Treebank Parsing and Knowledge of Language: A Cognitive Perspective

Sandiway Fong (sandiway@email.arizona.edu)
Departments of Linguistics and Computer Science, University of Arizona, Douglass 200E
Tucson, AZ 85721 USA

Robert C. Berwick (berwick@csail.mit.edu)
Departments of EECS and Brain and Cognitive Sciences, MIT, 32D-728, 77 Massachusetts Ave.
Cambridge, MA 02139 USA

Abstract

Over the past 15 years, there has been increasing use of linguistically annotated sentence collections, such as the Penn Tree Bank (PTB), for constructing statistically based parsers. While these parsers have generally been built for engineering purposes, more recently such approaches have been advanced as potentially cognitively relevant, e.g., for addressing the problem of human language acquisition. Here we examine this possibility critically: we assess how well these Treebank parsers actually approach human/child language competence. We find that such systems fail to replicate many, perhaps most, empirically attested grammaticality judgments; seem overly sensitive, rather than robust, to training data idiosyncrasies; and easily acquire "'unnatural" syntactic constructions, those never attested in any human language. Overall, we conclude that existing statistically based treebank parsers fail to incorporate much "knowledge of language" in these three senses.

Keywords: statistical parsing; psycholinguistics; language acquisition.

Introduction

Recently there has been considerable interest in advancing stochastic parsing systems, trained on Treebank corpora, as putative solutions to cognitively relevant questions such as language acquisition and sentence processing. A representative overview of this position is provided by Chater and Manning (2006):

> "*Probabilistic methods are providing new explanatory approaches to fundamental cognitive science questions of how humans structure, process and acquire language. . . . Probabilistic models can account for the learning and processing of language, while maintaining the sophistication of symbolic models.*"

Sproat and Lappin (2005) take a particularly optimistic view on this point:

> "*. . . the proposal that general learning and induction mechanisms, together with minimal assumptions concerning basic linguistic categories and rule hypothesis search spaces are sufficient to account for much (perhaps all) of the language acquisition task.*"

This basic position has been echoed by many authors since nearly the beginning of the modern era of generative linguistics; see, e.g., Suppes (1970) and Levelt (1974) among many others for representative earlier statements; and Abney (1996); Bod, Hay, and Jannedy (2003); Manning (2003); and Lappin and Shieber (2007) for more recent claims along these lines. Attempts have also been made to tie the continuum of probability scores accompanying the phrase structure recovered by such systems to "gradience" with respect to both grammaticality and performance. (See the discussion in Crocker and Keller (2006) and references cited therein.) Taking this line of reasoning further, (Bod, 2003) concludes:

> "*Language displays all the hallmarks of a probabilistic system. Grammaticality judgments and linguistic universals are probabilistic and stochastic grammars enhance learning. All evidence points to a probabilistic language faculty.*"

This paper takes such "cognitive fidelity" claims seriously. How closely do these systems replicate human acquisition and knowledge of language rather than the standardly-used precision/recall information retrieval engineering metrics? Roughly, we view this alternative as a kind of cognitive "Turing test": how well do these systems mirror the knowledge of language that we know adults and children possess? Note that we can approach this question and still strive to remain neutral about linguistic theory: we need not adopt any particular linguistic account, but rather draw on an empirically valid list of behaviors that we know children and adults exhibit. While there are many conceivable tests to probe such abilities, to at least begin the investigation in this paper, we consider three cognitively relevant ones: (1) the actual knowledge of language attained; (2) extreme sensitivity to perturbation in training data; and (3) acquisition of non-natural regularities in training data. More precisely, we consider the following three evaluation dimensions:

1. Do such systems attain a cognitively plausible knowledge of language when trained on the standard dataset (in the language engineering field) of 39,832 sentences from the Wall Street Journal (WSJ) section of the Penn Treebank (PTB) (Marcus, Santorini, & Marcinkiewicz, 1994)? For example, can they distinguish, as people do, between grammatical and ungrammatical sentences in the form of known minimal pairs? Further, in the case of "ungrammatical" input, do they yield the "right" wrong parse as the most probable analysis?[1]

[1] We do not intend the term ungrammatical here to carry any particular formal or theoretical weight. We use it simply as a familiar cover term. We discuss some of the subtleties of this position below.

