Constituent Structure for Filipino: Induction through Probabilistic Approaches*

Danniel Alcantara and Allan Borra

College of Computer Studies, De La Salle University, 2401 Taft Avenue,
1004 Manila, Philippines
{alcantarad, borraa}@ dlsu.edu.ph

Abstract. The current state of Philippine linguistic resources, which includes formal grammars, electronic dictionaries and corpora are not yet significant to address industrial-strength language technologies. This paper discusses a computational approach in automatically estimating constituent structures from a corpus using unsupervised probabilistic approaches. Two models are presented and results show an F1 measure of greater than 69%. Issues and phenomena of the Filipino language are identified and discussed

Keywords: Computational Linguistics, Probabilistic Approach, Constituent Structure, Context Free Grammar, Grammar Induction

1. Introduction

This paper discusses the algorithms used for the automatic induction of grammar for the Filipino language.

The rationale for the study stems on the minimal work done on the development of a computational grammar for the Filipino language for the development of robust and industrial-strength natural language analysis and technologies. Existing Filipino grammars can only handle a subset of declarative type sentences. Considering the difficulty to manually construct a robust grammar capable of parsing a broad scope of sentences, automatic grammar induction is a consideration that can be used for learning language structure.

Automated grammar induction systems deals with the generation of a *grammar* based from input corpora. Existing work for grammar induction fall under two categories based on their input constraints: Supervised and Unsupervised. *Slightly supervised* systems generate grammar rules from bracketed corpora or tree banks. *Bracketed corpora* are text documents that have been bracketed by a linguist to represent the skeletal syntactic structure of the sentences. *Treebanks* are large corpora that have been annotated with the part of speech tags, syntactic structure, and other functional attributes necessary. *Unsupervised* systems make use of non-bracketed corpora while applying searching and clustering algorithms to attempt to learn the language rules.

Works on slightly supervised grammar induction, such as the works of Lari and Young(1991), Brill(1993), Sekine and Grishman(1995), and Charniak(1996), present different output formalisms to represent the grammar. However, Filipino is currently a resource-limited language and does not have the computational resources necessary for the algorithms presented. There are existing corpora available for the language, but these have not yet been bracketed.

* The authors wish to thank Dr. Shirley Dita for providing the initial gold standard.

22nd Pacific Asia Conference on Language, Information and Computation, pages 113–122

Osborne and Briscoe(1997), Clark(2001), Klein and Manning,(2001) attempt to learn the grammar formalism for the language through the use of statistical analysis methods applied to a tagged corpus. Klein and Manning(2002) applies a context-constituency model and achieved promising results, precision rate of 55% and recall rate of 48%. Existing systems are designed for the English language and have had modifications applied that are specific to the right-biased nature and Subject-Verb-Object structure. The Filipino language does not follow these structures; rather it has free word order patterns, and Predicate-Topic phenomena, wherein the focus of a sentence is referred to as Topic rather than as a Subject.

Klein(2005) experimented with a combination of the said context-constituency model and a dependency model Klein and Manning(2004), which was then applied to English, German, and Chinese. The English language reached recall of 88% and precision of 69%. The German language had a corresponding recall value of 90%, but a considerably lower precision of 50%, caused by relatively flat gold standard corpora. The flat structure of the German language is attributed to free word order, a phenomena also identified in Filipino.

The aim is to develop an automated grammar induction system using an unsupervised approach. The approach chosen is influenced by the limitation of computational grammar resources in Filipino. Existing induction algorithms, usually for the English language, are modified to handle Filipino language phenomena.

2. Unsupervised Grammar Induction Approaches

Unsupervised grammar induction systems apply statistical or probabilistic analysis to estimate the necessary grammar formalism. Most unsupervised approaches apply language specific heuristics to improve computational reliability.

2.1.Current Solutions in Other Languages

Clark(2001) developed an unsupervised approach for extracting the phrase-structure from an input English corpus. Contiguous tag subsequences are identified from an input tagged corpus, and are considered as constituent candidates. Tag sequences that occur at least 5000 times are clustered together based on their context, the part of speech tag immediately preceding and following.

