JMLR: Workshop and Conference Proceedings 34:??–??, 2014 Proceedings of the 12th International Conference on Grammatical Inference

Evaluation of selection in context-free grammar learning systems

Menno van Zaanen M.M.vanZaanen@uvt.nl

Nanne van Noord N.J.E.vanNoord@uvt.nl

Tilburg University, Tilburg, The Netherlands

Editors: Alexander Clark, Makoto Kanazawa and Ryo Yoshinaka

Abstract

Grammatical inference deals with learning of grammars describing languages. Formal grammatical inference aims at identifying families of languages that have a shared property, which can be used to prove efficient learnability of the families formally. In contrast, in empirical grammatical inference research, practical systems are developed that are applied to languages. The effectiveness of these systems is measured by comparing the learned grammar against a Gold standard which indicates the ground truth. From successful empirical learnability results, either shared properties may be identified, leading to further formal learnability results, or modifications to the systems may be made, improving practical results. Proper evaluation of empirical systems is, therefore, essential. Here, we evaluate and compare existing state-of-the-art context-free grammar learning systems (and novel systems based on combinations of existing phases) in a standardized evaluation environment (on a corpus of plain natural language sentences), illustrating future directions for empirical grammatical inference research.

Keywords: Context-free grammars, empirical grammatical inference, evaluation.

1. Introduction

The aim of grammatical inference (GI) research is to identify families of languages that can be learned efficiently. A language is described by a finite representation, called a grammar, hence the name grammatical inference. Exactly when a language is learned (by building or identification) and what counts as "efficiently" is described by the learning setting.

Two different methodological approaches to GI are typically identified: formal GI and empirical GI. Both approaches lead to knowledge about learnability of languages, but the methods used are quite different.

Research in the field of formal GI takes a mathematical approach. First, a family, i.e. a collection of languages that share a common property, is identified. Next, an algorithm that describes the learning process is designed. The algorithm makes use of the common property of the language family. Using the algorithm, efficient learnability within a particular learning setting is then proved mathematically.

Empirical GI typically also starts from a family of languages that share a common property. However, in contrast to formal GI, often this property is hard to describe using mathematics. For instance, one might be interested in learning the family of natural language syntax. The fact that these languages are spoken or written by people does not help

greatly in identifying a mathematical property that describes this family. Again, like in formal GI, an algorithm is applied to the data, but in empirical GI, this is done based on example data from one or more languages from the family of languages to be learned.

The major difference between the two approaches is how bias is handled. Formal GI starts by identifying a family of languages and then modeling the bias of the learning algorithm to allow it to learn the family. In most empirical GI tasks, the family of languages to be learned is not formally described beforehand, so this approach relies on the underlying idea, or perhaps hope, that the learning system has an appropriate bias that allows for the efficient learning of the languages in the family.

When the bias of the learning algorithm does not match with the family of languages to be learned, imperfect learning will take place. To measure how different the learned language is from the language to be learned, evaluation is of extreme importance for empirical GI. In order to assess and compare the performance of learning systems, a proper evaluation is essential. However, in the past different evaluation approaches have been used leading to incompatible results. Furthermore, incomplete descriptions of the evaluation approach also means that results cannot be compared directly.

Here, we mention existing evaluation strategies, followed by a brief description of the current state-of-the-art context-free GI systems. Next, we identify two phases that allow for the clustering of approaches of the GI systems. We then provide a standardized evaluation method, and evaluate and compare current state-of-the-art empirical GI systems. This work builds on **?** and evaluates complete GI systems.

2. Background

Learning context-free grammars has received much attention within the research area of empirical GI. The difficulty with these more powerful families of languages (in contrast to regular or sub-regular languages) is that checking language equivalence is undecidable. However, given that much empirical GI research is performed in the area of context-free grammars, alternative evaluation approaches are required. Here, we briefly describe these approaches and also sketch the GI systems that will be evaluated in Section **??**.

2.1. Evaluation strategies

Several evaluation strategies can be used to evaluate GI systems. **?**, pp. 58–62 made a first attempt to cluster existing strategies into three groups. Later, four groups have been recognized (**?**): "Looks-good-to-me", "Rebuilding known grammars", "Compare against a treebank", and "Language membership". For an extensive overview of these groups, see **?**. As far as we know, all evaluations found in other publications fall in one of these groups.

