Bootstrapping an LFG F-structure Bank for Tagalog by Cross-Lingual Projection

Borra, Allan; Cada, Djesus Rex; Chen, HuaZong; Chan, Francis; Tan, Andrew Edwell
College of Computer Studies
De La Salle University
2401 Taft Ave.
1004 Manila, Philippines
(+632) 524-0402
borgz.borra@delasalle.ph;rex.cada@yahoo.com;
chenhuazong@yahoo.com;realfrancischan@gmail.com;edwelltan@gmail.com

ABSTRACT

This paper discusses a system that uses bootstrapping to induce Tagalog F-structures from the English F-structures. Bootstrapping was explored as an attempt to quickly produce a large corpus of F-structures for future research purposes. The core theory behind idea of bootstrapping is based on the Direct Correspondence Assumption that assumes that if two sentences are literal translations of each other and the words were aligned, then their syntactic relationship will be the same. A system that uses word alignment, parsing, and mapping was built to induce Tagalog F-structures from English F-structures. The system has achieved an F-score of 58%, and results seems to point that if two parallel sentences are translated as literally close as possible to each other, then the features of the induced F-structure are also correct.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language models, Language parsing and understanding.*

General Terms

Languages

Keywords

F-structures, Bootstrapping, Word Alignment, Cross-lingual Projection, LFG, NLP

1. INTRODUCTION

Machine translation, information retrieval, expert systems with natural language interface, and computer assisted language learning are different computer systems with very different functions and purposes. Though having those differences, these systems also have one thing in common – they deal with natural language. For any computer application which uses natural language processing, their performance and accuracy is directly related to the quality of linguistic resource available for them to be used. To accommodate the need for high quality linguistic resources, annotated corpus and grammars have been developed.

Generative grammars originated from the work of Noam Chomsky. The purpose of these grammars is to "develop formal mathematically explicit models of various aspects in the human language" [1]. Currently, the two biggest grammars are the Lexical Functional Grammar (LFG) and the Head-Driven

Phrase Structure Grammar (HPSG). The ParGram project used LFG to create grammars for English, French, Chinese, Arabic, German, Japanese, and Norwegian [2]. Delph-in, a community developing linguistic sources based on HPSG, has developed grammars for English, French, Japanese, Greek, Norwegian, and Spanish [3]. These grammars are used as the backbone for NLP applications. Information extraction tools, expert systems which perform querying based on the meaning of the text, and machine translation are just some of the possible applicable uses of the grammars.

An annotated corpus or more commonly known as a treebank is text annotated with parts-of-speech (POS) tags, syntactic structures and at times semantic functions. One of the largest and most used treebank is the Penn Treebank [4]. It has 4.5 million words of for American English and contains texts from Wall Street Journal, the Brown Corpus, Switchboard, and ATIS. The Penn Treebank has been used for language modelling, word sense disambiguation, POS tagging, statistical parsing, maximum entropy techniques and a lot more [5]. Other treebanks are the Prague Dependency Treebank [6] and the TIGER Treebank [7]. Parallel treebanks are basically treebanks with sentence pairs. From these sentence pairs the words are aligned and the constituents between sentence pairs are linked [8]. They are mainly used for machine translation and cross-lingual studies. Europarl is a well-known parallel corpora consisting of 11 European languages that contain documents from the European Parliament. The largest parallel corpus is the JRC-Acquis Multilingual Parallel Corpus which consists of legislative texts. Another possible resource for a parallel corpus is the Bible.

Treebanks and grammars are valuable resources for computational linguistics. However, the problem with these linguistic resources is that these are only available to a limited set of languages and creating them is time-consuming. The ParGram project [9], which began in 1994, is a consortium of researchers who develop hand-crafted LFG grammars. It took ParGram 15 years to be able to develop an industrial strength English grammar. The Penn Treebank was also hand-crafted, and according to [10], the rate at which an experienced annotator is able to work with is 700-1,000 words an hour which is roughly equivalent to 30-35 sentences an hour.

The two most common ways of creating grammars are either hand-crafting them or extracting them from a Treebank [11]. Clearly, hand-crafting a grammar is effort and time intensive. On the other hand, it is impossible to extract grammars from treebanks if a Treebank for the desired language does not exist. To solve this problem, a method called "bootstrapping" was used by [12] and [13]. According to the Oxford Dictionary, by definition bootstrapping is to get (oneself or something) into or out of a situation using existing resources. In bootstrapping, a pair of sentences, which are literal translations of each other, are aligned to project a syntactic structure. This is usually done by using a rich linguistic resource, like English, which is then translated to produce the projections for the desired language.

Tagalog is one of the languages that lack both a generative grammar and a treebank. As a result, the abovementioned lack of linguistic resources hinders growth and development of Tagalog NLP applications.

2. Tools and Related Techniques

2.1 A Rule-Based Tagalog Morphological Analyzer and Generator (MAG-Tagalog)

“The MAG-Tagalog System is a rule-based Tagalog Morphological Analyzer and Generator” (Aquino et al.). It has two modules, the Analyzer and Generation. The former permits the users to attain the root word of a given transformed word. The latter, on the other hand, generates a list of transformed words from a given root word. Morphological changes or morphological phenomena that transpired with a given word such as affixation and reduplication were distinguished by both modules.

The system supports nouns, verbs, and adjectives. 13,397 Tagalog words were used for testing the Generation module, while 16,540 Tagalog words were used for testing the Analyzer module. The system correctly generated 68.42% Tagalog words and 83.34% Tagalog words for analysis.

The system had issues with the effective handling of morphophonemic changes. For the Analyzer module, the system had difficulty in distinguishing root words and affixes as well. For the Generation module, a morphophonemic change called assimilation was the issue. The researchers recommend implementing Phonology, which includes stress and intonation, to address these issues because it is a vital component in applying morphophonemic changes on Tagalog words.

