, (8) web sources, and (9) others. Unique words from these collected texts were listed using Microsoft Excel. Cleaning of data was also done by removing redundant data and checking for the spelling of the words. The Parts of Speech tags from the Penn TreeBank were used to tag the listed words. It is also important to note that several words were tagged under several parts of speech. To further validate the corpus, Kankana-ey speakers were asked to review the corpus.

![Figure 1. Sample sentence with its corresponding POS Tag Structure](image)

3.2. Model Training
A total of 400 sentences were taken from the same texts used to create the corpus of Kankana-ey words. 80% of these sentences were used to determine the syntactic patterns of the Kankana-ey language. RNN was used through TensorFlow to identify these syntactic patterns. The remaining 20% of the sentences were used to check the accuracy of the identified syntactic patterns. Figure 1 shows a sample syntactic rule.

Table 1. A slightly more complex table with a narrow caption.

| Syntactic Rule | Frequency |
|----------------|-----------|
| VB – NN – JJ   | 300       |
| NN – VB – JJ   | 280       |
| PRP – VBZ – NNZ – IN – DT – NN | 243       |
| PRP – SBAR – PRP – MD – VB – MD | 184       |
| RB – VB – JJ   | 179       |
| NN – JJ – VB   | 188       |
| NN – VB – RB   | 187       |
| RB – DT – PRP – PRP – MD – VB – PRP – IN – DT – PRP – IN – DT – DT | 155       |
| VB – JJ – RB   | 149       |
| PRP – SBAR – PRP – MD – VB – MD | 131       |
| IN – DT – NN – RB – UB – VB – MD – PRP – MD – PRP – UH – DT – DT – IN – DT – NN – IN – PRP | 107       |
| PRP – MD – DT – PRP – DT – IN – DT | 90        |
| CC – DT – SBAR – DT – PRP – PRP – CC – PRP – DT – N – DT | 84        |
| CC – DT – SBAR | 50        |

4. Findings
A total of 3,412 words were identified and tagged during the creation of the Kankana-ey Corpus. These tagged words came from the texts that were collected. During the extraction of the syntactic rules, 1,722
rules were identified from the 320 evaluated sentences. Table 1 shows examples of the identified syntactic rules. The sentence pattern of VB-NN-JJ has the highest frequency of 300.

The researchers used 80 sentences for testing. 320 sentences were also used to train the model. The accuracy of the model is only 60%.

5. Discussions
The 3,412 words of the created Kankana-ey corpus has words with more than one tag. This is due to the possibility of a word to be used as different parts of speech. For example, a word can be a verb and another part of speech depending on how this is used. This is also common in other dialects or languages such as English. In English for example, fly can be used as a verb or a noun.

From the training process to extract the syntactic rules, a total of 1,722 rules were generated from just 320 sentences. This was encountered due to the fact that words from the corpus could have different parts of speech. This resulted to a sentence having different syntactic rules. In reality, the 1,722 rules are deemed valid on the basis of the arrangement of words but not necessarily on the semantics. The accuracy of the testing at 60% and can still be improved by increasing the training and testing data.

6. Conclusions
The corpus of the Kankana-ey dialect is composed of words that can be used as different parts of speech. This can thus result to a corpus with a greater number of words than the actual number of uniquely spelled words. As a consequence, the use of this corpus on the extraction of the syntactic rule of a sentence will lead to at least one rule. Although, some of these rules may not be acceptable if semantics is to be considered. As a general conclusion a certain level of semantic analysis may be useful in the study of syntactic rules.