The evaluation in empirical GI typically measures the effectiveness of a system on a particular task with the ultimate aim to show perfect learning. Only more recently have comparisons of the results of different systems been performed. The first comparison of two completely different practical systems can be found in **?**, where the ABL and EMILE GI systems were compared in a standardized setting.

The comparison between ABL and EMILE relies on the "Compare against treebank" approach. In this approach, a (manually) structured dataset serves as the Gold standard, or ground truth. From this treebank, plain sequences are extracted, which are given as

input to a GI learning system. The output of the learning system, which is a structured version of the input sequences, can then be compared against the Gold standard.

The "Compare against treebank" approach, which is now the de-facto standard evaluation approach for empirical GI systems, focuses on measuring structural or strong equivalence. This means that the learned *structures* of the two systems are compared. Typically, this kind of evaluation is unlabeled, which means that any labels assigned to the structures are not taken into account during evaluation.

The "Language membership" approach is the only other approach that can be used if the underlying grammar (or family) is unknown. GI systems indicate whether sequences are part of the language or not, which means that it can be used if, for instance, the learned structure is not relevant or not consistent between systems to be compared. This approach is often used in competitions or shared tasks, e.g. (**??**). This approach does not measure structural equivalence, but aims to measure language or weak equivalence.

One of the problems of the "Compare against treebank" approach is that several choices have to be made, such as exactly which structures are counted, which may result in unfair comparisons in case different choices have been made. **?** compared many of the evaluation choices and proposed a standard evaluation setting.

2.2. Comparing state-of-the-art

To allow for a comparison of the empirical GI systems that are currently considered state-of-the-art, **?** attempted to structure and compare the existing systems. It turns out that, in the existing systems, two phases can be identified: *generation*, which introduces structures, and *selection*, which selects or prunes structures from the structures introduced in the generation phase. The generation phase can be greedy, introducing all possible structures, or more gentle, slowly adding structure if enough evidence can be found. The selection phase may be extensive, especially in the greedy generation phase, or even non-existent, for instance, combined with very gentle generation phases.

In **?**, the generation phases of the state-of-the-art systems are evaluated and compared. We will briefly describe these systems, as many of these systems (and new combinations of the generation and selection phases) will be evaluated here. First, we described the systems that rely on the greedy generation phase, followed by the systems that use a more gentle generation phase.

2.3. Greedy generation

Constituent-Context Model (CCM) (**?**) and U-DOP (**??**) are both systems that rely on a greedy generation phase. This phase assigns all binary structures on input sequences. Since *all* possible structures are introduced, there are structures that overlap, resulting in structures that cannot be the result from a parse of a context-free grammar (which can only describe tree structures). These ambiguous structures are temporarily stored, but in order to end up with context-free structures, this ambiguity needs to be removed. This means that the selection phase should make sure that conflicting structures should be resolved.

The difference between the two systems can be found in the selection phase. Essentially, both systems take a similar approach: identify the mostly likely structures given their occurrences in the (complete) data. Their difference is in how this likelihood is calculated.

Both systems start from a uniform probability distribution over the structures. CCM then applies the expectation maximization (EM) algorithm (**?**) to improve the probabilities of the structures given the data. U-DOP relies on the (stronger) statistical model of Data-Oriented Parsing (**??**).

2.4. Gentle generation

There are several systems that start with a gentle generation phase. In particular, we focus on the Alignment-Based Learning (ABL) system (**???**), which will be evaluated here. However, alternative systems such as EMILE **?**, and ADIOS (**??**) rely on the same underlying notion to introduce structure only when enough evidence can be found.

The linguistic notion of substitutability states that elements of the same type are substitutable. For instance, if a noun is replaced by another noun in a particular sentence, this results in another syntactically correct sentence. If sentences can be found that show evidence of substitution, this information may be used to assign structure.

The ABL system combines a gentle generation phase with a selection phase. During the generation phase, structure is assigned, but because this process may introduce some overlapping structure (just like in the greedy generation systems), a selection phase is added to resolve these conflicting structures. The selection of structure is based on simple relative maximum likelihood probabilities of the structures.