2. Statistical models have been advanced as a way to avoid the "brittleness" of symbolic systems (Abney, 1996). However, all statistical language models must deal with sparse data to achieve robustness. Given this state of affairs, is it in fact true that statistical models are robust, in the sense of being unaffected by minute perturbations in training data? We note in passing that such sensitivity has typically not been taken as reflecting the state of affairs in child language acquisition (see, e.g., Pinker (1984), among many other sources).

3. Finally, one would expect a cognitively faithful model to have the property of being able to more easily acquire a natural language than one that violates unnatural constraints, i.e. constraints not present in any natural language: for example, the artificially constructed language Epun (Smith, Tsimpl, & Ouhalla, 1993), in which a particular emphatic form is based on counting rather than syntactic structure. Can Treebank systems acquire such non-natural languages easily, in contradistiction to human performance?

While this list plainly does not begin to exhaust the range of possible probes into the "competence" of statistical parsers, this paper aims to stimulate discussion of additional cognitively relevant stress testing beyond the simple PARSEVAL-style evaluation, which scores parser output by counting bracketing matches with respect to "gold standard" presumed ground-truth phrase structure from a human-vetted treebank.[2]

Methods and Results

In the following sections, unless otherwise noted all reported experiments have been performed using Bikel's (2002) reimplementation of Collins's (2003) lexicalized, head-driven statistical parser, henceforth abbreviated as B-Collins. We make use of both the top-ranking parse output and the associated logprob score reported by parser for a given input sentence.[3] Parse results using the Berkeley parser (Petrov & Klein, 2007) are also reported for one experiment described below. Both parsers have the critical property of having been extensively trained on the same subset of the Wall Street Journal (WSJ) portion of the Penn Treebank (PTB) dataset, henceforth, referred to simply as PTB.

The Effect of Sentence Length

Care must be taken when comparing sentence probability scores as derived from the stochastic context-free grammar frameworks used by these systems. There is a strong inverse correlation between probability and sentence length. For instance, as illustrated in (1), with the logprob score given at the right. Despite the preponderance of transitive over intransitive verb frames in the PTB, the sentence length effect overwhelmingly dominates the verb subcategorization difference (logprob scores given in parentheses, closer-to-zero number referring to higher likelihood):[4]

(1) a. The circus amused the children (-26.378)
b. * The children amused (-20.951)
c. The circus affected the children (-30.979)
d. * The children affected (-26.415)

In the case of the psych-verb *amuse*, the (ungrammatical) inchoative form (1b) has a smaller negative magnitude logprob score, and thus a higher probability, than its causative counterpart (1a). A similar size logprob gap is also obtained for *affect*.[5]

Given the structure of these systems, since the probability score assigned to a particular parse is typically constituted from many thousands of individual decisions, their generative history, down to the level of individual words and their frequency of occurrence in the PTB, it is quite challenging to control for all possible contrasts.[6] However, for the purposes of this initial study, in order to properly compensate for this effect, we employed examples with the same number of words, or in the case of minimal pair comparisons involving unequal sentence length, we point out cases where the sentence length effect has been unexpectedly neutralized or counteracted, e.g. in the case where an ungrammatical example is scored lower than a corresponding grammatical (but longer) counterpart.[7] With this background in mind, we now turn to some specific cases.

Assessing Attained Knowledge of Language

***Wh*-Movement** Consider the permutations shown in (2) for a *wh*-question counterpart to *Bill will solve the problem*. Highest logprob scores are given in the right-most column; the values shown are for the top-ranked parse only. (Note all sentences are of the same length; indeed, they contain exactly the same lexical items, just in different orders.)

[2] It is also clearly true that there are many aspects of "knowledge of language" that such systems do acquire, viz., what they have been trained to learn, namely, high replicability of the PTB bracketing; aspects of predicate-argument structure, and the like. However, the goal in this paper is to focus on "stress tests" to where the systems must be improved so as to be more cognitively plausible – and perhaps even improved from an engineering standpoint, since such an approach has long been in the repertoire of standard software engineering best practice.

[3] The term "logprob score" refers to the (natural) logarithm of the calculated probability for the top-ranking parse. Probability values range from 0 to 1, and the corresponding logprob values scale from $-\infty$ (zero probability) to 0 (absolute certainty, probability 1).

[4] The verb form VB is immediately followed by a NP complement in 39% of the cases vs. 8% for no complement (Collins, 2003).

[5] In the PTB *amuse* occurs less often but is scored higher than *affect* (freq(*amused*)=1, freq(*affected*)=62) due to the fact that its frequency count falls below a predetermined threshold value for retention of frequency information (freq$<$ 6) during training. Accordingly, *amused* is scored as an "unknown" word. Note that this special "unknown" category is thus accorded more probability mass than a verb that occurs 62 times, because so many more items occurring fewer than 6 times will fall into this particular bin.

