Learnability-based Syntactic Annotation Design

Roy Schwartz   Omri Abend   Ari Rappoport
Institute of Computer Science, Hebrew University of Jerusalem
{roys02|omria01|arir}@cs.huji.ac.il

Abstract

There is often more than one way to represent syntactic structures, even within a given formalism. Selecting one representation over another may affect parsing performance. Therefore, selecting between alternative syntactic representations (henceforth, *syntactic selection*) is an essential step in designing an annotation scheme. We present a methodology for syntactic selection and apply it to six central dependency structures. Our methodology compares pairs of annotation schemes that differ in the annotation of a single structure. It selects the more *learnable* scheme, namely the one that can be better learned using statistical parsers. We find that in three of the structures, one annotation is unequivocally better than the alternatives. Our results are consistent over various settings involving five parsers and two definitions of learnability. Furthermore, we show that the learnability gains incurred by our selections are both considerable (error reductions of up to 19.8%) and additive. The contribution of this work is in demonstrating that syntactic selection has a substantial and predictable effect on parsing performance, and showing that this effect can be effectively used in designing syntactic annotation schemes.

Keywords: Syntactic annotation design, Learnability, Parsing.

Proceedings of COLING 2012: Technical Papers, pages 2405–2422,
COLING 2012, Mumbai, December 2012.

1 Introduction

The formal manner in which syntactic relations are represented is at the core of the study of grammar. Numerous representations have been proposed over the years for expressing similar syntactic relations. This diversity of representations is expressed in a variety of syntactic annotation schemes currently in use in NLP. Examples include, for constituency annotation, schemes by (Marcus et al. 1993; Sampson 1995; Nelson et al. 2002, *inter alia)* and for dependency annotation, schemes by (Collins 1999; Rambow et al. 2002; Yamada and Matsumoto 2003; Johansson and Nugues 2007, *inter alia)*. Variation within the same formalism is expressed in structures that have several alternative annotations (henceforth *Varying Syntactic Structure* or *VSS*).

In this work we focus on dependency structures, where some of the most basic structures are VSS's. One example is prepositional phrases, which consist of a preposition followed by a noun phrase (e.g., "about everyone"). While some schemes select the preposition to head the NP (Collins 1999), others select the NP as the head of the preposition (Johansson and Nugues 2007) (see Figure 1). Other prominent VSS's include coordination structures and verb group constructions (see Section 3). In fact, more than 40% of the tokens in the Penn Treebank (Marcus et al. 1993) participate in at least one VSS (Schwartz et al. 2011).

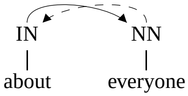


Figure 1: An example of a prepositional phrase – a Varying Syntactic Structure (VSS). Both annotation alternatives for this structure are plausible: either setting the preposition ("about" – solid line) as head, or the noun phrase ("everyone" – dashed line).

Despite the similar content represented by the alternative annotations to VSS's, selecting one over the other (*syntactic selection*) may have significant empirical implications. Previous work showed that syntactic selection can affect the parsing performance of a specific parser (see Section 7). In this work, we are the first to show that in some VSS's, syntactic selections improves parsing performance consistently across different parsers. As our findings are not parser-specific, they can be used to guide future syntactic annotation design.

The empirical implications of syntactic selection stem from the inter-relations between the VSS's and their surrounding structures. Figure 2 presents two alternative annotations for the sentence "he is sure about everyone". The alternatives differ in whether the preposition ("about") or the NP ("everyone") is selected to head the PP ("about everyone"). The two annotations can be deterministically derived from one another and express a similar syntactic relation, namely in both cases the PP is the complement of the adjective "sure". However, selecting one of the alternatives (the preposition) and not the other (the NP) results in a more learnable scheme.

Concretely, in dependency grammar, an adjective's complement is encoded by setting the head of the complement (either "about" or "everyone") as a dependent of the adjective ("sure"). It is plausible that a parser which is strongly guided by POS tags would not select an adjective ("sure") as the head of a noun ("everyone") as it is unlikely for adjectives to head nouns. This would result in a parsing error as in Figure 2(b). On the other hand, a similar parser would more likely select an adjective ("sure") to head a preposition ("about"), resulting in a correct parse as in Figure 2(a). Indeed, the MST parser (McDonald et al. 2005) exhibits such behavior.

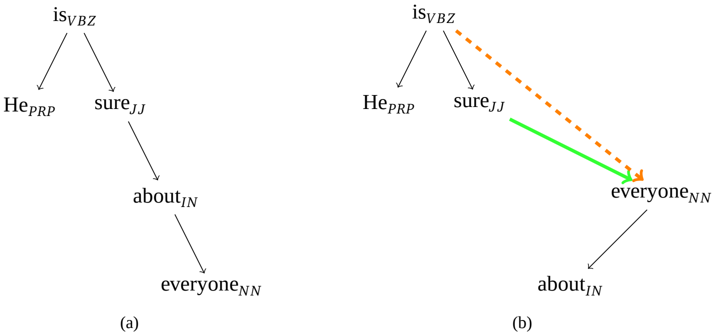


Figure 2: Exploring the effect of VSS annotation on neighboring structures. The sentence "He is sure about everyone", annotated when prepositions head PPs (Figure 2(a)) and when NPs head PPs (Figure 2(b)). Thin solid black lines mark gold+parser edges, thick green solid edges mark **gold edges**, and thick orange dashed lines mark **parser edges**.

The implications of syntactic selection underscore the importance of taking empirical considerations into account when designing an annotation scheme. In this work, we propose *learnability* as an empirical criterion for syntactic selection. The notion that more learnable schemes are preferable is motivated both practically and theoretically. Practically, more learnable schemes result in more accurate parsers. Theoretically, learnability has been a major consideration in the design of phrase structure grammar (Chomsky 2006), and can also be seen as a measure of simplicity, a fundamental principle in many other scientific fields (see Section 7).