References
[1] Zaccone G 2016 Getting Started with TensorFlow (Birmingham: Packt Publishing)
[2] Gelbukh A 2005 Natural Language Processing Fifth Int. Conf. on Hybrid Intelligent Systems p 1 DOI:10.1109/ICHIS.2005.79
[3] Zong Z and Hong C 2018 On Application of Natural Language Processing in Machine Translation 2018 3rd Int. Conf. on Mechanical, Control and Computer Engineering (ICMCCE) pp 506-10 DOI: 10.1109/ICMCCE.2018.00112
[4] Xihong Wu, Meng Zhang and Xiaojun Lin 2011 Parsing-based Chinese word segmentation integrating morphological and syntactic information 2011 7th Int. Conf. on Natural Language Processing and Knowledge Engineering pp 114-21 DOI:10.1109/NLPKE.2011.6138178
[5] Allen J L 2003 Operators and Clause Structure in Kankana-ey Philippine J. of Linguistics
[6] Zhao L and Li F 2008 Statistical Machine Learning in Natural Language Understanding: Object Constraint Language Translator for Business Process 2008 IEEE Int. Symp. on Knowledge Acquisition and Modeling Workshop pp 1056-9 DOI: 10.1109/KAMW.2008.4810674
[7] Kumar P, Srivastava S and Joshi M 2015 Syntax directed translator for English to Hindi language 2015 IEEE Int. Conf. on Research in Computational Intelligence and Communication Networks (ICRICCN) pp 455-9 DOI: 10.1109/ICRICCN.2015.7434282
[8] De Silva D et al 2008 Sinhala to English Language Translator 2008 4th Int. Conf. on Information and Automation for Sustainability pp 419-24 DOI: 10.1109/ICIAFS.2008.4783983
[9] Lane R and Bansal A 2017 An Adaptive Machine Translator for Multilingual Communication 2017 IEEE 26th Int. Conf. on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) pp 21-3 DOI: 10.1109/WETICE.2017.53
[10] Tapsai C 2018 Information Processing and Retrieval from CSV File by Natural Language 2018 IEEE 3rd Int. Conf. on Communication and Information Systems (ICCIS) pp 212-6 DOI: 10.1109/ICOMIS.2018.8644947
[11] Bates M and Weischedel R M 1993 Challenges in Natural Language Processing (Cambridge: Cambridge Press)
[12] Ishaq M 2019 What are some of the challenges we face in NLP today [online]
https://medium.com/datadriveninvestor/what-are-some-of-the-challenges-we-face-in-nlp-today-2e9d94da1f63

[13] Cambria E and White B 2014 Jumping NLP Curves: A Review of Natural Language Processing Research *IEEE Computational intelligence magazine* vol 9 no 2 pp 48-57 DOI: 10.1109/MCI.2014.2307227

[14] Sag I A, Baldwin T, Bond F, Copestake A and Flickinger D 2002 Multiword Expressions: A Pain in the Neck for NLP *Proc. of Int. Conf. on Intelligent Text Processing and Computational Linguistics* ed A Gelbukh (Berlin, Heidelberg: Springer) pp 1-15

[15] Li H 2011 *Learning to Rank for Information Retrieval and Natural Language Processing* (California: Morgan Claypool)

[16] Jones D B and Somers H L 1997 *New Methods in Language Processing* (London: UCL Press)

[17] Zanentin F 2011 *Hardwiring Corpus-Based Translation Studies: Corpus Encoding Corpus Based Translation Studies* (London: Bloomsbury Academics)

[18] Yao K, Zweig G, Hwang M, Shi Y and Yu D 2013 Recurrent Neural Networks for Language Understanding *Int. Symp. on Computer Architecture* DOI: 10.13140/2.1.2755.3285

[19] Arenas M, Maturana F, Riveros C and Vrgoc D 2016 A framework for annotating CSV-like data *Proc. of the VLDB* 9 11 pp 876-87

[20] Tapsai C, Meesad P and Haruechaiyasak C 2016 TLS-ART: Thai Language Segmentation by Automatic Ranking Trie *The 9th Int. Conf. Autonomous Systems*

[21] Marcus M P, Santorini B and Marcinkiewicz M A 1993 *Building a large annotated corpus of English: the Penn Treebank* Retrieved from https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html

[22] Granger S 2004 Computer learner corpus research: Current status and future prospects *Applied Corpus Linguistics: A Multidimensional Perspective* vol 52 pp 123–45

[23] Chun S 2011 Learners as corpus researchers *Foreign Languages Education* vol 18 no 3 pp 73-103

[24] Mason O 2008 Developing Software for Corpus Research *Int. J. of English Studies* vol 8 no 1 pp 141-56 DOI: 10.6018/ijes.8.1.49141

[25] Granger S, Gilquin G and Meunier F 2015 Introduction: Learner corpus research – past, present and future *Cambridge Handbook of Learner Corpus Research* ed S Granger *et al* (Cambridge: Cambridge University Press) pp 1-5