2.2 XLE

Xerox Linguistic Environment (XLE) is a "computational environment that assists in writing and debugging Lexical Functional Grammars" (LFGs) [2]. It consists of "cutting-edge algorithms for parsing and generating Lexical Functional Grammar (LFGs) along with a rich graphical user interface for writing and debugging such grammars" [14]. It is also for the Parallel Grammar Project, which is developing grammars for English, French, etc. [14]. Basically, linguists are provided with a tool for writing, testing, and editing syntactic rules and lexical entries. XLE also has finite-state morphological analysers in its interface.

XLE outputs the C-structures (whether or not they have valid F-structures),the chart containing all complete or incomplete bracketings of the input string that the grammar allows, the morphology which is all possible morphological analyses of each lexical item, and F-structures (including display of inconsistencies, incompletenesses, and incoherencies).

After installing a collection of syntactic rules and lexical entries into XLE, you can see whether those items are sufficient to analyse sentences or phrases in the language in question. You can also easily mix and match different sets of linguistic specifications as you experiment with different versions of particular rules and lexical entries, whether you have written them yourself or they have been provided by other users of the system [2].

XLE will be used by the proponents principally for annotating purposes. The English version of a bitext corpus will be inputted sentence by sentence. The outputted F-structures will be used soon after the process of word alignment.

2.3 Cross – Lingual Projection of LFG F-structures

The automatic induction of LFG grammars is the induction from the existing treebanks, but for the resource – poor language, manually construction of LFG grammars is expensive. The idea in the cross – lingual projection is that using a bilingual corpus, like English-Tagalog corpus, analysis tools are applied to the resource – rich language side, in this case, it's English. From the automatically produced word alignment links, the resulting annotations are projected to the resource – poor language which is Tagalog. The projection of syntactic dependencies is based on the Direct Correspondence Assumption, which states that the dependencies in a source sentence directly map to the syntactic relationships in the word-aligned target translation [15]. According to [12], the projected annotations are noisy, postcorrection rules and filtering methods may apply. Throughmapping of the induced F–structure to its appropriate C–structures, and using of the C- and F-structure bank, following the method of [16], a full – fledged LFG grammar can be obtained.

Two main characteristics of LFG make it especially suitable for this cross-lingual projection method:

1. Since LFG is a lexicalized theory, projection of annotations assigned to particular words can be sufficiently guided by word alignment.
2. F-structures constitute an abstract level of analysis that is largely invariant across languages,and thus perfectly suited for projection between languages with varying word order [15].

3. SYSTEM DESIGN

The corpora consisted of parallel texts from the Bible and Antoine de Saint-Exupéry's*The Little Prince*. The versions of the Bible used for this research are the *American Standard Version of the Bible [1901]* and Ang Dating Biblia [1905] for English and Filipino respectively. The Bible corpus consists of 28,791 sentences and 815,576 words for English and 29,512 sentences and 882,014 words for Tagalog. *The Little Prince* consists of 1,521 sentences and 16,802 words for English and 1,353 sentences and 15,145 words for Tagalog. The main reason for expanding the corpus is mainly to address accuracy issues which might happen during word alignment. The software used for the sentence and word count is an online tool. It can be located at *http://textmechanic.com/Count-Text.html*. These corpora will be stored in a plain text (.txt) file. Each line of the text file will consist of only one sentence.

Sentence and word alignment were done separately. Based on the sentence count, the sentence alignment is not 1:1. This inconsistency can be attributed to various reasons like miscount done by the software used or translation method used. Sentence alignment was done manually to produce the best alignment quality possible. Sentences with no possible alignments were discarded. The sentence-aligned corpus was used as the input for the XLE parser.

Parsing the English side of the corpus was done using XLE and the English grammar developed by PARC. XLE uses a hand-crafted English grammar from LFG. XLE was used to generate the English F-structures onto which Tagalog words were mapped to. The grammar used for parsing was English-2009-11-25. XLE has a function which allows for selecting the most probable analysis. This function was used to choose the most probable F-structure for a given sentence. The English corpus was parsed one sentence at a time. XLE outputs F-structures as Prolog (.pl) files. The types of sentence parsed were mainly dependent on the types of sentences present in the corpus. The grammatical function tags which used were also the tags used by XLE.

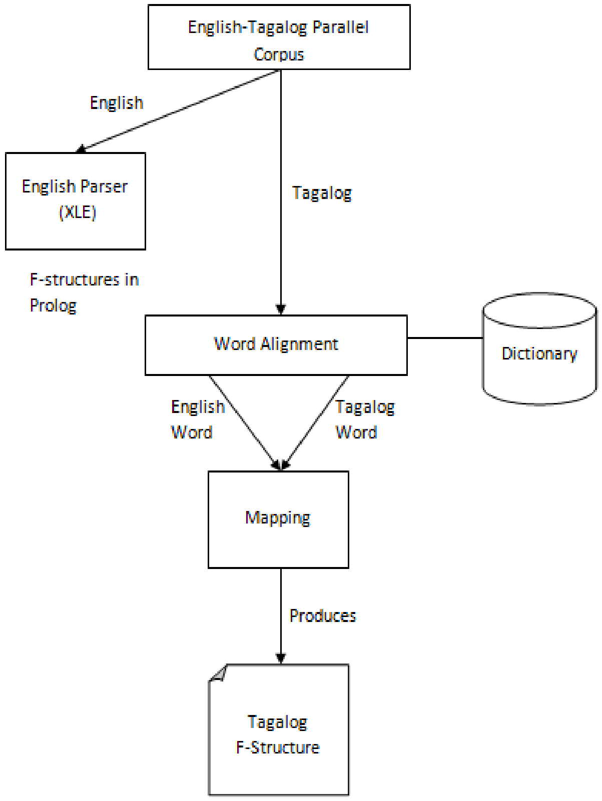


Figure 3: The System Architecture

The word alignment was done by extracting each English word from the F-structure's PRED. Each word was then translated by looking for a match in the dictionary, and when a match is found the corresponding word was used to replace the original English word. To determine the accuracy of the word alignment, manual inspection was done with the aid of existing tools such as bilingual dictionaries and translators. The inspection of the accuracy of the word alignment was manually done by native speakers. During disagreements to determine whether a word alignment is correct, the issue was discussed between all inspectors until a unanimous decision was achieved.

