ABSTRACT
Automated syntactic parsing of Philippine languages could be foundational to future machine translation systems. Rule-based systems for Philippine languages have typically not reached a level of wide coverage of language phenomena. The syntax parsing system described here uses the PAWS (Parser and Writer of Syntax) expert system to generate phrase structure rules. After customized rules common to most Philippine languages were added in the process of bringing a training data set up to 100% parsing, the auto-generated phrase structure rules were able to produce a correct parse for 81% of sentences in two Ayta Abellen native authored running texts. The customizations made for the training set helped further the development of the PAWS expert system for use with Philippine languages. The 81% parsing rate is significant in that it represents a wide range of coverage for a rules-based system.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Language parsing and understanding.

General Terms
Languages.

Keywords
Syntax parsing, Rules-based approach, Expert-system.

1. INTRODUCTION
In the 1980s various syntax parsing frameworks were developed such as LFG (Lexical Functional Grammar), FUG (functional unification grammar), DCG (definite-clause grammars), GPSG (generalized phrase structure grammar), and DIAGRAM. An underlying element in each of these was the concept of feature unification and all of these frameworks in a way influenced the development of PATR-II which Shieber calls "the simplest of the unification-based formalisms."[1] In the late 1990s the PATR-II formalism was implemented for the personal computer with PC-PATR by McConnel.[2]

The PAWS Starter Kit [3] developed by Dr. Cheryl A. Black and Dr. H. Andrew Black utilizes the PATR-II formalism and outputs phrase structure rules that work with the PC-PATR parser. PAWS does not do analysis in the same sense as some other NLP tools which aim to extract information about the syntax of a language by examining large text corpora. Rather, it is an expert system that gives the user explanations about various syntactic structures before asking the user to make choices about how the language works with respect to those structures, such as whether certain elements are separate words or affixes in the language and where an element occurs with respect to the head of the phrase. It also guides the user in marking features in the lexicon. PC-PATR implements a unification model which enables features to percolate up through the morphological parsing process so that they may become features of the phrase structures that are referenced in rules. Unification then enables PC-PATR to disambiguate wordforms for which the morphological parser produces two parses with different syntactic features.

PAWS generates articulated phrase structure rules according to PC-PATR specifications. A sample of the format used for rules by PC-PATR is shown below where a topic phrase (TopicP) is defined as a topic marker (TopicM) followed by a determiner phrase (DP). The lines that follow define the constraints on the arguments, namely that the head of the Topic phrase is the head of the DP phrase, that the DP head must be negative for the feature focus, negative for WH, negative for temporal, but that the TopicP must have the feature topic-marked.

rule {Topic Phrase}
TopicP = TopicM DP
<TopicP head> = <DP head>
<DP head type focus> = -
<DP head type wh> = -
<DP head type temporal> = -
<TopicP head type topic-marked> = +
Complement Clauses
Questions
Relative Clauses
Adverbial Clauses
Negative Constructions
Coordination Constructions
Focus and Topic Constructions
Exclamations and Greetings

At the end of each section of syntax questions there are instructions regarding “features” (eg. transitive, dative, etc.) that need to be marked on specific lexemes in the lexicon. When the questions for each area of syntax were answered and the lexicon correctly marked with features, the PSR grammar generated by PAWS and the Ayta Abellen lexicon were used by PC-PATR to perform syntactic parsing. A sample parse tree of one of the sentences used in the test is shown below.

The parser was first run on a training set of 40 sentences as the phrase structure rules generated by PAWS were extended and customized to handle the training set. Then the parser was run on a 49 sentence test from the text of two native authored Ayta Abellen texts. The results are detailed in section 6 below.

3. Formal Syntax for Philippine languages

Syntactic parsing in general is a vast field of study with special complications for Philippine-type languages. Most of the recent research on formal syntax has been done on Tagalog (eg. Kroeger [4], Maclachlan [5], Rackowski [6], Aldridge [7], and Sabbagh [8]), although work has also been done on Cebuano by Bell [9] and Kampampangan in Mirikitani [10]. In surveying the conclusions of formal syntax research on Philippine languages, I consider Kroeger, Bell, and Mirikitani as representative.