The clustering process identified sequences with clear syntactic correspondence. Accordingly, the procedure identified clusters with poor syntactic quality. Mutual Information criterion is applied to justify the valid clusters and filter out spurious candidates. Mutual Information indicates the importance of a relationship between the prior and subsequent part of speech tags. This intuitively implies that a correct constituent structure is highly sensitive on the context of usage and can appear in different contexts.

The presented algorithm introduces the possibility to induce grammatical structure from an unsupervised approach. The initial results are filtered through Mutual Information heuristics to remove spurious candidates. The implementation of the algorithm minimizes the requirement on input corpora, but is computationally expensive and requires sufficient memory to handle clustering.

Klein and Manning(2001) describes two systems for learning linguistic constituency in natural language grammars. Radford's study, as cited in Klein and Manning(2001) discussed two linguistic criteria for constituency identified that serves as the primary basis for the work: 1. External distribution: A constituent is a sequence of words which appears in various structural positions with larger constituents; and 2. Substitutability: A constituent is a sequence of words with (simple) variants which can be substituted for that sequence.

Klein and Manning take a tagged corpus and apply statistical analysis to identify most commonly occurring contiguous tag sequences. The tag sequences that reach the target criteria, based on a combination of entropy and divergence, are identified as constituent structures.

Another work by Klein and Manning(2002) develops a "Constituent Context Model" describing contiguous subsequences within a sentence. Each sentence is described by a list of

all possible subsequences, identified by their span, enclosed terminals, and context, terminals before and after.

Bracketing of the sentence is identified through probabilistic analysis of a class conditional independence model based on the input corpora. Subsequences are estimated as constituent structures or non-constituents, referred to as distituents.

Initial bracketing is approximated based on a random split mechanism that promotes generation of unbalanced binary trees. The conditional completion likelihood of the current state is computed based on the specified Expectation parameters. Bracketing is adjusted and Maximized to further improve the generated constituents.

This model focuses on identifying hidden bracketing structures, based on observable sentence and tag structures. A precision rate of 55% and recall rate of 48% was achieved through experimentation of the identified model, the best published unsupervised results at the time. The implementation of the EM algorithm maximizes the combinational likelihood of generating a correct bracket span.

Although currently beyond the scope of the study of grammar induction for Filipino, it is noteworthy to discuss the work by Klein and Manning(2004) which made use of dependency structure and extraction to model the language. Unlike standard approaches, where the tree or phrase structure is being identified, the work identifies dependencies between words. Each word is directly dependent on another word, and a single word dependent on the ROOT component of the sentence.

2.2.Probabilistic Induction for Filipino: Architectural Design

Figure 1 provides the overall architectural design for the system. The first module is the training module. In the process of training, the training corpora are tokenized into sentences. Quotations found within sentences, words enclosed within double quotes, are separated from their host sentence and are tokenized further. Tokenized sentences are then tagged through the use of an external part of speech tagger. The tagged sentences are passed through statistical analysis, in order to retrieve the necessary data used for probabilistic computations in rule generation.