Syntactic structure mapping was used to produce the F-structures. Using the Prolog output files of XLE the Filipino words were mapped to the syntactic structure based on word alignment. The researchers were the ones who designed the algorithm for mapping. The output file was still a Prolog file except that the previously English words are now Filipino words. All features and annotations from original English F-structure were still maintained after mapping.

4. DESIGN AND IMPLEMENTATION ISSUES

The system is divided into two phases: parsing and mapping. Parsing is handled by XLE's parser. Parsing is a simple three-step process. Basically it just inputs the English sentences, let XLE handle the parsing, then produce the English F-structure outputs.

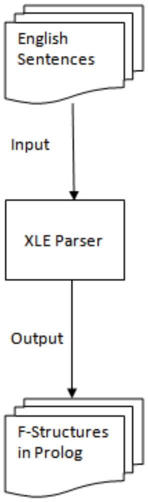


Figure 4: Diagram of the Parsing Process

The word alignment process begins immediately after the parsing process is completed. It requires two inputs, the Prolog representation of the F-structures produced during the parsing phase as well as a text file containing the parallel translation of the English sentences used as the input in the parsing. Each Prolog file is then processed to produce a Tagalog F-structure by mapping Tagalog words into the corresponding English words.

4.1 Bootstrapping Process

This section discusses the bootstrapping process done in the research.

4.1.1 The Dictionary

The system uses a parallel dictionary stored in a database. The dictionary contains an English word, an equivalent Filipino translation, and the words' Part-of-Speech. The initial entries of the dictionary were gathered from an English-Tagalog dictionary. The dictionary already contains the English words, Part-of-Speech, equivalent Tagalog words, usage examples, and the special cases where two English words are used to produce an equivalent meaning. Since, the system uses direct and literal translations to map the Filipino words to the English words in the F-structures, some information were deemed unnecessary, particularly the usage examples and the special words. The unnecessary information were removed and the dictionary was altered into a format which is easier to parse, and the entries were stored into a database.

A text parser was built for the sole purpose of reading the dictionary to be stored into the database. Building the tool was not a problem, but the real problem was cleaning the dictionary to a format that is readable by the parser. It took roughly around 30 hours for the researchers to clean a dictionary with 21,829 entries. The developed parser can be used to store new entries into the dictionary as long as it follows a prescribed format.

Each entry of the dictionary was then passed to Mag-Tagalog's morphological generator. This is done to increase the chances of finding a Tagalog word to match an English word. As of this paper, the dictionary contains 206,886 entries.

4.1.2 The Corpus

The corpus used for developing the F-structures was taken from a bilingual Tagalog-English parallel text of The Little Prince. These texts were prepared by separating the text such that it will be arranged having one sentence per line, this format will carry on for both texts. Also, it is to be made sure that a sentence in Tagalog in a certain line will be the equivalent to the English sentence on the same line. Aside from that, it is expected that both text files will contain the same amount of sentences.

Lines from the given corpus are discarded given that they are either incomplete sentences (fragments), or parallel sentence pairs that are too far apart in meaning. Other corpus can be used in place of The Little Prince, but should follow the format of having one sentence per line and the equivalent sentence on the same line of the other text file. In selecting sentences, it is advised to avoid dialogues like ““How are you?” asked Jane.” because given that the line is a sentence, XLE has a tendency to parse the line with the dialogue along with the line below it and counts it as one sentence.

4.1.3 Parsing

The system uses a wrapper to call XLE's parsing function. XLE uses the LFG grammar for English that was produced by Xerox PARC. The system heavily relies on XLE's output. Thus if XLE's parse is wrong and produces an inaccurate F-structure, the resulting error will also be carried over to the Tagalog F-structure.

4.1.4 Word Alignment

The researchers initially proposed and used GIZA++ as the third-party tool that will do the word alignment for the system. However, the aforesaid tool had issues, which will be discussed later in this chapter. This resulted to the development of the researchers' own implementation of a simple word alignment algorithm.

The word alignment starts by extracting the English words from the *Constraints* section of XLE's Prolog file. This is done by searching for a “semform(‘word’)“ value in the *Constraints* section. Once a “semform(“ marker is found, the English word it contains is extracted and the system tries to find a match in the dictionary.

In cases where no matches are found in the dictionary, the English word is retained and will serve as the equivalent Tagalog word. If there is only one entry that matches the current English word, the solitary Tagalog word is retrieved from the result set. In most cases, two or more matches are found in the dictionary. If this would be the case, the Tagalog words from the result set will be compared to every word of the Tagalog input sentence. Once any of the words from the Tagalog sentence matches the Tagalog words from the result set, that word will be retrieved and considered as the equivalent Tagalog word. On the other hand, if a match is not found, the first word from the result set will be retrieved and taken as the equivalent Tagalog word.

4.1.5 Word Alignment with POS

The addition of a parts of speech extractor is intended to further refine the results from the dictionary. The aim here is to get certain word's specific part-of-speech. As opposed to searching the dictionary for just the word which may return several results having multiple parts-of-speech, the English word and a matching part-of-speech which is taken from the English sentence, and will be passed on to the dictionary for a result that is more accurate.