We present a learnability-based methodology for syntactic selection and apply it to six central VSS's. We compare alternative annotations for each VSS, by examining pairs of schemes that differ only in their annotation of this VSS. For each pair, we pick the scheme that can be more easily learned using statistical parsers. We select an annotation for this VSS if we find that the schemes that use this annotation are consistently picked.

We experiment with five parsers of various types and using two different learnability measures. We obtain highly consistent results. Our experiments show that for three of the VSS's there is one alternative that is more learnable over all settings. That is, training any of the five parsers on an annotation scheme that uses the more learnable alternative results in higher parsing performance. The differences are substantial in magnitude, yielding error reductions that range between 2.4%-19.8%. Moreover, this gain is additive – using all three of the more learnable alternatives results in an even more accurate parser.

To further establish learnability as a coherent empirical criterion for syntactic selection, we show that our results are consistent with a parser-independent measure based on information theoretic notions.

The contribution of this paper is twofold. First, we present the first study focusing on syntactic selection and showing that it has a substantial and predictable effect on parsing performance. Second, we show that this effect can be used for designing syntactic annotation schemes. Specifically, our findings indicate that future dependency schemes should use (a) prepositions as heads of PPs (b) conjuncts as heads of coordination structures and (c) nouns (and not their determiners) as heads of NPs.

Section 2 describes our methodology. Section 3 discusses Varying Syntactic Structures (VSS). Experimental setup and results are described in Sections 4, 5. Section 6 discusses our methodology and presents further experiments that provide a wider context for understanding our findings. Section 7 surveys related work.

2 Methodology

We present a learnability-based methodology for selecting between alternative annotations for VSS's.

2.1 Notation

In the following we give a formal definition of an annotation scheme. We then turn to describe the different settings in which our methodology conducts experiments.

Our methodology experiments with a set S of VSS's. For each $s \in S$, we examine a set of alternative annotations. For clarity of presentation we assume each VSS has exactly two alternatives and denote them α_s, β_s. Let k denote the size of S ($k = 6$ in our experiments).

An annotation scheme is defined as a selection of an annotation for each of the structures in the language. It therefore includes a (fixed) annotation for non-VSS's, as well as a selected annotation for each of the VSS's. We can thus represent a scheme $\mathscr{A}$ as a k-tuple that selects one of the alternative annotations for each of the VSS's in S (all in all, 2^k schemes). Table 1 shows an example of two annotation schemes that differ in the annotation of exactly one structure (s_4).

Structure	s_1	s_2	s_3	$\boldsymbol{s_4}$	s_5	s_6
$\mathscr{A}_1$	α	β	α	$\boldsymbol{\alpha}$	α	β
$\mathscr{A}_2$	α	β	α	$\boldsymbol{\beta}$	α	β

Table 1: Applying our methodology to VSS's ($s_1, \ldots, s_6$): $\mathscr{A}_1, \mathscr{A}_2$ are annotation schemes that are identical in their annotation of all the VSS's but s_4 (bold red **column**). α, β are short for $\alpha_{s_i}, \beta_{s_i}$ respectively.

To obtain robust results, our methodology repeats each experiment in different settings, each determined by a parser and a learnability measure. We use P (L) to refer to the set of parsers (learnability measures). We use $|P| = 5, |L| = 2$, all in all $5 \times 2 = 10$ settings.

2.2 Learnability Measures

We propose two straightforward definitions of learnability. They are both defined with respect to a parser p and an annotation scheme $\mathscr{A}$ (as defined above). Both measures assume a fixed corpus partitioned into a training set and a test set.

The first measure is "Accuracy-Learnability". To compute it, we train p on the training set annotated according to $\mathscr{A}$, parse the test set, and evaluate it against the annotation determined by $\mathscr{A}$. We use attachment score for evaluation, which is the standard measure for dependency parsing evaluation. An annotation for which p receives a higher attachment score is considered more learnable.

The second measure is "Rate-Learnability" that measures the rate in which the different annotation schemes can be learned to a given accuracy. We define a target attachment score β. We train p on a corpus annotated with $\mathscr{A}$ several times, using an increasingly larger number of samples. We then evaluate the trained parser on our test data (annotated with $\mathscr{A}$) and create a learning curve of p and $\mathscr{A}$. An annotation for which p reaches β using less training samples is considered more learnable.

2.3 Learnability-based Methodology

We turn to describing a methodology for selecting learnable annotations for VSS's. The methodology runs a set of experiments, each using a parser p, a learnability measure l and a scheme $\mathscr{A}$. In each experiment, we compute the learnability of $\mathscr{A}$ with respect to p and l.

For every $s \in S$ and alternative annotations α_s, β_s, there are $2^k/2$ pairs of schemes that differ only in their annotation of s, one using α_s, and the other using β_s (see Table 1 for an example). Given a parser p and a learnability measure l, we compute the learnability of each pair of schemes and pick the more learnable scheme (see Table 2). We count the number of pairs in which the scheme using α_s is picked and the number of pairs in which the scheme using β_s is picked. We thus receive two figures that sum up to $2^k/2$ (32 in our experiments).