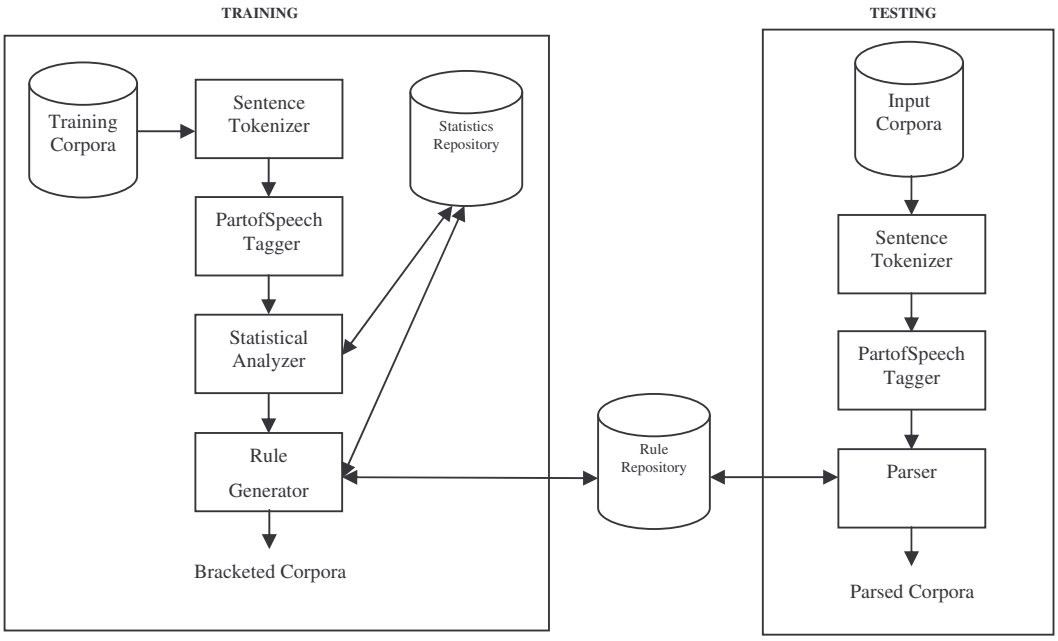

Figure 1: Overall Architectural Design of the Probabilistic Constituent Structure Induction System for Filipino

The statistical analyzer identifies all part of speech tag sequences that occurred in the training corpus. The symbol α will denote a specific part of speech tag sequence. The occurrence count of each α is scored and this value is one of the primary measurements used in rule generation.

All occurrences of α is identified within a context of two adjacent tags or sentence boundaries: $x\ \alpha\ y$. x denotes the symbol directly before α, and y directly after. These two contexts may be merged together to form a local linear context, x-y. The distribution of contexts in relation to α will be denoted by $\sigma(\alpha, x)$, $\sigma(\alpha, y)$, and $\sigma(\alpha, x\text{-}y)$ for pre, post, and linear respectively.

According to Klein and Manning(2001), a study by Radford states that "a constituent is a sequence of words which appears in various structural positions within larger constituents". The phrase "various structural positions" suggests that the constituency of α can be identified by the entropy of its linear context, $H(\sigma(\alpha, x\text{-}y))$.

However, entropy alone is not enough to recognize a constituent. This is due to uncommon but possible linear contexts having little bearing on the entropy. A scaled entropy, (listing 1) takes into consideration the uniform distribution of contexts in relation to α, $\sigma u(\alpha, x\text{-}y)$, and the uniform distribution of all possible contexts u.

$$Hs(\sigma(\alpha, x\text{-}y)) = H(\sigma(\alpha, x\text{-}y))\ [H(\sigma u(\alpha, x\text{-}y)) \,/\, H(u)\ ] \quad (1)$$

Clark(2001) discussed that if one were to consider the tag previous of a true constituent, one should be able to presume the tag directly after. However for non-constituents, possibly referred to as distituents, the two contexts form independent distributions. Therefore, one measure of constituency is identified by the Mutual Information between the pre and post contexts, $MI(\sigma(\alpha, x), \sigma(\alpha, y))$.

Towards the end of the training phase, the resulting probabilities are applied to the input sentences and an estimated parse is proposed. Two models are implemented for bracket estimation and rule generation. The first is a Greedy Selection approach that brackets the highest ranked sequence per iteration based on the measurement identified.

Algorithm Skeleton for the Greedy Selection Model.

```
Rules GreedySelection (Corpus corpus)
  apply statistical analysis of corpus
  while (at least one sequence meets filter criteria)
    maxSeq = filtered sequence with maximum measurement
    generate production rule based on maxSeq
    for (all sentences in Corpus)
      parse sentence with generated production rule
    apply statistical analysis based on new parse
  return Rules
end.
```

The selection and filtering criterion is based on available measurements of the sequence. The simplest filtering criterion is requiring the sequence size to be at least two. Different types of measurements can be applied for selection and filtering, such as requiring Mutual Information of 0.2 while obtaining the maxSeq based on occurrence.

The second model designed for training is a modification of the "Constituent Context Model" of Klein and Manning(2002). Sequences σ and linear contexts x-y are assigned probabilities of being a constituent or a distituent. Initial probabilities are derived from enumerating all possible valid bracketings of all sentences of the input corpora. Valid bracketing requires the bracket result in a binary tree. To minimize execution time, a three-dimensional dynamic matrix was implemented for said completions.