[6] Bikel (2004) notes: "*it may come as a surprise that the decoder needs to access more than 219 million probabilities during the course of parsing the 1,917 sentences of Section 00* [of the PTB]."

[7] As far as we have been able to determine, there is no straightforward mapping between logprob scores and some simple notion of grammaticality.

(2) a. Bill will solve which problem? (-41.058)
b. Which problem will Bill solve? (-56.858)
c. * Which problem Bill will solve? (-53.381)
d. * Bill solve which will problem? (-53.267)
e. * Bill solve which problem will? (-56.735)
f. * Which problem Bill solve will? (-62.3)

Not only is the *wh*-question (2b) dispreferred by B-Collins, it has the 2nd worst logprob score shown. In fact, in some cases, the top-ranked parse is not what a native speaker might have in mind. For example, B-Collins returns the parse shown in Figure 1 for (2f) with the modal part of speech tag for *will* re-tagged as a noun by the parser.[8]

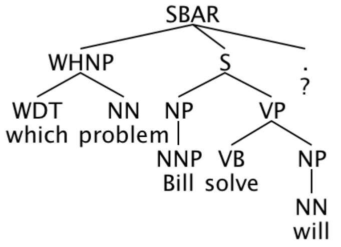

Figure 1: B-Collins parse for (2f).

Case Theory Consider the declarative and *wh*-question pairs in (3) below, again with logprob scores on the right.

(3) a. John is likely to win the race (-31.931)
b. * John is likely will win the race (-45.355)
c. Who is it likely will win the race? (-45.251)
d. * Who is it likely to win the race? (-36.299)

In the declarative raising case shown in (3a) and (3b), an infinitival (but not a tensed) embedded clause is permitted. However, exactly the reverse is true in the *wh*-question environment in (3c) and (3d) . The B-Collins parser exhibits a consistent preference for the infinitival environment, which works for the declarative case. Unfortunately, this also means that it ineluctably, and incorrectly, signals a preference for the embedded infinitival clause in the case of *wh*-questions. It cannot succeed in both situations.

Tense-marking ? (?) in a Linguistic Society of America pamphlet considered a "text reading" puzzle as an example of what is impossible for a computer to accomplish without knowledge of language: in particular, the task of determining the pronunciation of the orthographical form *read*, which can be pronounced as *red* or *reed* depending on context. The sentences considered by Jackendoff are reproduced in (4).

(4) a. The girls will read the paper. (*reed*)
b. The girls have read the paper. (*red*)
c. Will the girls read the paper? (*reed*)
d. Have any men of good will read the paper? (*red*)
e. Have the executors of the will read the paper? (*red*)
f. Have the girls who will be on vacation next week read the paper yet? (*red*)
g. Please have the girls read the paper. (*reed*)
h. Have the girls read the paper? (*red*)

It should be clear from the examples in (4) that a computer program needs to possess knowledge of the English auxiliary/main verb system along with basic properties of sentence phrase structure in order to correctly carry out this task. The PTB assumes a part of speech tagset that identifies and distinguishes among different forms of a verb as shown in Table 1, and these indeed ought to be sufficient, since these values suffice to fix a deterministic decision procedure to pronounce *read* correctly, as is evident.

Table 1: Penn Treebank (PTB) verb tagset.

Tag	**Description**	**Example**
VB	Verb, base form	*write*
VBD	Verb, past form	*wrote*
VBG	Verb, gerund or present participle	*writing*
VBN	Verb, past participle	*written*
VBP	Verb, non-3rd person singular present	*write*
VBZ	Verb, 3rd person singular present	*writes*

One might reasonably expect a stochastic parser trained on nearly 40,000 sentences to have acquired basic English sentence structure and properties of the auxiliary and verbal system, and thus be able to decode the examples in (4), correctly identifying the appropriate tag for *read* in each case, thereby solving the "text reading machine problem" posed by Jackendoff. For example, the parse tree recovered by the Berkeley parser in the case of (4b), correctly identifying *read* asVBN, is given in Figure 2. (In the case of *read*, only the VBD and VBN forms should be pronounced as *red*.)

However, this does not seem to be true. Figure 3 illustrates the corresponding parse for (4h). The sentence has been properly identified as an interrogative (category label SQ) but the parser has failed to assign the correct VBN tag to *read*. (The assigned tag VB will result in a pronounciation of *reed*.)