Annotation	s_1	s_2	s_3	s_4	s_5	s_6	score
$\mathscr{A}_1$	α	α	α	α	α	α	0.91
$\mathscr{A}_2$	α	α	α	α	α	β	**0.92**
$\mathscr{A}_3$	α	α	α	α	β	α	**0.94**
$\mathscr{A}_4$	α	α	α	α	β	β	0.935
$\vdots$							
$\mathscr{A}_{2^k-1}$	β	β	β	β	β	α	0.892
$\mathscr{A}_{2^k}$	β	β	β	β	β	β	**0.896**

Table 2: Applying our methodology for selecting a syntactic annotation for VSS s_6, under parser p and learnability measure l: each row in the table is an experiment with annotation scheme $\mathscr{A}_i$. The experiment compares the learnability (last column) of pairs of annotation schemes that differ only in their annotation of s_6 (where the annotations for $s_1, \ldots, s_5$ are fixed). For each pair of annotation schemes, the more learnable annotation for s_6 is in boldface (blue for $\boldsymbol{\alpha}$, red for $\boldsymbol{\beta}$).

We then define a significance value $1 \geq r \gg 0.5$. If one annotation (say α_s) is more learnable than the other (with respect to p, l) in a relative portion r of these pairs, we say that p is r-*biased* towards α_s with respect to l.

If for some $s \in S$, it holds that for every $p \in P, l \in L$, p is r-biased ($r \gg 0.5$) with respect to l towards the same annotation (say, α_s), we say there is a *unanimous* r-*bias* towards α_s. Consequently, α_s is the *empirically preferred annotation* of s.

3 Varying Dependency Structures

Varying syntactic structures are prevalent in many syntactic formalisms (see Section 7). In this section we focus on dependency structures.

Dependency structures receive varying annotation when the identity of the structure's head is debatable. This stems from the multiple, occasionally conflicting, criteria for defining a head. A few of the more generally acknowledged criteria for defining H to be the head of D in constituent C are (Kübler et al. 2009):

1. H determines the syntactic category of C and can often replace C.
2. H determines the semantic category of C; D gives semantic specification.
3. The form of D depends on H.

These definitions can often be applied to determine the identity of the head. For example, according to (1,2) a noun is the head of its modifying adjective (e.g., “cat” in “big cat”) and a verb is the head of its adverb (e.g., “eat” in “eat quickly”).

In VSS’s, these criteria are either inapplicable or conflicting. For example, in a sequence of proper nouns (e.g., “John Smith”), neither criterion is applicable. In a verb group construction (e.g., “can eat”), the main verb should be the head according to (2). On the other hand, the preceding modal restricts the main verb to be in infinitive form, and thus should be the head according to (3) (e.g., “he can eat” vs. “he eat**s**”).

Such structures have led to the creation of several dependency schemes, each taking a different approach to annotating them (Collins 1999; Rambow et al. 2002; Yamada and Matsumoto 2003; Johansson and Nugues 2007, *inter alia)*. We turn to describing the VSS’s that we experiment with and the alternatives annotations we consider for them. All of these annotations are in use in NLP. They are also plausible from a theoretical standpoint. Figure 3 shows a diagram for each of the structures, along with their possible annotations.

Coordination Structures are composed of two words, separated by a conjunction (e.g., “John and Mary”). It is not clear which token should be the head of this structure, if any (Nilsson et al. 2006). We consider two alternative annotations: (a) setting the conjunction as head, and both conjuncts as its dependents and (b) setting either of the conjuncts as head, selected according to the specific structure type (e.g., noun phrase, verb phrase).

Infinitive Verbs are verb phrases that contain the sequence “to” + infinitive verb (e.g., “to eat”). In (Yamada and Matsumoto 2003) the verb is the head, while in (Collins 1999) the “to” token is the head. We consider both annotations.

Noun Phrases that contain a determiner and a noun (e.g., “the apple” or “a dog”). Either the determiner (Bosco and Lombardo 2004) or the noun (Collins 1999) may serve as the head. We consider both annotations.

Noun Sequences are noun phrases that contain sequences of more than one noun (e.g., “John Doe”). Various alternative annotations for this structure include (Collins 1999), which takes the last noun as head, and BIO’s scheme which is somewhat more complex (Dredze et al. 2007). We consider either the rightmost or the leftmost noun as head, and mark all other nouns as its dependents.

Prepositional Phrases consist of a preposition and a noun phrase (e.g., “in a bag” or “of Rome”). Complement clauses that contain a subordinating conjunction (e.g., “after you go”) are also included[1]. Either the preposition/subordinating conjunction (Collins 1999) or the NP/clause (Johansson and Nugues 2007) can be the head. We consider both alternatives.

Verb Groups are composed of a verb and a modal verb (e.g., “can come”). Some schemes select the modal as head (Collins 1999), others select the verb (Rambow et al. 2002). We consider both

[1] For brevity, we use the term Prepositional Phrases to refer to both structures.

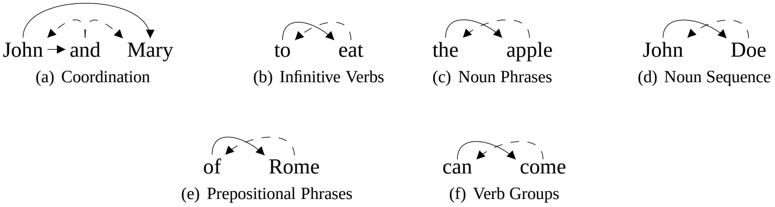


Figure 3: The VSS's with which we experiment. The possible annotations for each structure are marked using solid and dashed lines.

alternatives[2].

4 Experimental Setup

4.1 The Parsers

In this work we experiment with five parsers of different types. We briefly describe them.

Dependency Model with Valence (DMV) (Klein and Manning 2004) is a generative parser that defines a probabilistic grammar for unlabeled dependency structures. This parser is widely used in the field of *unsupervised* dependency parsing, where the great majority of recent works are in fact elaborations of this model (e.g., (Cohen and Smith 2009; Headden III et al. 2009)). In our experiments we use a *supervised* version of this parser, by training it using maximum likelihood estimation (MLE). This approach was used in various previous works as an upper bound for the unsupervised model (Blunsom and Cohn 2010; Spitkovsky et al. 2011). Decoding is performed using the Viterbi algorithm[3].