3. Experiments

To measure the performance of current state-of-the-art systems, we perform experiments of several systems within a standardized environment, making sure that the comparison is fair.[1] To be able to build on existing work, we follow the procedure as used in **?**, which only focused on the generation phase. This work is extended here by considering selection phases as well. By evaluating all combinations of generation and selection phases, we also introduce new empirical GI systems that have never been evaluated before.

3.1. Dataset

All GI systems are applied to the Wall Street Journal (WSJ) sections of the Penn Treebank 3 (**?**). This treebank contains newspaper articles, which may contain relatively long sentences. To be able to investigate the effect of sentence length, we make selections based on maximum sentence length. If we set the maximum sentence length to x, then WSJx describes all sentences of maximum length x. Previous work has mostly evaluated on WSJ10 (**??**). Here, we also provide information on WSJ20 and WSJ50 (an overview of the size of these datasets can be found in Table **??**) and show results by plotting information from WSJ3 to WSJ50.

To calculate the evaluation metrics, the entire dataset is used. For instance, evaluating WSJ10, all sentences up to a length of ten words are provided as input to the GI systems and the learned structures are compared against the corresponding Gold standard data.

1. In previous publications, results are not always comparable due to differences in evaluation choices.

Table 1: Properties of the WSJ treebank. Structures shows the number of pairs of brackets. Sentences displays the number of sequences containing tokens (or words). Types denotes the number of unique tokens.

Dataset	Structures	Sentences	Tokens	Tokens/sentence	Types
WSJ10	82,508	7,092	49,625	7.0	10,707
WSJ20	559,995	25,017	330,806	13.2	29,158
WSJ50	1,728,360	48,766	1,018,846	20.9	48,385

3.2. Metrics

Measuring the similarity of the structure between the Gold standard and the output of the learned systems is done using three well-known metrics taken from the field of information retrieval (**?**). Firstly, precision (Eq. **??**) measures the percentage of correctly learned structures (i.e., pairs of brackets).[2] Secondly, recall (Eq. **??**) shows the percentage of the Gold standard structures that are also found in the learned structures. Thirdly, F-score (Eq. **??**) is the geometric mean of precision and recall, providing an overall evaluation score.

$$\text{Precision} = \frac{\sum_{s \in structure} |correct(gold(s),\ learned(s))|}{\sum_{s \in structure} |learned(s)|} \tag{1}$$

$$\text{Recall} = \frac{\sum_{s \in structure} |correct(gold(s),\ learned(s))|}{\sum_{s \in structure} |gold(s)|} \tag{2}$$

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

3.3. Systems

We compare the performance of several systems. Due to space restrictions, we select two systems: one system that uses the greedy generation method and one that is based on a gentle generation method. These systems are split into their generation and selection phases and each combination of generation and selection phase is evaluated. Additionally, these results are compared against two baseline systems.

From the greedy generation systems, we selected CCM. For this system, an implementation (developed by **?**) is freely available. The system required minor modifications to allow for the system to perform the selection phase on data generated by non-greedy generation phases and on words rather than POS tags. The generation phase used in CCM is called "binary" and the selection phase is called "ccm".

From the gentle generation systems, the ABL system is used. An implementation of ABL is available (**??**), which allows both phases to be applied separately.[3] Here, we use the "wm" generation method and the "leaf" and "branch" selection methods as described in **?**.

2. In this evaluation, pairs of brackets are counted, so partially correct tree structures improve the evaluation scores.
3. We are aware of the implementations of EMILE and ADIOS. However, EMILE has a wide range of parameters that greatly influence the results and the free implementation ADIOS has limitations on the size of the training data.

The "wm" generation method relies on the edit distance algorithm (**?**) to identify alignments (words or phrases that can be found in two sentences). The "leaf" and "branch" selection methods compute probabilities for each of the introduced hypotheses (i.e., potential constituents or pairs of brackets) and resolves overlapping hypotheses by selecting only those hypotheses with the highest probabilities. The probability computed using the "leaf" method normalizes the number of times the words within the hypothesis occur together as a hypothesis by the total number of hypotheses. The "branch" method also takes the non-terminal of the hypothesis into account.

As generation phase baselines, we use the "left" and "right" systems, that assign binary tree structures that extend to the left or to the right respectively. Because these baselines result in proper tree structures (in contrast to the "binary" and "wm" systems), the best performing system ("right") is also used as baseline for the selection phase.