Overview of the Constituent Context Model Induction.

```
Rules CCMInduction (Corpus corpus)
  apply statistical analysis of corpus
  Collection inside = sequence probabilities
  Collection outside = context probabilities
  do
    for (all sentences in Corpus)
      bracketSentence =
        InsideOutside(sentence, inside, outside, 0, sentence.size-1)
    apply statistical analysis base on new parse
    inside = bracketed sequence probabilities
    outside = bracketed context probabilities
  while inside changes and outside changes
  combine inside and outside to generate Rules
  return Rules
end
```

The Inside Outside algorithm is utilized to dynamically estimate the optimum binary bracketing of each sentence. Sequence probabilities σ represent the inside computations and linear contexts *x-y* represent the outside. The process applies a divide and conquer approach to identifying the bracketing that will maximize the resulting probability. The estimated probabilities can also be weighted based on the previously identified measurements.

Skeleton of the Inside Outside Algorithm

```
 [bracketSentence,prob] InsideOutside
    (Sentence sentence, Collection inside, Collection outside,
      int nStart, //starting index token or tag
      int nEnd)   //end index token or tag
  String yield = sentence[nStart..nEnd]
  String context = sentence[nStart-1] + sentence[nEnd+1]
  double probC = inside[yield] * outside[context]
  double probD = (1-inside[yield]) * (1-outside[context]
  double prob = weight(yield) * probC / (probC + probD)
  if(nStart != nEnd)
    for( k = nStart ; k < nEnd ; k++)
      left = InsideOutside(sentence,inside,outside,nStart,k)
      right = InsideOutside(sentence,inside,outside,k+1,end)
      maxProb = max( left.prob * right.prob )
    prob *= maxProb
    bSentence = merge left.bracket with right.bracket based on max
  return [bSentence, prob]
end
```

After initial bracketing of all sentences, probabilities are recomputed and the process iterates in the form of an Expectation-Maximization algorithm. This will induce structure to the corpora and create a globally higher score as compared to the previous model.

The second module parses the input sentences utilizing the rules estimated during training. Prior to the actual parsing process, the input corpora is processed by the sentence tokenizer and part of speech tagger, equivalent to the learning phase.

The architecture makes use of three data sources, the Training Corpora, the Statistics Repository, and the Rule Repository. The Training Corpora is a non-bracketed data source that contains sentences for Grammar generation. It is possible that the training corpora be pre-tagged for correctness, but that is the only amount of preprocessing necessary for the unsupervised approach. The Statistics Repository will contain data relating to the training corpora, this includes the occurrence of tag sequences, the contexts of these sequences, and the corresponding occurrence. The Rule Repository will be the storage facility for grammar rules in the learning phase and the source of these rules for the testing phase.

3. Linguistic Data

The system presented takes as input a collection of sentences. The part of speech tags for each word are then identified and the algorithms only take into consideration the tags rather than their surface form. Sentences were manually verified, to minimize tagging error; however this is the only amount of preprocessing for both learning and parsing phases. A bracketed version of the corpus is available, and this is used for evaluation purposes only.

The part of speech tagset used contains 66 word classes, each falling into one of 9 supersets (noun, pronoun, determiner, conjunction, verb, adjective, adverb, cardinal, punctuation), and one added word class used to signify a quotation. The tagset is based on the updated tag set of TPOST. Punctuation marks are not removed from the input corpora, however statistical data is not gathered for sequences that contain punctuation marks. The punctuations are part of the final grammar rules, but they do not directly contribute to the estimation of constituents.

The corpus used for training was the Filipino translation of The Little Prince by Antoine de Saint – Exupéry. The selected corpus contains 1,685 sentences composed of approximately 15,000 words, and can be categorized under the domain of literature. The dataset contains a range of sentences, mostly simple and declarative sentences.

Due to the fictional narrative nature of the input corpora, a high percentage of the sentences serve a declarative purpose. There are blocks of dialogue found inside the story; this introduces the other types of sentence structures. "The Little Prince" is generally a children's book, the reason for the majority of simple sentences, but compound and complex sentences are also identified in the narration.

The experiments presented are done using sentence lengths of 1-10. Majority of the simple sentence structures can be found within this range. Compound and complex sentences are variations of the simple sentence construct and experimentation focused on simple sentences only is imperative.