MST Parser (McDonald et al. 2005)[4] formulates dependency parsing as a search for a maximum spanning tree (MST). It uses online training and extends the Margin Infused Relaxed Algorithm (MIRA) (Crammer and Singer 2003) to learning with structured outputs.

Clear Parser (Choi and Nicolov 2009)[5] is a fast transition-based parser that uses the robust risk minimization technique (Zhang et al. 2002). k-best ranking is used to prune the next state in decoding.

S_u Parser (Nivre 2009)[6] is a transition-based parser and an extension of the MALT parser (Nivre et al. 2006). The parser starts by constructing arcs between adjacent words and then swaps the order of input words in order to learn more complex structures. It uses the *stackeager* algorithm, and is trained using various linear classifiers (including SVM).

NonDir Parser (Goldberg and Elhadad 2010)[7] is a non-directional, easy-first parser, which is greedy and deterministic. It first attempts to induce a non-directional version of the easiest arcs in

[2] Some definitions of verb groups also include auxiliaries. We choose to exclude them from our definition since we use the PTB POS set, which distinguishes modals, but not auxiliaries, from other verbs.

[3] http://www.cs.columbia.edu/~scohen/parser.html

[4] http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html

[5] http://code.google.com/p/clearparser/

[6] http://maltparser.org/

[7] http://www.cs.bgu.ac.il/~yoavg/software/easyfirst/

a dependency structure, and continues by iteratively selecting the best pair of neighbors to connect, until a complete dependency tree is created.

These parsers span the major approaches to statistical dependency parsing. The two main approaches are (Kübler et al. 2009) (a) *transition-based* methods that use state machines to map sentences to dependency graphs, attempting to reach the optimal state; and (b) *graph-based* methods, which try to find the best scoring dependency graph in some graph space. *Clear Parser* and S_u *Parser* are examples of (a), while *MST Parser* and *DMV* are examples of (b). NonDir takes a somewhat different parsing approach.

4.2 Technical Details

Following standard practice in English, used in the great majority of recent works, all the corpora are generated by converting constituency annotation to dependency using a set of head percolation rules[8]. Using these rules is also suitable here since they can easily be manipulated to create the different corpora required for applying our methodology.

Parsers are trained on sections 2–21 of the Penn TreeBank (PTB) WSJ corpus (Marcus et al. 1993), and are tested on section 23. We use the default feature set for each of the parsers. Evaluation is done using unlabeled attachment score, a common evaluation measure for dependency parsing.

For the Rate-Learnability measure, we select a different β value for each parser, due to their different performance levels; β is set to be the attachment score of the least learnable annotation for that parser, as determined by our experiments with the Accuracy-Learnability measure. This is the highest value of β that all schemes would reach at some point along their learning curve.

5 Results

Table 3 shows our results. In three out of the six structures, a strong unanimous bias is found. A unanimous 0.9-bias is found towards (a) selecting the preposition as head of prepositional phrases, and (b) selecting either of the conjuncts as head of coordination structures. A unanimous 0.7-bias is found towards the noun in noun phrases. For these structures, one annotation is clearly more learnable than the other, independently of the selected annotations for the other structures. This gives an empirical motivation for using these annotations.

In two of the remaining structures (verb groups and noun sequences), we find a trend towards one of the annotations; in five of the settings a 0.7-bias is found towards one alternative (modal and leftmost noun, respectively). In the other five settings no strong bias is found towards either alternative. In these structures, it might be the case that certain modeling assumptions incorporated into the parsers affect whether one alternative is preferred or not. This calls for a more detailed investigation, which we defer to future work.

Finally, no considerable bias is found in the infinitive verb structures, as a 0.7-bias towards any alternative is found in only one setting. Thus, our experiments suggest no preference towards either alternative in this case.

[8] We use a slightly modified version of the *pennconvertor*, tailored for our experimental setup (`http://nlp.cs.lth.se/software/treebank_converter/`) (Johansson and Nugues 2007).

Structure	*Setting / Annotation*	*DMV*		*MST*		*Clear*		S_u		*N.D.*		
		A.L.	*R.L.*	*A.L.*	*R.L.*	*A.L.*	*R.L.*	*A.L.*	*R.L.*	*A.L.*	*R.L.*	
Coord.	**CONJ**	32	30.5	32	32	32	32	32	32	32	32	◄
	CC	0	1.5	0	0	0	0	0	0	0	0	
Inf. Verbs	TO	16	17	19	17	21	17.5	25	19	18.5	13	
	VB	16	15	13	15	11	14.5	7	13	13.5	19	
NP	**NN**	24	24	32	24	32	23	32	24.5	30	23.5	◄
	DT	8	8	0	8	0	9	0	7.5	2	8.5	
N. Seq.	LEFT	25.5	24	29	21.5	32	31.5	21.5	18	11.5	12	
	RIGHT	6.5	8	3	10.5	0	0.5	10.5	14	20.5	20	
PP	**IN**	32	28.5	32	32	32	32	32	32	32	32	◄
	NP	0	3.5	0	0	0	0	0	0	0	0	
Verb Gr.	MD	32	23	28	20	23.5	20	15	17	24	17	
	VB	0	9	4	12	8.5	12	17	15	8	15	