4. Results

For sake of completeness, we first briefly discuss the results of the generation phase, which can be found in **?**. We focus only on the best generation systems, which in turn form the basis of the evaluation of the selection phases. These results are provided following the discussion of the generation phase results.

4.1. Generation

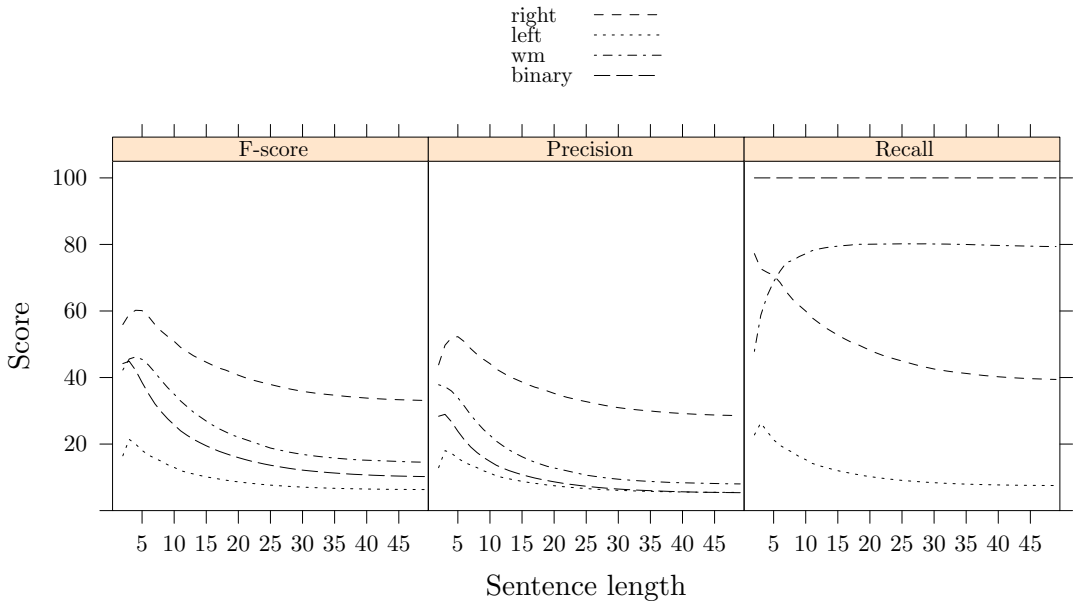

Figure 1: F-score, precision and recall results on subsets of WSJ dataset (the x-axes indicate the maximum sentence length of the subset) for a variety of generation systems.

Figure **??** shows the performance of the two generation phases ("wm" and "binary") as well as the two baselines ("left" and "right"). The different generation systems all have significantly different results ($p < .001$).

Looking at the two baselines ("right" and "left"), we see that the "right" baseline performs very well on all metrics. This shows that the syntax of English is mostly right-branching.

The greedy "binary" generation system has perfect recall. This shows that all possible structures are introduced. Many incorrect structures are also introduced, which can be derived from the low precision results. This should not be seen as a problem, as the task of the selection phase is to remove incorrect structures, which should improve precision while retaining high recall.

The recall of the gentle "wm" system improves when more data is available. This is to be expected because more evidence for substitutability can be found if more examples are present. However, the recall does not reach 100%, which is a problem, because the selection phase will not introduce any more structures, hence the recall cannot improve anymore. The precision is higher than that of the greedy generation system, which means that fewer incorrect structures are introduced.

4.2. Selection

Evaluating the selection systems can now be done based on the results of the different generation systems. Here, we will concentrate on the impact of the selection systems based on the two generation systems described in the previous section: "wm" and "binary".

Figure **??** shows the performance of the three selection systems ("leaf", "branch", and "ccm") as well as the output of the "binary" generation phase without selection, so with overlapping structures ("initial"). "Initial" corresponds to the upper bound on recall and should be seen as the lower bound on precision (as the aim of selection is to remove incorrect structures). Additionally, we show the "right" baseline, as that also produces tree structures and leads to the best results on the generation phase.

All three systems remove a large amount of possible structures. Some of these structures are correctly removed (increasing the precision over the "initial" system), but some correct structures are (incorrectly) removed as well, which is illustrated by the drop in recall with respect to "initial". Until around WSJ10, the selection systems do not translate in an improvement of the F-score for any of the systems and corpora containing with longer sentences, only minor improvements can be found.