4. Preliminary Results and Analysis

This section discusses the results and analysis from statistical data as well as evaluating proposed bracketings produced by the system against a linguistically verified gold standard.

4.1.Analysis of Statistical Data

Preliminary analysis is focused on the statistical analyzer, and identifies the appropriateness of the identified measurements. **Table 1** lists the top ten most frequently found constituents in the Gold corpus.

Table 1. Selected constituent sequences found in the gold corpus with corresponding statistics.

Sequence	Gold Probability	Gold Count	Occurrence	Raw Entropy	Scaled Entropy	Mutual Information
CCB NNC	87.16%	129	148	4.495	2.376	0.760
JJD NNC	94.16%	129	137	3.956	1.976	0.691
DTCP NNC	83.00%	83	100	4.670	2.381	0.839
CCB JJD NNC	100.00%	71	71	2.944	1.168	0.375
DTC NNC	52.46%	64	122	5.787	3.440	1.435
CCT PRSP	76.25%	61	80	4.908	2.534	1.132
NNC PRS	71.21%	47	66	4.875	2.452	1.696
PRSP NNC	86.79%	46	53	3.466	1.375	0.548
CCT NNC	67.74%	42	62	4.930	2.514	1.498
VBW CCB JJDNNC	93.18%	41	44	0.994	0.150	0.000
RBI PRS	3.57%	2	56	5.222	2.680	2.292
CCB JJD	2.63%	2	76	2.158	0.782	0.307
RBF PRS	1.43%	1	70	4.759	2.341	0.739

DTC NNC has a lower gold probability due to modifiers attached to the noun such as in the example [ang [kopya [ng drowing]]]. Ang kopya with no consideration of context can be considered as a constituent, however in this scenario the constituency should be primarily placed on kopya ng drowing, before the ang is included in the bracketing.

The occurrence count heuristic is the measurement that most closely resembles the gold count. However, as is the case with CCB JJD and CCB JJD NNC, sequences found inside constituents will have as high or possibly higher occurrence counts, but not be actual constituents.

The scaled entropy adjustment to the raw entropy measurement improves the criteria slightly. Both entropies appear to address the issue of occurrence concerning CCB JJD and CCB JJD NNC. However, both entropy measures rank noun phrases considerably lower, while considering conjunctive and prepositional phrases considerably high as is the case for RBI PRS and RBF PRS which ranked in the top ten of both entropies.

The mutual information measurement fares poorly in the listing, none of the high constituency sequences are correctly identified. This heuristic was originally proposed as a filter rather than a selection criterion.

Identifying a solid measurement to duplicate the actual rankings can potentially increase the quality of bracketing, especially when using the Greedy Selection Model.

4.2.Evaluation of Proposed Bracketing

Further discussion focuses on evaluating the output of the system. The system produces two output, the proposed bracketing and estimated rules. The estimated rules are the more representative output and may be compared to an existing formalism of the language. However when dealing with natural languages, especially resource limited languages, the grammar formalism is debatable and is not available for evaluation purposes. Furthermore, there are no predefined metrics useable for comparing two grammars, one can only state if the two grammars are the same or not.

The alternative output of grammar induction systems are the predicted structures the grammar rules produce. This is implemented through bracketing of input text and evaluated through comparison with a linguistically motivated gold-bracketed corpus. The gold standard represents the linguist's grammar, and equivalence to the gold-standard is equivalent to meeting the linguist's standard.

Precision, recall, and their harmonic mean, F1 measure are quantifiable statistical measures for identifying similarity, and in this case correctness, between the proposed and gold brackets.

Precision: the percentage of non-terminal bracketings in the predicted parse that also appeared in the tree bank parse

Recall: the percentage of non-empty non-terminal bracketings from the tree bank that also appeared in the predicted parse

F1 Measure: the combination of Precision and Recall

$$F1 = 2 * Precision * Recall / (Precision + Recall) \qquad (2)$$

Estimating entirely left-branching and right-branching bracketing of a structure is used to serve as a baseline for the resulting scores. The English language has been proven by both Brill(1993) and Klein(2005) to realize good results using a right-branching structure. This is illustrated with the example sentence "[I[am[responsible[for[my[rose]]]]]]". This can be translated into "[Pananagutan ko[ang[rosas ko]]]", which also implies a right biased heuristic for the Filipino language.