Table 3: Exploring the learnability of the different annotation schemes. Each row pair corresponds to a pair of annotations for a given VSS, and each column pair corresponds to a parser, under Accuracy-Learnability (*A.L.*) and Rate-Learnability (*R.L.*) (see Section 2). For a given VSS, learnability measure and parser, we show the number of times one annotation is more learnable than the alternative. There are 32 experiments with each such combination, each has a single winner, resulting in a pair of numbers that sums up to 32. Gray cells mark settings in which the annotation is substantially more learnable than the alternative (dark/light gray correspond to $r = 0.9/0.7$ respectively). Rows marked with an arrow (◄) mark annotations that are *unanimously* biased. The annotations (see Section 3): Coordinations – headed by one of the conjuncts (CONJ) or by the conjunction (CC) ; Infinitive Verbs – headed by "to" (TO) or by the Verb (VB) ; Noun Phrases – headed by the noun (Noun) or by the determiner (DT) ; Noun Sequences – headed by the left/rightmost noun (LEFT/RIGHT); Prepositional Phrases – headed by the preposition (IN) or by the noun phrase (NP) ; Verb Groups – headed by the modal (MD) or by the Verb (VB). The Parsers (see Section 4.1): *DMV* (Klein and Manning 2004) ; *MST* (McDonald et al. 2005) ; *Clear* (Choi and Nicolov 2009) ; S_u (Nivre 2009) ; *N.D.* – NonDir (Goldberg and Elhadad 2010).

5.1 Analysis

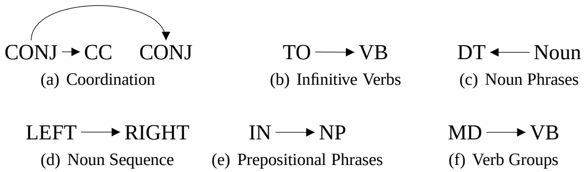
The empirically preferred annotations cannot be reduced to any simple, intuitive rule. For example, they do not match simple distinctions such as the one between closed and open classes: some of the more learnable annotations select closed class tags as heads (e.g., the preposition in prepositional phrases), while others select open class tags (e.g., the noun in noun phrases). Similarly, it is also not necessarily the rightmost or the leftmost word in the structure that is preferred.

Our results indicate that the biases are substantial. Table 4 shows that the difference between the accuracies of the most learnable annotation and the least learnable annotation for each parser under the Accuracy-Learnability measure. The accuracies range between 2.5-8.3%, which correspond to 22.2-35.3% error reduction. Table 4 also shows the the average performance gain from selecting each of the three empirically preferred annotations. These gains are substantial and yield error reductions that range between 3.7-19.8%, 2.4-4.8% and 7.4-15.3% for Coordinations, NPs and PPs respectively. Moreover, the gains are additive. That is, selecting all three of the empirically preferred annotations results in a gain similar to the sum of the average gains in the individual structures.

	Struct.	*DMV*	*MST*	*Clear*	S_u	*N.D.*	Err. Red.
Avg. Per. Diff.	Coord.	1.3%	1.2%	2.1%	1.6%	0.9%	**3.7-19.8%**
	NP	1.6%	0.2%	0.3%	0.5%	0.2%	**2.4-4.8%**
	PP	2.6%	1.6%	1.1%	1.0%	0.9%	**7.4-15.3%**
Best – Worst		8.3%	3.4%	4.2%	3.4%	2.5%	**22.2-35.3%**
Avg. Per.		66.2%	90.1%	90.2%	89.2%	90.4%	—

Table 4: The average performance gain incurred by selecting the empirically preferred annotations for the VSS's for which a unanimous bias is found. The last column is the error reduction range. The last row shows the mean attachment score of each parser when averaging over all schemes. The row before shows the difference between the lowest scoring and the highest scoring scheme for each parser. Annotation abbreviations (see Section 3): Coord. – Coordinations, NP – Noun Phrases, PP – Prepositional Phrases. Parser names are taken from Table 3.

Another natural question to ask is whether there is a single scheme that receives the highest score in all settings. We find that in fact this is the case. Figure 4 shows this scheme. The obtained scheme does not exactly match any of the commonly used annotation schemes, although it closely resembles that of (Collins 1999), differing only in the annotation of noun sequences. We note that since we addressed a particular set of VSS's, the winning scheme presented here is optimal only with respect to this selection.



(a) Coordination (b) Infinitive Verbs (c) Noun Phrases

(d) Noun Sequence (e) Prepositional Phrases (f) Verb Groups

Figure 4: The scheme that receives the highest score under all settings. Annotation abbreviations are taken from Table 3.

Correlation between Settings. We aim to show that our results are independent of the setting, and can therefore be seen as reflecting underlying phenomena. The parsers and the specific learnability measures can thus be seen as proxies by which these phenomena are observed.

In each setting (i.e., parser + learnability measure), we sort the different schemes according to their learnability (a total of 2^k values per ordering). The ten different settings (5 parsers × 2 learnability measures) yield ten relative orderings. To assess their similarity we compute the Kendall rank correlation coefficient (Kendall 1938)[9] between each pair of relative orderings ($\binom{10}{2} = 45$ pairs). The coefficient receives values in $[-1, 1]$, where 1 indicates equality, 0 indicates no correlation, and -1 indicates anticorrelation. We also compute a significance p-value, which is the probability for obtaining a given correlation at random (Abdi 2007).

Results show that the relative orderings obtained in the different settings are very much in concordance. The obtained Kendall τ correlation coefficients range between $(0.46, 0.88)$. Interestingly, when excluding DMV, results are even more significant (correlation in $(0.64, 0.88)$). This corresponds to p-values smaller than 10^{-7} and smaller than 10^{-13} if we exclude DMV.

[9] This is a commonly used measure in NLP (Lapata 2006; Brody and Kantor 2011).