When comparing the three selection systems, "ccm" performs best until approximately WSJ20. When longer sentences are found in the dataset, "leaf" performs best, closely followed by "branch". This is due to a stronger decline in both recall and precision for "ccm", showing that this method is more reliable on shorter sentences.

Figure **??** shows the performance of the three selection systems ("leaf", "branch", and "ccm") as well as the results before selection ("initial") and the right branching baseline ("right") on the output of the "wm" generation system.

Similarly to the performance on the "binary" generation output, the performance of all three systems drops as the sentences grow longer. However, there are some notable differences. Firstly, the F-score of all three systems does not come above the results of the

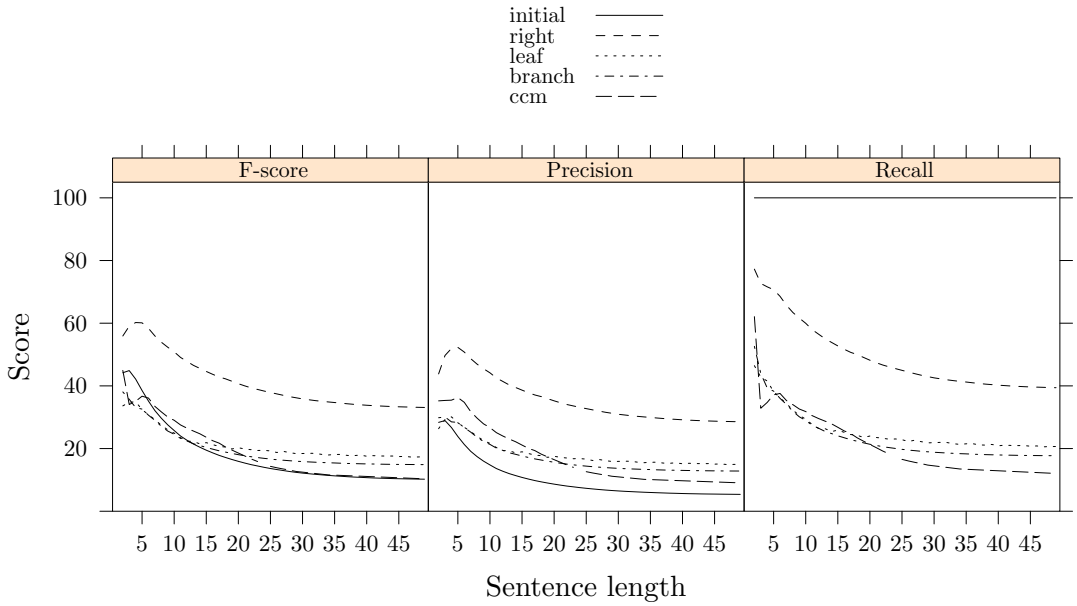

Figure 2: F-score, precision and recall results on subsets of WSJ dataset (the x-axes indicate the maximum sentence length of the subset) for a variety of selection systems based on the binary generation system.