Kroeger, Bell, and Mirikitani have each described the respective Philippine languages as non-configurational and lacking a traditional VP. It is helpful to see how the basic sentence structures of these languages have been analyzed. Kroeger described the basic Tagalog sentence with this tree:

For each of these Philippine languages something other than a traditional VP is proposed. In a traditional VP there is a close relationship between the verb and the object. But Philippine language syntax is different as even the identification of what constitutes the subject is problematic and has been debated by many linguists. What is clear is that the verb normally precedes one or more noun phrases and hence the notion that Philippine languages are VSO. With this background it is possible to see that the sample parse tree generated below for a transitive Ayta Abellen sentence most closely resembles Bell’s model tree for Cebuano.

4. Syntax parsing challenges

4.1 Case, Voice and Grammatical Relations

PAWS queries the user about the case system of the language. Since linguists have come to varying conclusions about whether Philippine languages are more nominative-accusative or
ergative-absolutive with regard to case system, this question is not easy to answer for Ayta Abellen.

PAWS assumes that the grammatical relations are SUBJ, Direct Object (DO) and Indirect Object (IO). Selecting a straight Nominative-accusative case system in PAWS will not work because these grammatical relations don't have a one-to-one correspondence with these case markers. Voice alternations on the verb cause the labels to not work for all sentence types. A solution to this is to use Kroeger's analysis where the nominative marked argument (ang in Tagalog) is regarded as the subject and other arguments are labelled OBJ1 and OBJ2. This solution requires customization to the auto-generated PAWS rules because a different set of feature labels is needed: SUBJ, OBJ1, OBJ2. An Ayta Abellen grammar file has been developed following this model.

The other option is to choose Ergative-absolutive. This solution almost works except for the case of transitive active voice sentences where the arguments are reversed. With a customized rule to handle this special case, the original grammatical relations labels of PAWS (SUBJ, DO, IO) can be retained. A set of phrase structure rules has also been created following this model.

4.2 2nd Position Clitics
Ayta Abellen, like Tagalog, has pronoun and particle clitics that occur in what is called the second position of the sentence, meaning that they usually attach to a sentence initial verb or adverb. In reference to this phenomenon in Tagalog, Anderson has said “Tagalog clitics pose problems of varying severity for an account based on purely syntactic mechanisms” [11]. The second position is problematic to almost all theories of syntax as the positions of the arguments can move based on whether they are full noun phrases or pronouns. This can be seen below where the direct object occurs after the actor when the object is a full noun phrase but occurs before the actor when referenced with a pronoun.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibiyay nan Pabling ye lanom kanan Emilio give GEN Pabling NOM water DAT Emilio</td>
<td>Pabling will give the water to Emilio.</td>
</tr>
<tr>
<td>Ibiyay na yan Pabling kanan Emilio. Give 3S 3S Pabling DAT Emilio</td>
<td>Pabling will give it to Emilio.</td>
</tr>
</tbody>
</table>

Normally the noun phrase arguments occur after the verb, but with a sentence-initial adverb, pronominal arguments occur in the second position, in front of the verb seen below.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tampol na ibiyay ye lanom kanan Emilio, quickly 3S give NOM water DAT Emilio</td>
<td>Quickly he will give the water to Emilio.</td>
</tr>
</tbody>
</table>

If there are two pronouns with a sentence-initial adverb, both occur before the verb.

This use of the second position in the sentence is not limited to pronouns. Some clitics or particles also occur in second position. This second position is problematic for syntactic parsing and many just choose to call it a special case not handled by current syntactic theories. Since the goal of my research is not to make general statements about the underlying syntax structures but rather to determine what parts of the language could be parsed, custom rules were created to handle these preverbal possibilities. Because of the frequency of pronouns and particles in the data this seemed a better solution than regularizing all the data to full noun phrases and eliminating all particles.