Relation between Learnability Measures. In order to explore the relations between the two learnability measures, we focused on pairs of orderings that use the same parser, but different learnability measures ($|P| = 5$ pairs). The Kendall τ values in this case range between (0.75, 0.82), which corresponds to p-values $< 10^{-18}$.

Despite the high correlation between the measures, the biases discovered under the Accuracy-Learnability measure are stronger than the ones discovered under the Rate-Learnability measure. This demonstrates the somewhat different perspectives obtained by using different definitions of learnability.

6 Discussion

6.1 Syntactic Selection in a Wider Context

This paper presents a methodology for syntactic selection using learnability. The use of learnability is justified both for theoretical (see Section 7) and practical reasons, as it has direct implications on parsing technology. Namely, it is advantageous to train parsers on schemes that are inherently more learnable.

In the following we define a different, simplified empirical measure for syntactic selection and show that it correlates with learnability. The proposed measure is conceptually simpler than learnability and can therefore be used to partially explain the learnability results. However, it will be argued that learnability has several advantages over it as an empirical measure for syntactic selection.

We define the *predictability* of a scheme as minus the entropy of a parent given its child (unlike learnability, *predictability* is not defined with respect to a parser). We represent a word (x) as its POS tag, and a parent ($Pa(x)$) as the conjunction of its POS tag and the direction of the parent relative to the child (left or right). Concretely (P denotes the set of POS tags):

$$predictability = -H(Pa(x)|x) =$$

$$\sum_{x \in P} \sum_{\substack{Pa(x) \in \\ P \times \{L,R\}}} Pr(Pa(x)|x) \cdot \log Pr(Pa(x)|x)$$

Intuitively, *predictability* measures how easy it is to predict a word's head. On the face of it, higher *predictability* is likely to imply higher learnability. For example, if predictability is very high (i.e., the entropy is very low), words generally determine their parents, which facilitates learning. The opposite case, when predictability is very low, occurs when given a word, any other word is equally probable to serve as its parent. It is likely that such a scheme would be hard to learn.

We repeated the experiments described in Section 4, this time using predictability instead of learnability[10]. Results show that in the three structures where a unanimous learnability bias is found, a strong predictability bias is also found. Predictability yields similar results to learnability in the infinitive verb structure as well, both showing no strong bias. However, in the two other structures results diverge. In noun sequences, predictability shows a strong bias towards the left noun, while learnability showed a weaker trend (with no unanimous bias). In verb groups, a strong predictability bias is found in the opposite direction to the non-unanimous one found with learnability.

In addition, we derive a relative ordering of the different schemes (see Section 5.1). We compare this ordering to each of the orderings obtained in the learnability experiments. The obtained Kendall τ values range between (0.38, 0.66), which corresponds to p-values $< 10^{-4}$.

[10] Note that this time there is only one setting.

Predictability is a simple measure to understand and compute. The fact that it correlates with learnability can provide a partial explanation to the learnability results. However, it has several disadvantages compared to learnability. First, learnability relates directly to parsing technology, as parsers trained on more learnable schemes are likely to obtain higher results. Second, learnability is better motivated theoretically – it has been used extensively as a deciding factor in both linguistics and cognitive science (see Section 7). Third, predictability only quantifies a specific aspect of an annotation scheme (namely the POS tag and direction of the parent relative to its child), while parsers tend to take into account many other factors. These factors are captured by learnability.

Looking at our results, we observe that while correlations between predictability and learnability orderings are relatively high (mean Kendall τ value 0.51), they are generally lower than the correlations between the different settings of our learnability experiments (mean Kendall τ value 0.67). We conclude that predictability does give partial explanation to our results, but that further research is required in order to fully comprehend why exactly are some schemes more learnable than others.

6.2 The Methodology

Our methodology is designed for deciding between several alternatives, each having equal a-priori plausibility. It is therefore applicable for deciding between alternative annotations in VSS's.

Although we compare performance against different test sets, we find the comparison meaningful. Presumably, had there been no preference to either of the annotations, the performance on all these data sets should have been equivalent. Our experiments show that this is usually not the case, and by this reveal a non-trivial property of both the parser and, in those cases where the bias is unanimous, of the structures in question.

The consistent results obtained across five parsers using two learnability measures, which are in turn consistent with the results of a parser-independent predictability experiment, demonstrate the robustness of our results. However, it is still possible that a future parser will exhibit different patterns. Such a parser would very likely be fundamentally different, in some way, from the set of parsers used in this work. Our methodology can thus be used to discover an interplay between parser families and their empirically preferred annotations, an interesting topic in its own right.

Finally, we remark that learnability cannot by itself be used as a criterion for the quality of a scheme. For example, consider the simple right-branching scheme, where each word receives the word to its right as its head. It is trivial to learn despite its inferiority as a dependency scheme. We address this issue by applying our methodology only to compare between annotations that aim to represent the same structure and that were proposed as valid dependency annotation schemes. All considered schemes are derived by combining annotations to VSS's that were proposed in the literature (see Section 3). It is exactly because of the lack of consensus with regard to these structures that applying a complementary criterion, such as learnability, is required.

7 Related Work

7.1 Varying Syntactic Structures

The exact formal manner in which syntax should be represented has been the subject of endless debates. The diversity of approaches yielded a variety of annotation schemes for encoding similar structures.

Representational variation can be seen in virtually any formalism for syntactic annotation. In the field of POS tagging, the Brown Corpus (Francis 1964), the Penn Treebank (PTB) (Marcus et al.