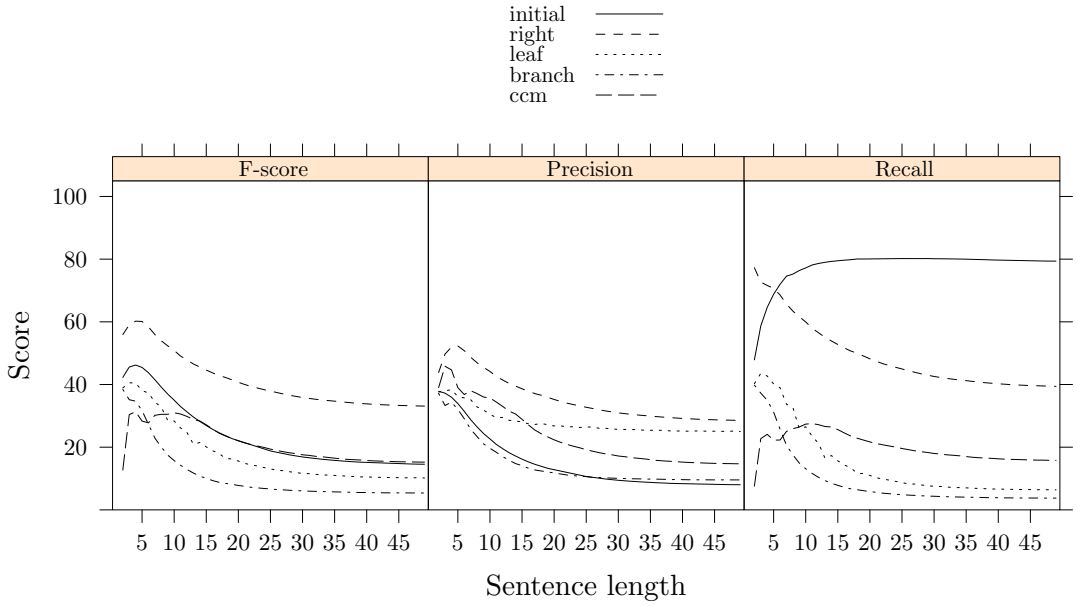

Figure 3: F-score, precision and recall results on subsets of WSJ dataset (the x-axes indicate the maximum sentence length of the subset) for a variety of selection systems based on the wm generation system.

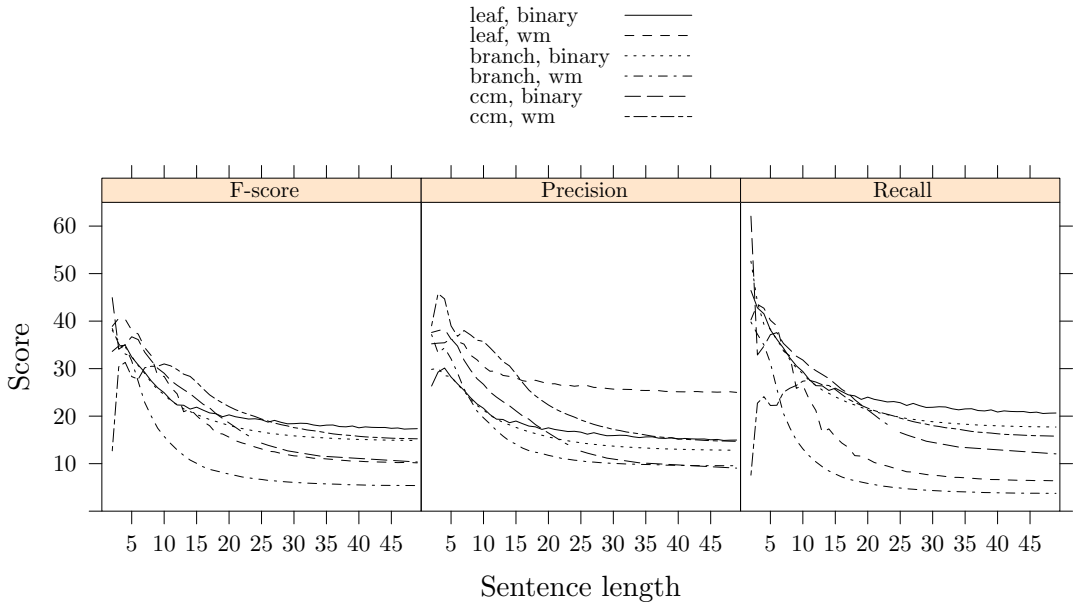

Figure 4: F-score, precision and recall results on subsets of WSJ dataset (the x-axes indicate the maximum sentence length of the subset) for a variety of selection systems based on the wm and binary generation systems.

generation phase, indicating that the drop in recall is not compensated by the increased precision. Secondly, the recall of "ccm" is higher than both "leaf" and "branch" on longer sentences (from around WSJ12), whereas for "binary" generation this was the other way around. We expect that a positive interaction between the "wm" and "ccm" systems may be the reason that the recall of "ccm" on longer sentence is somewhat higher. However, the precision of the "leaf" system outperforms the other selection systems. The precision of "branch" is approximately equal to the results of the generation phase, which means it does not remove any incorrect structures. Finally, the F-score results of the "ccm" method stays approximately the same as that of the "initial" system, and the other selection systems have lower results. This is also in contrast to the results of the "binary" generation phase, where "leaf" and "branch" outperform the "ccm" system. It is disappointing to see that none of the selection systems manage to outperform the (rather strong) "right" baseline.