This solution is developed because despite the previous algorithm working properly, it is found to be weak in cases where there are multiple results and ends up selecting the first result -that comes up. With incorporation of a part-of-speech extractor, it resolves that case where instead of selecting the first result that is generated, it looks for a result with a matching part-of-speech.

Parts-of-speech that the system covers are verbs, nouns, determiners, adverbs, adjectives, pronouns, prepositions, and conjunctions. Below is a flow of how the word alignment with the POS extractor functions:

1. The algorithm scans the C-structure portion of the Prolog file for the word

2.On the same line, the word will have a corresponding ID surrounded by '[]'; that will be extracted

3.The part-of-speech of the matching the current ID will be extracted

4. The dictionary will be queried to check if it contains a word with the extracted parts-of-speech.

5. After the database is queried, the algorithm runs similarly with the algorithm mentioned in section 5.2.4

6.There are cases where a word would have no part-of-speech in the C-structure. For cases like this, the dictionary is queried for the English word without supplying a part-of-speech to narrow down the search

4.2 Issues Encountered and Limitations of the System

Throughout the development process the researchers have encountered issues that have acted as hindrances to the whole research progress. The limitations would be the dictionary itself, while the issues include a lost cause for developing the system around Giza++ along with a hundred year old corpus using antiquated English and Tagalog.

4.2.1 Issues Encountered with Bible Corpus

In the middle of development, the group discontinued the production of Tagalog LFG using the Bible corpus. This is because the Bible corpus was too old. Aside from that there is a significant population of the sentences used were incomplete, and contains a huge amount of dialogues which the researchers have mentioned to be not usable.

In the middle of development, the group discontinued the production of Tagalog LFG using the Bible corpus. This is because the Bible corpus was too old. Aside from that, there is a significant population of the sentences used were incomplete, and contains a huge amount of dialogues which we have mentioned to be not usable.

The issue regarding the corpus being too old was in terms of the words used in its sentences for both Tagalog and English, and was very evident at first. But later on was figured to be ineffective to build an F-structure bank using sentences/phrases/words that most of the people don't use anymore.

In addition to the issue of the old corpus, many of the sentences in the Bible corpus were incomplete or dialogues. As dialogues and incomplete sentences would be removed, this would significantly decrease the size of the F-structure bank that will be produced.

Another reason for discontinuing the usage of the Bible corpus was due to the fact that there seems to be a lot of sentences that XLE was unable to parse properly. These parsing failures are basically brought about by the text being old, and being incomplete (but more on the corpus being too old).

4.2.2 Issues with Giza++

The researchers initially worked with a word aligner called Giza++. Giza++ needed parallel texts to be aligned in the sentence level. Giza++'s input needed to be formatted to contain only one sentence per line. The first corpus which was used for Giza++ was the Bible corpus. The Bible corpus contained 28,791 for English and 29,512 sentences for Tagalog. However, the original format of the Bible corpus was one verse per line. A tool was developed by the researchers to separate the sentences into lines. Although the process was partially automated, all the books were not thoroughly cleaned. As stated earlier, the

Tagalog Bible contained more sentences than the English Bible. The researchers had to manually check each book of the Bible and manually align the sentences. This whole process of took almost three weeks to complete.

For the first testing of the system using Giza++ and the Bible corpus, the results were poor. The researchers thought that the corpus didn't have enough sentences and decided to add The Little Prince into corpus. The Little Prince also had to be sentence aligned, but it took only 4 days to completely align 1,521 and 1,353 English and Tagalog sentences respectively.

A second test was conducted and the word alignment and mapping results were still poor. The researchers suspected that the reason for the poor results was because the Bible corpus was out of date and used old English and Tagalog. Further testing was conducted to see if using only The Little Prince yielded better results, but that was also a failure.

Since the researchers no longer had access to other parallel corpus, new parallel texts had to be created. The researchers translated ten Wikipedia articles which took roughly 3 weeks to complete. The translated text contained 1,294 sentences. These sentences were added on top of The Little Prince, but the results didn't improve. Upon further inspection, the F-structures would have been correct for some sentences if the word alignment was correct.

Steps were done to find out how word alignment accuracy could be increased. Research suggested that increasing the size of the corpus increased the accuracy of the word alignment. The researchers tried joining the entire corpus into one text file as input for Giza++, but the produced F-structures were still poor. Other word aligners were tested but they also produced poor results and didn't contain sentence-per-sentence word alignment that Giza++ produces. The researchers also tried using Giza++'s option to input a dictionary while training the inputs, but it made no difference. Giza++ produced the same alignment with or without the dictionary.

Since Giza++ wasn't showing any improvement and the researchers had access to a dictionary, a word alignment system using the dictionary was developed. This was initially done to test if it a dictionary based word alignment can produce better results, and initial testing showed that it did. This led to the researchers abandoning Giza++ and focusing on improving alignment using the dictionary.

Having incorrect word alignment results greatly affects the accuracy of the induced F-structures. The resulting F-structure is poor because of this incorrect word alignment. Unlike the dictionary based word alignment, errors in Giza++'s word alignment are not so easy to correct or improve because the system has no direct way of influencing how words are to be word aligned.

4.2.3 Limitations of the System

The dictionary, although it acts as a tool to further refine the results and quality of the Tagalog F-structures that will be produced, also serves as a huge limitation to the system. It limits the system or the quality of output according to how big the dictionary is or if it contains the equivalent Tagalog word that is being searched. Knowing that the system goes through the dictionary by looking for the Tagalog equivalent of the English word which is done word for word, it is an issue that the dictionary won't be able to handle two or more words that are connected or bound by a certain context such as the words "take off" the system would look it up one word at a time, so it would eventually end up with something like "kuhapatay" because it didn't count "take off" as a word or something that is connected to produce something more of "lumipad". So in a nutshell the system's ability to produce the Tagalog F-structures is as good as its dictionary.