Each of the statistical measurements were used as the selection mechanism for the Greedy Selection Model and scored accordingly. As seen in Table 2, the occurrence measurement proves the better heuristic of the three, entropy was excluded in favor of scaled entropy. However, the difference between occurrence and scaled entropy is negligible. Mutual

Information produced significantly lower results, which can be stemmed back to the discrepancy identified in table 1.

Table 2. Precision, Recall and F1 Measures for various settings of Greedy Selection Model

	Precision	Recall	F1 Measure
Left branching	26.32%	37.69%	30.99%
Right branching	55.78%	62.17%	58.80%
Occurrence	70.91%	66.95%	68.87%
Raw Entropy	68.61%	62.05%	65.17%
Scaled Entropy	70.91%	66.40%	68.58%
Mutual Information	65.89%	55.32%	60.14%

The Greedy Selection Model emphasizes precision, a higher percentile of proposed constituents, but it identifies less bracketings thus a lower recall value. The precision of the Selection Model can be attributed to the accuracy of the selection metric in identifying the most likely constituent from the corpus.

The Greedy Selection Model identifies the locally optimum sequence to be bracketed per iteration. Once a sequence has been bracketed, this bracketing can not be undone, and results in the issues identified earlier with `CCB JJD` and `CCB JJD NNC`. To address this issue, the Constituent Context Model was designed that implements the Inside Outside algorithm.

Table 3. Precision, Recall and F1 Measures for various settings of Constituent Context Model

	Precision	Recall	F1 Measure
Left branching	26.32%	37.69%	30.99%
Right branching	55.78%	62.17%	58.80%
None	55.52%	60.38%	57.86%
Occurrence Weight	63.64%	69.20%	66.31%
Scaled Entropy Weight	66.73%	72.56%	69.52%
Mutual Information Weight	66.75%	72.58%	69.54%

Table 3 presents the results of the Constituent Context Model, with the inclusion of weighting system to increase bias towards sequences that are higher in the identified measurement.

The Constituent Context Model always results in a binary bracketing, increasing the amount of proposed constituents. This increases recall significantly, and the decrease in precision is not as significant, thus achieving the highest F1 measure, when weighted.

Without any additional heuristics, the output score of the Constituent Context Model is low. This is because, the model is highly reliant on a descriptive initial seed, and the probability generated from all possible binary trees is not sufficient to model the language. This produces lower results due to the lack of information concerning the actual sequences.

Occurrence count is an improvement to the basic model, but the commonly occurring sequences are given too high priority, and the high precision that was found in the Greedy Selection Model, is negated by the recall-oriented nature of the Constituent Context Model. Scaled entropy and mutual information produce the best overall results, because of the consideration for contextual information of each sequence.

The low precision is attributed to the fact that Filipino is relatively flatter than English. Consider the example sentence “[Pananagutan ko [ang [rosas ko]]]”, translated to “[I [am [responsible [for [my [rose]]]]]]”. The English translation produces a binary tree, while Filipino commonly produces a ternary tree, and it is even possible for a tree with 4 children to a single node.

The common problem identified with both models is handling nouns with describing phrases. Among the most frequently under proposed sequences is `NNC CCB NNC`. The `CCB NNC` serving as a phrase describing the primary `NNC`. In the example, `ng tren` describes the `riles`, and `ng riles` describes the `tagapaglihis`. Neither dependency is sufficiently identified by the models, rather `CCB NNC` is given priority on all occasions.

```
Gold:       QOT [sabi [ng [tagapaglihis [ng [riles [ng tren]]]]]]
Selection:  QOT [[[sabi [ng tagapaglihis]] [ng riles ]] [ng tren]]
CCM:        QOT [[[sabi [ng tagapaglihis]] [ng riles ]] [ng tren]]
English:    QOT [ said [ the [ railway signalman ] ] ]
```

5. Conclusion

Two models for the unsupervised constituent structure induction were presented. One of which is a Greedy Selection Model focused on maximizing the specified measurement on a local level. The other is a simplified Constituent Context Model modified to make use of identified heuristics in order to produce results taking into consideration the entire sentence structure. Both models are capable of estimating constituent structure rules from unbracketed corpora and producing promising scores.