We summarize the results of the *read* pronunciation task in Table 2 (incorrectly tagged cases are starred (*)). As the results indicate, both parsers get 4 out of a total of 8 cases correct. For comparison, an assignment based purely on tag frequency would yield a crude baseline of 3 out of 8 correct on this task, as VB and VBN occur 45% and 19% of the time in the training set for *read*.

[8] It does not help to force B-Collins to retain the correct part of speech tag: the corresponding parse with *will* retaining its correct modal part of specch tag has an even less favorable logprob score of -64.695.

Table 2: Berkeley and B-Collins results for the *read* pronunciation task.

Example	(4a)	(4b)	(4c)	(4d)	(4e)	(4f)	(4g)	(4h)
Berkeley	VB	VBN	VB	*VB	*VB	*VB	VB	*VB
B-Collins	VB	VBN	VB	*VB	*VB	VBN	*VBN	*VB

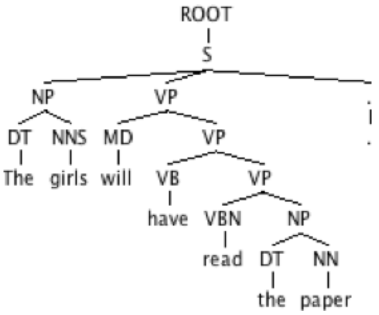

Figure 2: Berkeley parse for (4b).

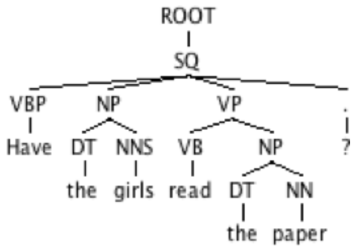

Figure 3: Berkeley parse for (4h).

One might properly ask here whether the “blame” for the incorrect results is due to improper tagging or rather the parser itself. We can control for this factor by forcing the system to use the “ground truth” tags.

Robustness and Sensitivity to Perturbation

It is sometimes tacitly assumed that statistical systems are inherently less brittle than symbolic models, in the sense that they can assign parses to less-than-grammatical input, as well as being robust in the face of input “noise” equivalently, low sensitivity to small alterations in a large training set. It is this latter property that it is explored below.

The Milk Example Consider the set of sentences in (5).

(5) a. Herman mixed the water with the milk
b. Herman mixed the milk with the water
c. Herman drank the water with the milk
d. Herman drank the milk with the water

Each of these sentences should receive the same basic parse, with the prepositional phrase (PP) headed by *with* attaching high, at the verb phrase (VP) level, as exhibited in Figure 4 in the case of sentence (5a).

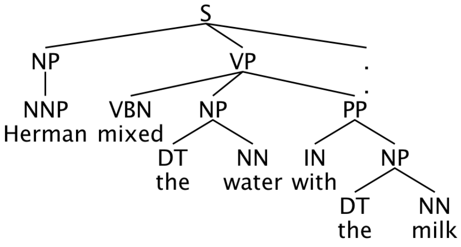

Figure 4: High attachment for (5a).

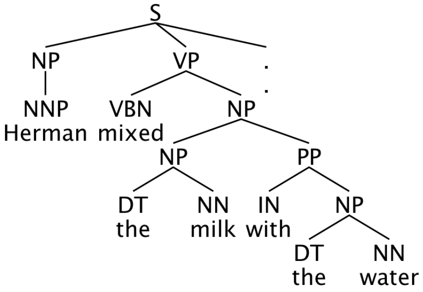

Figure 5: High attachment for (5b).

However, in the case of both (5b) and (5d), in which the order of *milk* and *water* is reversed, quite unexpectedly a low attachment for the PP is preferred by the parser. (Figure 5 displays the corresponding B-Collins parse for (5b).) Why? This bias is a property of the training set. It turns out there is exactly one sentence, reproduced in (6) below, where the PTB contains PP-attachment information for *milk*:[9]

(6) Borden even tested [$_{NP}$ [$_{NP}$ a milk] [$_{PP}$ with 4% butterfat]] in the South but decided the market was too small.

It is straightforward to verify that this one example can control low vs. high PP-attachment in the case of *milk*. For instance, we performed an experiment in which the only change to the PTB was that the PP *with 4% butterfat* in (6) was eliminated from the training set. After re-training, the attachment preference was reversed, i.e. high attachment obtains for (5b). Since the sentence in (6) is just one out of 39,832 training examples, this experiment indicates that B-Collins shows a surprising sensitivity to perturbation.