A post-correction rule module was supposed to be implemented into the mapping system. However, because the produced F-structures by using the Giza++ word alignment module were poor, the researchers felt that the F-structures were not yet in a state where it could be evaluated by a linguist to find any useful information. During that stage, most of the rules which could be implemented were only rules to correct the misaligned words, not to correct any grammatical or language discrepancies. However, improving the word alignment to an acceptable level took longer than expected and the post-correction rule module was not implemented.

Currently, the F-structures produced are all based on the English F-structure. If a discrepancy is found, the only way to fix that is use text editors to edit the Prologfiles directly. Adding or removing features can only be done in the Prolog file which requires a good understanding of the syntax used by LFG.

5. RESULTS AND OBSERVATIONS

This section presents the comparison of Giza++ and the Dictionary based Word Aligner accuracy in word alignment. This section also presents discrepancies found as a result of directly mapping Tagalog words to English F-structures. It also compares the F-score of the current experiment with the F-score of other bootstrapping process made for another language.

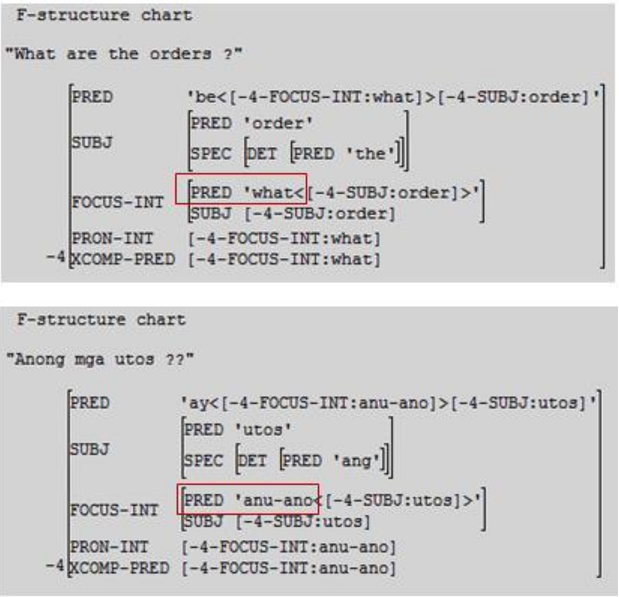
5.1 The Corpus Used for Evaluation

The corpus contains 25 sentences with a total of 160 words. The sentences on this corpus were randomly selected from both the Little Prince and some books of the Bible. These sentences were used in evaluating both the word alignment accuracy and linguistic evaluation.

5.2 Comparison of Giza++ and Dictionary-based Word Aligner

The initial plan of the system was to rely on Giza++ to do the word alignment for the mapping of the F-structures. However, when the results were analyzed, Giza++'s word alignment was unsatisfactory. There were cases in which it seemed that the features of the induced Tagalog F-structures were wrong, but it was actually only because the word alignment was incorrect. For this reason, the researchers developed their own word alignment algorithm, and found out that it performed better than Giza++. The actual outputs can be found in Appendix A.

Both tests were conducted on the same set of corpus. For each sentence, the number of correctly translated words, mistranslated words, and untranslated words are counted. Untranslated words are words which were kept in English because there were no entries found in the dictionary or the equivalent translation was a "null." Mistranslated words are words which were translated but are incorrect because the selected word's definition is out of context or simply because the selected word is incorrect. Translations are considered correct if the selected word exists in the original sentence, or the selected word's usage and definition is similar to the original word. Figure 5 is an example where the English word "what" was translated into "anu-ano." Although "anu-ano" is different from the original word, "Ano", of the sentence, its usage and definition is correct and is therefore counted as a correct translation.



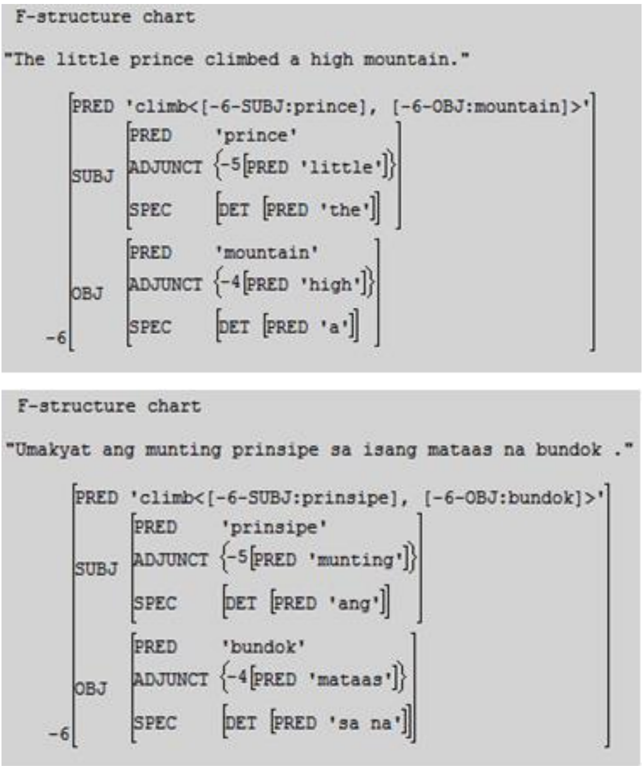
```
F-structure chart

"What are the orders ?"

     PRED         'be<[-4-FOCUS-INT:what]>[-4-SUBJ:order]'
                  PRED 'order'
     SUBJ
                  SPEC DET PRED 'the'
                  PRED 'what<[-4-SUBJ:order]>'
     FOCUS-INT    SUBJ [-4-SUBJ:order]
     PRON-INT     [-4-FOCUS-INT:what]
   -4 XCOMP-PRED  [-4-FOCUS-INT:what]
```

```
F-structure chart

"Anong mga utos ??"

     PRED         'ay<[-4-FOCUS-INT:anu-ano]>[-4-SUBJ:utos]'
                  PRED 'utos'
     SUBJ
                  SPEC DET PRED 'ang'
                  PRED 'anu-ano<[-4-SUBJ:utos]>'
     FOCUS-INT    SUBJ [-4-SUBJ:utos]
     PRON-INT     [-4-FOCUS-INT:anu-ano]
   -4 XCOMP-PRED  [-4-FOCUS-INT:anu-ano]
```

Figure 5: Evaluation Corpus Sentence 2 F-structure

The words to be translated are not the words of the sentence but only the words in the "PRED: Value" format. This is because XLE uses the lemmatized words as values in its PRED. XLE also omits some of the words like conjunctions and helping verbs. The omitted words are not translated.