1993), the British National Corpus (BNC) (Aston and Burnard 1998) and the SUSANNE corpus (Sampson 1995) all proposed different schemes for representing grammatical categories. Another example is the different annotation schemes used for noun compounds (Nastase and Szpakowicz 2003; Moldovan et al. 2004). In the field of constituency annotation, (Marcus et al. 1993; Sampson 1995; Kim et al. 2003) vary in the details of their representation of English syntax. Variation in dependency annotation, the focus of this paper, was discussed in (Ivanova et al. 2012) and is described in detail in Section 3. While these examples are all taken from English, variation is found in any language for which sufficient resources are available (Zeman et al. 2012).

Many previous works addressed the difficulties imposed by the lack of established standards for syntactic representation. Jiang and Liu (2009) adapted statistical tools trained with one annotation standard to another. Other works proposed to normalize the different representations into a standard scheme (Ide and Bunt 2010; Zeman et al. 2012). Parsing evaluation is also highly affected by VSS's. Schwartz et al. (2011) suggested Neutral Edge Direction (NED), an evaluation measure for unsupervised dependency parsing that accepts more than one plausible annotation for dependency VSS's. Tsarfaty et al. (2011) suggested a new evaluation measure for supervised dependency parsing to address representational variation. The measure is based on tree edit distance. Tsarfaty et al. (2012) extended this measure for comparing between annotations from different formalisms.

The emphasis of all the above works was mainly to overcome the problems incurred by the lack of standard, and not to select the most advantageous annotation according to some empirical criterion. In contrast, other works addressed the advantages some schemes have over their alternatives, and selected a scheme which best suited their needs.

One of the motivations behind the LTH dependency scheme (Johansson and Nugues 2007) was to facilitate semantic-role-labeling (SRL). They showed that an SRL tool that used their scheme performed better than a tool that used an alternative dependency scheme. While their method provides empirical reasons for using the LTH scheme, our work has a few advantages: first, our methodology examines each VSS individually, while they only compared an annotation scheme as a whole; second, while they performed the comparison on a single (basic) SRL tool, we compared the schemes on five different parsers (four of them state-of-the-art) using two definitions of learnability. Stanford dependencies (de Marneffe et al. 2006) were also designed using empirical considerations, namely to facilitate information extraction. However, they did not attempt to propose a methodology for syntactic selection.

Nilsson et al. (2006) modified the gold standard dependency annotations of two VSS's in order to improve parsing accuracy. They were able to improve performance by training a parser on a transformed corpus, parsing, and re-transforming the induced parse. While their work evaluated against a fixed gold standard, our work provides a methodology for designing an optimal gold standard with respect to learnability considerations. Furthermore, while they experimented with a single parser[11], our experiments use five parsers of different types and two learnability measures. As a result, their findings may be parser-specific, while our consistent results reveal a property of the scheme itself. Therefore, our results are directly applicable to annotation design. Last, our work is more extensive in terms of the number of examined structures (six vs. two). Kübler (2005) and Maier (2006) conducted similar experiments in constituency parsing.

[11] They experimented with two variants of the same parser.

7.2 Learnability

The notion that simpler or more *learnable* structures should be preferred is a recurring theme, both in theoretical linguistics (Chomsky 2006; Clark 2010) and more generally in the discussion of representations in cognitive science (Chater and Vitányi 2003). In the context of language learning, learnability refers to the question of what biases are required in order to learn a language, and in particular its grammar (Pinker 1989). In formal linguistics, learnability using distributional methods has been used as an important consideration in designing the phrase structure formalism (Chomsky 2006).

In Machine Learning, the term learnability refers to the question of whether, under certain assumptions, an underlying hypothesis may be learned given sufficient training samples (prominently, PAC-learnability (Valiant 1984)).

An empirical study by Perfors et al. (2011) used learnability considerations to decide between different syntactic formalisms. This line of research bears resemblance to model selection techniques in statistics, which aim to find which *model* best explains a fixed data set. Our work takes a similar direction. However, our methodology assumes the parsers are acceptable models for the given formalism, and tries to find the most suitable *annotation* from a set of a-priori equally likely alternatives. To the best of our knowledge, no previous work has tackled a similar task.

Predictability. Previous works used information theoretic measures to quantify sentence complexity, taking into account its syntactic representation. Hale (2006) explored a similar measure to predictability in the context of context-free-grammar. Hale (2001) and Levy (2008) explored a different measure ("surprisal"). These works demonstrated that their complexity measures correlate with human judgments on sentence comprehension difficulty.

Conclusion and Perspectives

In this paper we showed that selecting between alternative syntactic representations (syntactic selection) has a substantial and predictable effect on parsing performance. We presented a novel learnability-based methodology for syntactic selection and applied it to six central dependency structures that have several alternative annotations. Our methodology produced highly consistent results, and revealed a unanimity among all parsers in three of the structures. We showed that the gain from selecting the empirically preferred annotations is both substantial (error reduction of up to 19.8%) and additive. That is, selecting all three results in an even more accurate parser.

The higher learnability of the preferred annotations can be seen as an indication for their consistency with the rest of the scheme and has direct implications for parsing performance. We therefore suggest using the preferred annotations when designing future dependency schemes.

Future work will include applying our methodology to languages other than English, in order to assess whether the biases discovered in this work generalize cross-linguistically. We also plan to apply it to deciding between alternative annotations in other syntactic formalisms (such as constituency parsing) and in other NLP tasks such as POS tagging and noun-compound annotation.

Acknowledgments

Omri Abend is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. We would also like to thank Roi Reichart, Shai Shalev-Shwartz and Valentin I. Spitkovsky for their useful comments. We would also like to thank Richard Johansson for providing us with the code for the *pennconvertor* script.