The Greedy Selection Model was shown to achieve a precision rate of 70.9% and an overall performance of 68.9%. Analysis of the model shows that the occurrence of a sequence is currently the most effective measurement for identify constituency.

The Constituent Context Model estimates a globally optimized bracketing based on constituency probabilities placed on the sequences and contexts found. This model achieved a recall value of 72.6% and an overall performance of 69.5%, when applying either scaled entropy or mutual information as a weighting heuristic.

Experimentation identified that the Filipino language does not follow as strict a binary structure as English, but they are similar in the right-biased nature of the languages.

Future work include consideration of the dependency between words, incorporating the use of the right branching tree as a base case, improving on the statistical measurements identified, and implementing a slightly supervised approach to grammar induction.

References

Brill, E. 1993. Automatic grammar induction and parsing free text: A transformation based approach [online]. Proceedings of the 31st annual meeting on Association for Computational Linguistics: 1993, 259-265. Available: http://acl.ldc.upenn.edu/P/P93/P93-1035.pdf (April 1, 2008).

Charniak, E. 1996. Tree-bank grammars [online]. Proceedings of the Thirteenth National Conference on Artificial Intelligence: 1996, 1031-1036. Available: http://citeseer.comp.nus.edu.sg/cache/papers/cs/138/ftp:zSzzSzftp.cs.brown.eduzSzpubzSztechreportszSz96zSzcs96-02.ps.gz/charniak96treebank.ps.gz (April 1, 2008).

Chomsky, N. 1965. Aspects of the theory of syntax. MIT Press, Cambridge, MA.

Clark, A. 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering [online]. Proceedings of the 2001 workshop on Computational Natural Language Learning: 2001, 1-8. Available: http://wing.comp.nus.edu.sg/acl/W/W01/W01-0713.pdf (April 1, 2008).

Hopcroft, J., Motwani, R., & Ullman, J. 2001. Introduction to automata theory, languages, and computation. Addison-Wesley.

Lari, K. & Young, S. J. 1991. Applications of stochastic context-free grammars using the Inside-Outside algorithm [online]. Computer Speech and Language: 1991, 237-257. Available: http://64.233.179.104/scholar?hl=en&lr=&safe=off&q=cache:rVtjtefvq8J:staff.science.uva.nl/~jzuidema/temp/lariYoung90csl-insideOutside.pdf (April 1, 2008).

Klein, D. 2005. The unsupervised learning of natural language structure [online]. PhD thesis, Stanford University. Available: http://www.cs.berkeley.edu/~klein/papers/klein_thesis.pdf (April 1, 2008)

Klein, D. & Manning, C. D. 2001. Distributional phrase structure induction [online]. Proceedings of the Fifth Conference on Natural Language Learning: 2001, 113-120. Available: http://ucrel.lancs.ac.uk/acl/W/W01/W01-0714.pdf (April 1, 2008).

Klein, D. & Manning, C. D. 2002. A generative constituent-context model for improved grammar induction [online]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics: 2002, 128-135. Available: http://www-nlp.stanford.edu/~manning/papers/KleinManningACL2002.pdf (April 1, 2008).

Klein, D. & Manning, C. D. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency [online]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: 2004. Available: http://acl.ldc.upenn.edu/acl2004/main/pdf/341_pdf_2-col.pdf (April 1, 2008).

Kroeger, P. 1993. Phrase Structure and Grammatical Relations in Tagalog. Stanford: CSLI Publications.

Osborne, M. & Briscoe, T. 1997. Learning stochastic categorial grammar [online]. Proceedings of CoNLL97: Computational Natural Language Learning: 1997, 80-87. Available: http://acl.ldc.upenn.edu/W/W97/W97-1010.pdf (April 1, 2008).

Schachter, P. & Otanes, F. 1972. Tagalog reference grammar. Berkeley : University of California Press.

Sekine, S. & Grishman, R. 1995. A corpus-based probabilistic grammar with only two non-terminals [online]. Proceedings Fourth International Workshop on Parsing Technologies: 1995, 216-223. Available: http://www.cs.nyu.edu/~sekine/papers/iwpt95.pdf (April 1, 2008).

Steedman, M. 1993. Categorial Grammar. Lingua, 90: 221-258. Available: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1490&context=cis_reports