The dictionary based algorithm was able to correctly translate 76.25% of the total words while Giza++ was only able to correctly translate 25.63% of the total words. This huge difference is what made the researchers drop the use of Giza++ as the primary tool for word aligning. However, Giza has also produced well aligned words. In **Figure 6**, besides the word "climb" which was left translated, and the word "a" which was mistranslated, Giza++'s result for this sentence can be considered acceptable. This may mean that once Giza++'s technology is improved, it may be a viable option for quickly word aligning without the need of a dictionary.



```
F-structure chart

"The little prince climbed a high mountain."

      PRED 'climb<[-6-SUBJ:prince], [-6-OBJ:mountain]>'
            PRED     'prince'
      SUBJ  ADJUNCT {-5[PRED 'little']}
            SPEC     [DET [PRED 'the']]
            PRED     'mountain'
      OBJ   ADJUNCT {-4[PRED 'high']}
   -6       SPEC     [DET [PRED 'a']]
```

```
F-structure chart

"Umakyat ang munting prinsipe sa isang mataas na bundok ."

      PRED 'climb<[-6-SUBJ:prinsipe], [-6-OBJ:bundok]>'
            PRED     'prinsipe'
      SUBJ  ADJUNCT {-5[PRED 'munting']}
            SPEC     [DET [PRED 'ang']]
            PRED     'bundok'
      OBJ   ADJUNCT {-4[PRED 'mataas']}
   -6       SPEC     [DET [PRED 'sa na']]
```

Figure 6: Mapped F-structures Using Giza++

The researchers tried to further improve the accuracy of word alignment by considering the parts of speech of the words to be aligned. However, it slightly improved the accuracy by 3.13% as far as the correctly aligned words are concerned. This is because the usage of the words (parts of speech) is now considered, thus letting the word alignment algorithm select from a narrowed set of words with the same parts of speech. For the errors, the total number of untranslated words was lessened, but the total number of mistranslated words increased.

5.3 Linguistic Analysis

According to the linguist, Dr. ArieneBorlongan, the produced F-structures are fairly accurate. That is the feature of the F-structure correctly performs their functions. Adjectives and adverbs are correctly labeled as ADJUNCTS and directly describe or modify the SUBJ or OBJ. However, since the translation or mapping is direct, there are some discrepancies between the two languages.The errors found during the evaluation of the F-structures are (1) untranslated words, (2) mistranslations, (3) extra adjuncts, (4) subject not captured, and (5) changed focus.

Mistranslation is to be expected due to how the word alignment algorithm was implemented. From the 25 sentences in the corpus, 13 (52%) of them had mistranslations. The word which was most commonly mistranslated was "a", and the other words were mistranslated simply because there were no correct entries in the dictionary.

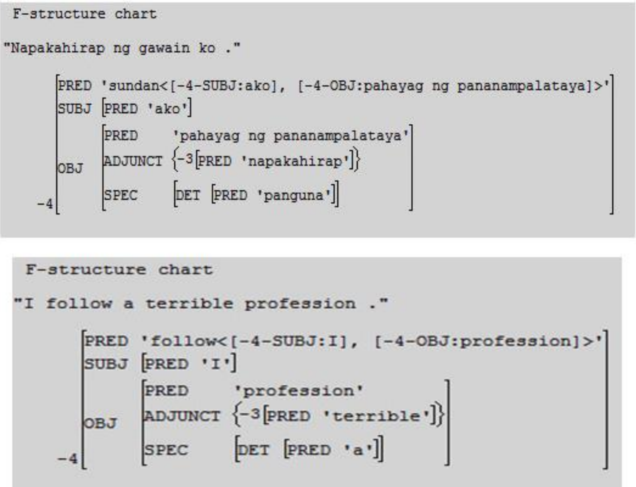
Sentences having untranslated words are also to be expected. 5 out of the 25 (20%) evaluation sentences contained untranslated words. The untranslated words are a result of the English words having no literal Tagalog translations.

Extra ADJUNCTS are similar to untranslated words, but for this case, instead of leaving the extra English words untranslated, they were properly translated but had no corresponding word in the original Tagalog sentence. S1 of the evaluation corpus showed this particular discrepancy. It can be seen that the sentence uses "Kakaibang-kakaiba" to describe "planeta", but the English sentence uses only "strange" to describe "planet". Comparing the "Kakaibang-kakaiba" and "strange", both of them do have the same meaning but their degree of comparison is different. In order for "strange" to be in the same degree of "Kakaibang-kakaiba", it needs the adverb, "very", to compliment it.

However, some induced F-structures it seem that the features are correct. Instead of using the translation of the original sentence, the direct translation was used. The translation for "kakaibang-kakaiba" has been changed to "tunay" & "kakaiba". The resulting sentence would then be "Tunaynakakakaibaangikalimangplaneta.", which is still similar to the original sentence. This phenomenon can also be observed when Tagalog words have the "Napaka-" affix or when a Tagalog word is repeated like "matamisnamatamis" (very sweet) or "makulitnamakulit" (very persistent).