4.3 Topicalization
One unique feature of Ayta Abellen is the use of left dislocated topic phrases which are marked with hiyay or demonstratives. Topicalization occurs frequently as can be seen in Stone [12] where a twenty sentence natural text had seven instances of topicalization. Because the topic is marked explicitly with a separate word and the position of the topic phrase is always before the verb and bounded by a comma, this phenomena in Ayta Abellen is relatively easy to model using the custom rules listed below:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TopicM = Dem Linker</td>
<td></td>
</tr>
<tr>
<td>TopicP = (Deg) TopicM VP</td>
<td></td>
</tr>
<tr>
<td>TopicP = TopicM IP</td>
<td></td>
</tr>
<tr>
<td>TopicP = TopicP_1 (DP) Conj TopicP_2</td>
<td></td>
</tr>
<tr>
<td>VP = TopicP</td>
<td></td>
</tr>
</tbody>
</table>

5. Customized rules
The PAWS starter kit, using the data supplied by the user, generated 121 rules for Ayta Abellen. In the process of bringing the training set of data up to 100% parsing capacity, 28 custom rules were written. These custom rules were created to handle language specific structures that were not included with the original rules generated by PAWS. Of those 28 rules, 11 were related to accounting for linkers, 9 for second position clitics, 6 for topic constructions, 1 for adjoining the case markers to DP, and 1 for handling passive/stative constructions.

One of the first discoveries of a feature not covered in the original set of rules generated by PAWS was the a linker which connects a variety of constructions such as modifiers to nouns, degree words to modifiers, modifiers to modifiers, etc. This linker is not optional and the lack of an account for it in the phrase structure rules blocked all adjective phrases from being parsed. In all, eleven custom rules were created and appended to the end of the grammar file to account for the linker.

Custom rules were added to handle the second position pronouns and clitics. The basic approach taken in these nine rules is to place manner adverbs and negative auxiliaries in the inflection position I and allow pronouns and clitics to attach to the head. When there is no adverb or negative auxiliary present...
then the pronouns and clitics remain in their normal position and adjoin to the verb.

Five custom rules were added to handle various topic constructions. The first rule added the specification that a demonstrative with a linker could also be a topic marker. Various nominalized elements that could occur inside a topic phrase were added such as a VP or an IP or conjoined topics. Equative sentences with a topic marker were handled with a rule to allow the topic phrase to be equated with the VP.

Custom rules were also written for case markers. The case markers could have been analyzed as prepositions resulting in PP arguments. But since the case markers are really proclitics, it is more optimal to create a custom rule that will adjoin the case marker to DPs rather than the default PP. Another rule was added to accomplish the same for nominalized VPs with a case marker, that they would be parsed as VPs also.

A custom rule was added for passive/statives so that they will not be parsed as actor voice. Since passive/stative verbs have an argument structure similar to that of object voice, it was best for the custom rule to link these sentences to the object voice rule.

The Ayta Abellen language uses pauses in natural speech to make preverbal phrase boundaries explicit. This was symbolized in the input text with commas so a feature “comma” was added to the word that comes before the comma. The comma feature was then used to constrain other rules in the grammar so that the syntactic parser could recognize phrase boundaries. So, while this was not done with a separate custom phrase structure rule, it was added as a custom constraint to several existing rules.

6. Results on natural texts

After the syntactic parser was working for all sentences in a training set of 40 sentences, the parser was tested on a natural, running text. To do this evaluation I chose two native authored texts which were different than the 40 sentence example set. One was a 21 sentence narrative about a hunting experience told by a man from the northern part of the language area. The other was a 28 sentence narrative told by a man in the central part of the language area where he recounts the traditional story of the Tower of Babel as he heard it from his father. So, a total of 49 sentences were used.

As I began testing the syntactic parser with natural texts, I found that there were phrase structure rules generated to address certain constructions, but the sentences containing those constructions were not parsing for various reasons. For example, there was one place in the natural texts where the speaker added material to explain the meaning of a word for the sake of the audience. This addition was semantically and syntactically redundant. Removing the phrase did not affect the grammaticality or meaning of the sentence and made the sentence better conform to a written style.