The true picture is actually even more unstable than described above. By cycling through combinations of verbs and

[9] There are a total of 24 sentences in the PTB training set containing *milk*, 21 as a noun.

nouns, one can get a range of different PP-attachment behaviors as shown in Table 3.

Table 3: Attachment sensitivity of *milk*.

Verb	**Noun**	**Attachment**	
		Verb + *milk with* + noun	Verb + noun + *with milk*
drank[10]	water	high	high
mixed	water	low	high
mixed	computer	low	low

This merely hints at a much broader problem in the training of such systems and the determination of high-low modifier attachment, a key problem that has drawn much psycholinguistic and computational attention over the years and which remains a source of a great deal of the error in parsing systems, no matter what metric is used. The problem is that statistically-trained systems must inherently rely on sparse data and smoothing. When examined more closely, we find that many modifier relations – like the PP attachment example above – are learned on the basis of a *single* example, fixing an alternative to some default attachment point. If the above experiment is on the right track, then each of these cases would prove to be data-sensitive; indeed, as mentioned, this is precisely one of the areas where current systems behave poorly. Deeper examination of such cases may reveal whether it will ever be possible to improve on modifier attachment without resort to additional data or resources.

Assessing Non-Natural Language Acquisition

In the following experiments we explore the question of how well Treebank-trained (and developed) parsers work when faced with non-natural training data. The basic strategy we employed was to modify the PTB training set by applying a series of basic (yet clearly humanly unattested) phrase-order transformations.

Experiment 1 Verb-complement constituent order can be viewed as a basic parameter of natural language: for head-initial languages such as English, verbs precede their complements; in head-final languages like Japanese, verbs follow their complements. For some verb-second languages, e.g. German, the verb must be the second phrase in matrix clauses, but head-final in subordinate clauses. However, in no natural language we are aware of does a speaker utter one sentence adhering to a head-initial parameterization, and then in the next sentence follow head-final order, in some such random fashion.[11] To emulate this unattested situation, in the following experiment we created a deliberately un-natural training set: we inverted verb-complement order so that every other sentence was verb-initial, and the intervening sentences verb-final. We then re-trained using this transformed Treebank.

Experiment 2 Another basic parameter of language often advanced is the constituent order of arguments and adjuncts. For example, in English, VP adjunct phrases tend to respect verb-complement adjacency, and are consistently attached at the edge of the VP (following the complement); thus we find *John ate the ice-cream while on the table* but much less frequently, *John ate while on the table the ice-cream*. To mimic this effect, in this second experiment, we swapped the order of arguments and adjunct phrases for every other sentence in the PTB, so that adjuncts become adjacent to the verb and arguments therefore non-adjacent.

Experiment 3 The final experiment simply combines the transforms of the two prior experiments, resulting in extremely "unnatural" sentence phrase structure, such as **the proposed changes also executives later and less often report exercises of options allow would*.

After training and testing following standard methods, the experimental results are summarized in Table 4:[12]

Table 4: B-Collins evaluation on non-natural training data.

Experiment	**Precision/Recall**	**F-measure**
Baseline (original data)	88.1 / 88.3	88.2
(1) Verb $\leftrightharpoons$ complement	88.7 / 86.7	87.7
(2) Adjunct $\leftrightharpoons$ argument	88.6 / 86.5	87.5
(1) + (2)	88.5 / 85.8	87.1

In each case, bracketing precision and recall are as defined in (Harrison & al, 1991) and was computed over the *same* set of held-out test sentences as in the original (unmodified) dataset.[13] It is evident that B-Collins appears to perform nearly as well after the training set is liberally sprinkled with extremely unnaturally modified PTB data. One possible reason for this may stem from the genre of the data employed: despite the size of the Treebank, WSJ sentences are relatively self-similar.[14]

Discussion

Let us now revisit the three basic questions outlined earlier and take stock of the results:

(1) Have state-of-the-art statistical parsers attained "knowledge of language"?

[11] This individual stochastic behavior has sometimes been suggested as an account of historical and/or idiolect variation, and, while logically possible, to the best of our knowledge remains speculative.

[12] The F-measure reported is the harmonic mean of bracketing recall and precision.

[13] As is standard in the PTB literature, training is performed on WSJ sections 2–21 (nearly 40,000 sentences), and evaluation on section 23 (approx. 2500 sentences).