Another issue that was encountered was the English F-structures labeling the subjects of imperative sentences as "null_pron". It is a known fact that the subject of imperative sentences is always "you", but it is not explicitly written in a sentence. Instead of placing "you" in its subject field, XLE inserts "null_pron" in the subject field and annotates it to indicate that "null_pron" is a second person pronoun. For this reason, the Tagalog F-structures also have "null_pron" as the subjects even though a form of "you" is explicitly stated in the Tagalog sentence. This error can be found in S13 and S19 of the evaluation corpus.

Figure 7 is an example of a Tagalog sentence whose F-structure was not properly represented by the English F-structure. The focus both sentences are different. In the English sentence, the subject of the sentence is "I". However, for the Tagalog sentence, the focus is not "ako", but rather on "gawain". This problem is due to the fact that the translations of the two parallel sentences are too different from each other.



```
F-structure chart

"Napakahirap ng gawain ko ."

     PRED 'sundan<[-4-SUBJ:ako], [-4-OBJ:pahayag ng pananampalataya]>'
     SUBJ [PRED 'ako']
          PRED      'pahayag ng pananampalataya'
     OBJ  ADJUNCT  {-3[PRED 'napakahirap']}
          SPEC      [DET [PRED 'panguna']]
  -4
```

```
F-structure chart

"I follow a terrible profession ."

     PRED 'follow<[-4-SUBJ:I], [-4-OBJ:profession]>'
     SUBJ [PRED 'I']
          PRED      'profession'
     OBJ  ADJUNCT  {-3[PRED 'terrible']}
          SPEC      [DET [PRED 'a']]
  -4
```

Figure 7: F-structure of Sentence: "I follow a terrible profession."

5.4 F-measure

The following list contains the categories that have been identified as the result of analyzing the F-structures:

- Dictionary - This is an error caused by the dictionary not containing the correct entry for a Tagalog word.
- Corpus – This is an error caused because the sentences used are too different in translation. This means that the Tagalog sentence was constructed in such a way that the context of the English sentence was kept, but some words were removed or added into the Tagalog sentence which caused the sentences translation to be too far apart.
- Corpus(Lexical Difference) – This is an error caused because an English word doesn't have a direct equivalent Tagalog word. Some Tagalog words can only be described by two English words. This requires the English F-structures to be restructured to fit the Tagalog Word.
- XLE – This is an error caused by XLE. XLE's parsing is not 100% perfect and sometimes produces errors in features

The F-measure was calculated by using the following formula (Listing 1):

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (1)$$

$$Precision = \frac{Geranated\ Feature\ \cap\ LFG\ Feature}{(All\ Preds\ -\ Preds\ of\ those\ 0\ correct\ features)} \quad (2)$$

$$Recall = \frac{Generated\ Feature\ \cap\ LFG\ Feature}{All\ Preds} \quad (3)$$

Listing 1: Formulae Used for Calculating F-scores

Experiments	Language	F-Score
Wroblewska et al.	English - Polish	49.98%
Hwa et al.	English - Spanish	33.9%
Current Experiment	English - Tagalog	58.59%

Table 1: Comparison of F-scores

The F-scores of Wroblewska et al. and Hwa et al. used for the comparison are from the F-scores of their Direct-Alignment without correction rules. It can be seen that the F-scores of the current experiment is higher than the other works, as shown in Table 1. This could be due to the fact that different measuring criteria were used to measure precision and recall. Neither the methods nor the formulas used for calculating the F-scores werementioned in both works. It could also indicate that English and Tagalog language has more similar grammatical characteristics that made it easier to produce more accurate word alignments.

6. CONCLUSION AND RECOMMENDATIONS

This section contains the things that were discovered while doing this research. This part also contains recommendations for improving the word alignment system as well as proposals for future works.

6.1 Conclusion

In summary, the study has produced a system for parsing English text to produce F-structures, perform automatic word alignment by using direct translations from a dictionary, and then inducing the English F-structures, to produce Tagalog F-structures. The researchers have found out that it is possible to produce Tagalog F-structures by inducing the English F-structures produced by XLE. For many simple sentences, the features of English and Tagalog are similar, and the tags of the English F-structures can simply be renamed to its Tagalog equivalent. The researchers have also determined that the quality of the word alignment plays a significant role for producing an accurate induced F-structure. However, upon the researchers' and the linguist's evaluation and analysis of the F-structures, there were also errors found such as untranslated words, mistranslations, extra adjuncts, not captured subject, and changed focus. These errors were mainly caused by the quality of the corpora, size of the dictionary, and the subject-focus phenomena of the languages. The use of part of speech tags resulted into slightly better results, as it reduced the number of untranslated words by 5.63%, increased the number of mistranslations by 2.5%, and increased the number of correctly aligned words by 3.13%.

Using corpora written in Old English is not recommended as XLE is not able to completely capture their F-structures. Also, sentences written in dialogue format are not recommended because XLE has doesn't capture them as sentences.

It has also been determined that a dictionary-based word alignment performs significantly better than a statistical word aligning tool like Giza++. In fact, the dictionary-based word alignment had an accuracy of 76.25%, compared to only 25.63% of Giza++. Because the correctness of the induced F-structures is greatly dependent on the quality of the word alignment, it is important that the word alignment process be fairly accurate. Not all of the induced F-structures are perfect because there are some discrepancies produced like mistranslations, untranslated words, extra adjuncts, and incorrect feature tags. Parallel sentences whose translations are as literally close as possible produce the most accurate F-structures.