References

Abdi, H. (2007). Kendall rank correlation. In Salkind, N. J., editor, *Encyclopedia of Measurement and Statistics*, pages 508–510. SAGE. 10

Aston, G. and Burnard, L. (1998). *The BNC handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press. 13

Blunsom, P. and Cohn, T. (2010). Unsupervised induction of tree substitution grammars for dependency parsing. In *Proc. of EMNLP*. 7

Bosco, C. and Lombardo, V. (2004). Dependency and relational structure in treebank annotation. In *Proc. of the Coling Workshop on Recent Advances in Dependency Grammar*. 6

Brody, S. and Kantor, P. (2011). Automatic assessment of coverage quality in intelligence reports. In *Proc. of ACL-HLT*. 10

Chater, N. and Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science. *Trends in Cognitive Science*, 7:19–22. 14

Choi, J. D. and Nicolov, N. (2009). K-best, locally pruned, transition-based dependency parsing using robust risk minimization. In Nicolov, N., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*, pages 205–216. John Benjamins, Amsterdam & Philadelphia. 7, 9

Chomsky, N. (2006). *Language and Mind*. Cambridge University Press, third edition. 3, 14

Clark, A. (2010). Three learnable models for the description of language. In *LATA*, pages 16–31. 14

Cohen, S. B. and Smith, N. A. (2009). Shared logistic normal distributions for soft parameter tying. In *Proc. of HLT-NAACL*. 7

Collins, M. J. (1999). *Head-driven statistical models for natural language parsing*. PhD thesis, University of Pennsylvania, Philadelphia. 2, 6, 10

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *JMLR*, 3:951–991. 7

de Marneffe, M. C., Maccartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proc. of LREC*. 13

Dredze, M., Blitzer, J., Talukdar, P. P., Ganchev, K., Graça, J. V., and Pereira, F. (2007). Frustratingly hard domain adaptation for dependency parsing. In *Proc. of the CoNLL 2007 Shared Task. EMNLP-CoNLL*. 6

Francis, W. (1964). *A standard sample of present-day English for use with digital computers*. Brown University. 12

Goldberg, Y. and Elhadad, M. (2010). An efficient algorithm for easy-first non-directional dependency parsing. In *Proc. of HLT-NAACL*. 7, 9

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proc. of NAACL*. 14

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30:643–672. 14

Headden III, W. P., Johnson, M., and McClosky, D. (2009). Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proc. of HLT-NAACL*. 7

Ide, N. and Bunt, H. (2010). Anatomy of annotation schemes: mapping to GrAF. In *Proc. of the Fourth Linguistic Annotation Workshop (LAW)*. 13

Ivanova, A., Oepen, S., Øvrelid, L., and Flickinger, D. (2012). Who did what to whom? a contrastive study of syntacto-semantic dependencies. In *Proc. of the Sixth Linguistic Annotation Workshop (LAW)*. 13

Jiang, W. and Liu, Q. (2009). Automatic adaptation of annotation standards for dependency parsing: using projected treebank as source corpus. In *Proc. of IWPT*. 13

Johansson, R. and Nugues, P. (2007). Extended constituent-to-dependency conversion for English. In *Proc. of NODALIDA*. 2, 6, 8, 13

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30:81–93. 10

Kim, J.-D., Ohta, T., Tateisi, Y., and ichi Tsujii, J. (2003). GENIA corpus – a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*. 13

Klein, D. and Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. of ACL*. 7, 9

Kübler, S. (2005). How do treebank annotation schemes influence parsing results? or how not to compare apples and oranges. In *Proc. of RANLP*. 13

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*. Morgan And Claypool Publishers. 5, 8

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484. 10

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. 14

Maier, W. (2006). Annotation schemes and their influence on parsing results. In *Proc. of the ACL Student Research Workshop*. 13

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330. 2, 8, 12, 13

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proc. of HLT-EMNLP*. 2, 7, 9

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., and Girju, R. (2004). Models for the semantic classification of noun phrases. In *Proc. of the HLT-NAACL Workshop on Computational Lexical Semantics*. 13

Nastase, V. and Szpakowicz, S. (2003). Exploring noun-modifier semantic relations. In *Proc. of IWCS*. 13

Nelson, G., Wallis, S., and Aarts, B. (2002). *Exploring natural language: working with the British component of the international corpus of English*. John Benjamins Pub. Co. 2

Nilsson, J., Nivre, J., and Hall, J. (2006). Graph transformations in data-driven dependency parsing. In *Proc. of ACL*. 6, 13

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *Proc. of ACL-IJCNLP*. 7, 9

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proc. of LREC*. 7

Perfors, A., Tenenbaum, J. B., and Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338. 14

Pinker, S. (1989). *Learnability and Cognition*. MIT Press. 14

Rambow, O., Creswell, C., Szekely, R., Tauber, H., and Walker, M. (2002). A dependency treebank for English. In *Proc. of LREC*. 2, 6

Sampson, G. (1995). *English for the Computer: The Susanne Corpus and Analytic Scheme*. Clarendon Press. 2, 13

Schwartz, R., Abend, O., Reichart, R., and Rappoport, A. (2011). Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proc. of ACL-HLT*. 2, 13

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2011). Punctuation: Making a point in unsupervised dependency parsing. In *Proc. of CoNLL*. 7

Tsarfaty, R., Nivre, J., and Andersson, E. (2011). Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *Proc. of EMNLP*. 13

Tsarfaty, R., Nivre, J., and Andersson, E. (2012). Cross-framework evaluation for statistical parsing. In *Proc. of EACL*. 13

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142. 14

Yamada, H. and Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proc. of IWPT*. 2, 6

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). HamleDT: To parse or not to parse? In *Proc. of LREC*. 13

Zhang, T., Damerau, F., and Johnson, D. (2002). Text chunking based on a generalization of winnow. *JMLR*, 2:615–637. 7