To investigate the differences between the selection systems, Figure **??** shows the performance of the three selection systems ("leaf", "branch", and "ccm") on both the output of the "binary" and "wm" generation systems. All selection systems have significantly different results ($p < .001$). Both "leaf" and "branch" systems have highest precision when combined with the "wm" method, but the recall of that combination drops more with respect to the "binary" systems on longer sentences, leading to a lower F-score around WSJ10 (compared against the "binary" generation system). The same trend can be found for the "branch" selection system. For the "ccm" selection method, the reverse is true. Around WSJ10, the F-score of the system combined with the "wm" generation phase starts outperforming that

of the combination with "binary". Since these results may be difficult to identify from the figure, we also show F-score results of the systems in Table **??**.

5. Discussion

From the results presented in the previous section we can observe that the selection task becomes more complex as more data and longer sentences are processed, which negatively impacts the performance. This means that instead of (only) presenting results on WSJ10 as is done in the past, results including longer sentences are required to properly show the performance of the empirical GI system.

Instead of only evaluation existing systems, results of new system combinations have been presented here as well. The results of the combinations of "wm" and "ccm" as well as the "leaf" and "branch" combined with "binary" have not been published before. Interestingly, on longer sentences, the new "binary" and "leaf" combination outperforms all other systems.

The "ccm" model performs disappointingly (in particular on longer sentences). However, "ccm" has been designed to work with Part-of-Speech (POS) tags, rather than words. This has a major impact on the amount of possible types (i.e. unique words). As can be seen in Table **??** the amount of types increases as sentences grow longer. This could be a possible cause for the relatively poor performance of the "ccm" system. The underlying statistical model might perform better when there are fewer types to consider. Future work will have to show whether this is the case or not.

6. Conclusion

Evaluation of empirical GI systems is of extreme importance. In cases where perfect learning of language is (as of yet) unfeasible, it is essential to know how well the systems work.

By setting up a standardized evaluation environment, we have evaluated and compared existing state-of-the-art context-free grammar learning systems. We have identified generation and selection phases in each system, which means that new, previously untested, combinations can be evaluated as well. On datasets containing longer sentences, a novel combination of phases outperforms all other systems.

The differences in how the systems perform illustrates strengths and weaknesses of the existing approaches. Based on these results, new systems can be developed that build on the strengths, but try to resolve the weaknesses.

7. Future work

The work presented here may serve as the starting point for a range of future work directions. In particular, we realize that the research presented here only evaluates a selection of the existing system. Additional systems, including EMILE and ADIOS, can now be evaluated within this evaluation framework as well.

The evaluation of the CCM model as performed here may be considered unfair as CCM is designed to be evaluated on part-of-speech sequences instead of sequences of words. At the moment, this evaluation has not yet been performed, but given the framework, actually

Table 2: F-score results of generation systems (top part) and selection systems (bottom part) of sentence length 10, 20, and 50. Best results are highlighted.

Generation	Selection	10	20	50
right		**52.33**	**41.48**	**33.12**
left		13.77	8.87	6.33
wm		36.94	22.75	14.56
binary		27.57	16.52	10.23
wm	leaf	28.54	16.25	10.17
wm	branch	17.47	8.14	5.39
wm	ccm	**30.50**	**22.76**	15.23
binary	leaf	25.98	19.88	**17.36**
binary	branch	25.59	18.45	14.88
binary	ccm	30.07	19.52	10.35

performing the evaluation is straightforward. Additionally, evaluations based on part-of-speech sequences that have been learned in an unsupervised way may be done in the same way, allowing for a comparison of the impact of the quality of the part-of-speech tags.

GI systems that focus on learning dependency relations (such as the DMV model (**?**) or related systems, see, for instance, **?**) have not been evaluated within this framework either.

Another direction for future research is to use a range of treebanks, preferably in different languages. Initial results on Chinese (**?**) and Arabic (**?**) treebanks show that the evaluated systems show similar trends on other languages as well.

The structures that are the result of the GI systems are currently only evaluated against the structure of the Gold standard (found in the treebank). However, more generic information about the grammars may provide additional insight in the performance of the GI systems. For instance, the size of the grammar, theoretical or practical execution time and memory usage are also properties that may be important properties for instance when deciding on which GI system to use in a practical situation.