The researchers have developed a tool that is able to automatically induce Tagalog F-structures from English. From this small scale experiment, the researchers have discovered that there are some Tagalog sentences which cannot take an English F-structure due to monolingual restrictions. The scale of the experiment should be increased to find more types of sentences which can and cannot be directly mapped. The system is not limited to only producing a Tagalog F-structure bank. It can also produce F-structure banks for other languages as long as a parallel corpus and dictionary, with English as the source language, are provided. This can open up research for producing LFG for other Filipino languages such as Cebuano, Ilongo, or Bikolano.

6.2 Recommendations

The most common errors found in the induced Tagalog F-structures are mistranslations and untranslated words. This is because either there was no entry in the dictionary, or because there is no direct translation equivalent for that particular English word (e.g. the English word "a" is usually not translated by the system). The first problem can be solved by simply increasing the number of entries the dictionary. In the current implementation of the system, if no match was found in the dictionary, the system simply keeps the original English word as the translation for the Tagalog word. The proponents would like to recommend the use of existing statistical machine translators to translate the unknown words. XLE also provides annotation for the English words, it would be good to look at those annotations to see if there are ways to use those annotations to improve word alignment.

Another way in which word alignment can be improved is by using a lemmatizer instead of a morphological generator. Since XLE lemmatizes the English words, the translated Tagalog words should also be lemmatized. The dictionary should be run through a lemmatizer which would produce a smaller and cleaner dictionary and improve the time for finding matches. Since the system's word alignment algorithm uses the parallel Tagalog sentence as a basis for determining which translation to use, the Tagalog sentences should also be processed by the lemmatizer to increase the chances of finding a match.

After the word alignment issues are fixed or it is able to produce satisfactory results, the next area that needs to be looked into is creating correction rules for catching the differences in English and Tagalog. According to Hwa, et.al (2005), they were able to build a parser from using projected Chinese treebanks that is only a few points below a parser made from one to two years of manually annotated treebank. It takes significantly less time, requiring less than one person-month, to write manual correction rules to account for limitations in projecting dependencies from English. The researchers would like to propose collecting a small corpus containing the discrepancies of English and Tagalog structures. The corpus will contain only one type of discrepancy at a time, and develop rules to capture, then slowly expand the corpus until all discrepancies can be covered.

XLE is able to annotate the English words, and these annotations are also carried over in the induced Tagalog F-structures. Assuming that the word alignment is satisfactory and the discrepancies between English and Tagalog can already be captured, it would be useful to look into creating a treebank from by using extracting the annotations from the F-structures. Future works could use the same bootstrapping method used in this study to produce Tagalog F-structures or simply work with the F-structures produced in this study to build their own treebank.

7. References

[1] Falk, Y. (1999). *Philippine subjects in a monostratal framework.*Sixth Annual Conference of the Austronesian Formal Linguistics Association.University of Toronto.

[2] Natural Language Theory and Technology (NLTT). (2011b, July).*XLE project.*http://www2.parc.com/isl/groups/nltt/xle/

[3] DELPH-IN. (n.d.).*Deep linguistic processing with HPSG.* Retrieved from http://www.delph-in.net/

[4] Marcus, M., Marcinkiewicz, M. &Santorini, B. *Building a large annotated corpus of English: The Penn Treebank.* Retrieved from http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf

[5] Rocha, M., & Sánchez, J. (2009). *Machine Translation of the Penn Treebank to Spanish.*

[6] Hajic, J. (1998). *Building a syntactically annotated corpus: The prague dependency treebank.* In: *Issues of Valency and Meaning.*, (pp. 106-132). Karolinum, Praha.

[7] Brants, T., Dipper, S., Lezius, W., Plaehn, O., & Smith, G. (2001). *The TIGER treebank.*, (pp. 24-41).

[8] Tiedemann, J., &Kotz´e, G. (2009). *Building a large machine-aligned parallel treebank.* Proceedings of the *8th International Workshop on Treebanks and Linguistic Theories (TLT '08)*, (pp. 197-208). Milano, Italy.

[9] Butt, M., Dalrymple, M., Dipper, S., Dyvik, H., Kaplan, R., King, T. H., Marcotte, J., Masuichi, H., Ohkuma T., Rohrer C., Rosen V., &Zaenen A. (n.d.).*Pargram lfg02 abstract* [online]. Retrieved from http://cslipublications.stanford.edu/LFG/7/lfg02pargram-abs.html

[10] Volk, M., & Samuelsson, Y. (2004).*Bootstrapping parallel treebanks.*Proceedings of the *7th Conference of the Workshop on Linguistically Interpreted Corpora (LINC)*, 71-77.

[11] Clement, L. &Kinyon, A. (2003).*Generating LFGs with a metagrammar.*InProceedings of the *LFG03 conference*. Retrieved from http://cslipublications.stanford.edu/LFG/8/lfg03clementkinyon.pdf

[12] Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., &Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. Nat. Lang. Eng., 11(3):311-325.

[13] Wroblewska, A., & Frank, A. (2009). Cross-lingual projection of LFG F-structures: Building an F-structure bank for Polish. In Eighth International Workshop on Treebanks and Linguistic Theories (p. 209). Retrieved from http://www.cl.uni-heidelberg.de/~frank/papers/tlt8_wroblewska_frank.pdf

[14] Natural Language Theory and Technology (NLTT). (2011a, July).*About nltt.*http://www2.parc.com/isl/groups/nltt/

[15] Frank, A., Sadler, L., van Genabith, J., & Way, A. (n.d.).*From treebank resources to LFG F-structures.*Automatic F-structure Annotation of Treebank Trees and CFGs extracted from Treebanks. Retrieved from http://www.cl.uni-heidelberg.de/~frank/papers/kluwer-tb-21.pdf

[16] Burke, M., Cahill, A., McCarthy, M., O'Donovan, R., van Genabith, J., & Way, A. (2004) *Evaluating automatic LFG F-structure annotation for the Penn-II treebank. Research on Language , Computation*, 2 (4). pp. 523-547. ISSN 1570-7075