The quote margin wana ‘he said’ can occur both quote initial and quote medial. Ayta Abellen speakers frequently use the quote margin for the discourse function of reminding the listener who is talking and quote medial quote margins are not predictable. Because of its random position, several rules would be needed to try to predict its location inside a quotation. For this reason quote medial margins were removed and the quotation was split off into another sentence.

Ayta Abellen has a particle kano that can roughly be translated ‘it is said’. It normally occurs in the second position of the clause but like Walrod [13] noted about the use of kano in Gadang, its distribution can best be analyzed from a discourse perspective rather than syntax. The frequency of its use can be related to peak development in both Gadang and Ayta Abellen as in Stone [14]. For this reason, it can sometimes occur at the end of a sentence. Because of this irregular positioning, the particle kano was deleted from the source texts.

The last adjustment made was for the vocative anak ko ‘my child’. Because a vocative like this could be placed almost anywhere in the sentence, it was decided to not attempt coverage for vocatives in the syntactic parser and it was removed from the source text.

Four different tests were run on the PAWS generated rules. The first test used the original rules from PAWS with the original, unadjusted source text. The second test used the original source text but used the custom rules added by the developer. The third and fourth tests used the adjusted source text. Predictably the test run with the adjusted source text and including the custom rules produced the highest percentage of sentences parsed, 81% with the lowest parsing percentage of 38% for the original source text and the original PAWS generated rules. The results are shown below.

Table 1. Percentage of sentences parsed on native authored running texts (49 sentences)

<table>
<thead>
<tr>
<th></th>
<th>Original source text</th>
<th>Adjusted source text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original PAWS rules</td>
<td>38%</td>
<td>65%</td>
</tr>
<tr>
<td>Original + customized rules</td>
<td>60%</td>
<td>81%</td>
</tr>
</tbody>
</table>

It should be mentioned that these kinds of results are significant for Philippine languages. Previous research by others using a rule-based approach for Philippine languages have not been tested on natural text. Numerous NLP (Natural Language Processing) researchers have worked on the problem of syntactic parsing for Tagalog. Giganto [15] developed PinoyMMT which has a syntactic parser component for both Tagalog and Cebuano. However, the system was only tested on 16 base sentences. Another syntax parsing system for Tagalog was developed by Borra et al. [16] for use in a machine translation system between English and Tagalog. This work used the LFG framework for 12 Tagalog sentences but it was concluded that the system was only as good as the linguistic resources available to it. The need for “the formal grammar for languages involved” was cited. The LFG framework was also used by Manguilimotan for parsing a dataset that “covered only simple Tausug verbal sentences” [17].

7. Conclusion and Recommendations

This research, using the Ayta Abellen language of the Philippines, has enabled the developer of PAWS to make needed adjustments to the tool for use with Philippine
languages. The next version of PAWS will incorporate these customizations.

But research is still needed. The custom rules should be expanded to cover all the structures of the Ayta Abellen language so that higher percentages of natural texts can be accurately parsed. Currently custom rules are being developed to account for syntactic structures in the test set that were not handled by the phrase structure rules.

Second, it is hypothesized that most of the PAWS answers for Ayta Abellen would be the same for Tagalog and many other Philippine languages. This hypothesis should be tested through the use of the PAWS Starter Kit with another Philippine language such as Tagalog or Cebuano. Comparing the resulting phrase structure rules with those for Ayta Abellen would show the level of similarity between these languages.

Third, it would also be helpful to have a customized version of PAWS for Philippine languages. Since the underlying template files that drive PAWS are in XML, it would be possible to modify them and make a version specifically designed for Philippine or Austronesian languages where default values are set to match proto-Austronesian syntax.

8. ACKNOWLEDGMENTS

Special thanks to Dr. Cheryl Black and Dr. Andrew Black for their encouragement and technical support throughout my research.

9. REFERENCES