[14] Indeed, this self-similarity is also supported by cross-validation analysis that yields nearly the same F-scores with as little as 10% of the training data; space does not permit the reproduction of these extensive results here.

Current state-of-the-art systems, such the B-Collins (and Berkeley) parser reviewed in this paper, score close to the 90%-level (on withheld PTB data) when evaluated on phrase structure bracketing fidelity (Collins, 2003).[15] However, merely being able to bracket sentences "accurately" evidently does not constitute full "knowledge of language." Rather, knowledge of language is multi-dimensional and cannot be conveniently summarized in terms of a single number, an F-measure. Similarly, grammaticality cannot be described in terms of a simple logprob score. Such conclusions may seem obvious from the outset, but the goal in applying the kinds of stress tests described in this paper is to discover exactly where these systems fail. As such, these experiments and test data are merely diagnostic aids. We have shown through stress testing over *wh*-questions, subject raising, and auxiliary fronting that these statistical parsing systems, despite access to 40,000 training examples, fail to learn many grammatical generalizations that all native speakers possess. The paper focused on certain cases of syntactica and lexical relation effects (as in PP attachment) because this has often been advanced as one of the strengths of such systems. (For reasons of space we have omitted many other similar constraints that also fail.) Indeed, the challenge would seem to be to discover, out of the very long list of grammatical generalizations that linguists have accumulated over the past sixty years, some certainly more valid than others, which, if any of these constraints these parsers do capture. The challenge for future research is whether these (or other similar) diagnostics can be exploited to advance the state-of-the-art in statistical parsing.

(2) Are statistical parsers 'robust?

As the milk example illustrates, the modification of a *single* example can overturn high/low modifier attachment preferences. Thus, these systems can be extremely fragile despite their inherently statistical nature. One possible reason for this stems from the size of the parameter estimation problem for training. All statistical parsers must employ a variety of smoothing methods to counteract the "sparse data problem" — methods for estimating phrase structure rule probabilities for which none, or very few, examples exist in the training set.

(3) Do statistical parsers mirror human limits on acquisition?

On the one hand, the fact that these systems often fail to acquire generalizations of the sort discussed earlier points to weaknesses in acquisition, despite state-of-the-art bracketing fidelity; on the other hand, the fact that these systems can perform robustly when constituent order parameterization is pseudo-randomized points to a non-human-like acquisition ability. Thus these systems seem at the same time to be both too weak and too strong. It is this lack of fit to human-like abilities, a "cognitive gap," that would seem most important to remedy if one indeed wants to take such systems seriously as models for human language acquisition and cognition.

[15] Bracketing is not the only possible evaluation metric. Predicate-argument and modifier-modifee (or dependency) relations are other clear choices, as has been discussed in the literature.

Acknowledgments

We thank Michael Coen, Ali Mohammed, and Beracah Yankama for assistance and valuable suggestions.

References

Abney, S. (1996). Statistical methods and linguistics. In J. Klavans & P. Resnik (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). Cambridge, Massachusetts: The MIT Press.

Bikel, D. M. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the second international conference on human language technology research* (pp. 178–182). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Bikel, D. M. (2004). Intricacies of Collins' parsing model. *Computational Linguistics*, *30*(4), 479–511.

Bod, R. (2003). *Is there evidence for a probabilistic language faculty?*

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. Cambridge, Massachusetts: MIT Press.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335–344.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Comput. Linguist.*, *29*(4), 589–637.

Crocker, M., & Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In G. Fanselow & et al (Eds.), *Gradience in grammar: Generative perspectives.* Oxford: Oxford University Press.

Harrison, P., & al et. (1991). Evaluating syntax performance of parser/grammars of English. In *Proceedings of the workshop on evaluating natural language processing systems.* Association for Computational Linguistics.

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, *43*(2), 393-427.

Levelt, W. J. M. (1974). *Formal grammars in linguistics and psycholinguistics (vol. 1): An introduction to the theory of formal languages and automata*. The Hague: Mouton.

Manning, C. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289–341). Cambridge, Massachusetts: MIT Press.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, *19*(2), 313–330.

Petrov, S., & Klein, D. (2007). Learning and inference for hierarchically split PCFGs. In *Aaai 2007 (nectar track).*

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Smith, N., Tsimpl, I.-M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua*, *91*, 279–347.

Sproat, R., & Lappin, S. (2005). *Re: A challenge to the minimalist community.* Linguist List 16.143.

Suppes, P. (1970). Probabilistic grammars for natural languages. *Synthese*, *22*, 95–116.