Empiricist Solutions to Nativist Puzzles by means of Unsupervised TSG

Rens Bod
Institute for Logic, Language & Computation
University of Amsterdam
Science Park 904, 1098XH Amsterdam, NL
rens.bod@uva.nl

Margaux Smets
Institute for Logic, Language & Computation
University of Amsterdam
Science Park 904, 1098XH Amsterdam, NL
margauxsmets@gmail.com

Abstract

While the debate between nativism and empiricism exists since several decades, surprisingly few common learning problems have been proposed for assessing the two opposing views. Most empiricist researchers have focused on a relatively small number of linguistic problems, such as *Auxiliary Fronting* or *Anaphoric One*. In the current paper we extend the number of common test cases to a much larger series of problems related to *wh-questions, relative clause formation, topicalization, extraposition from NP* and *left dislocation.* We show that these hard cases can be empirically solved by an unsupervised tree-substitution grammar inferred from child-directed input in the Adam corpus (Childes database).

1 Nativism versus Empiricism

How much knowledge of language is innate and how much is learned through experience? The *nativist* view endorses that there is an innate language-specific component and that human language acquisition is guided by innate rules and constraints ("Universal Grammar"). The *empiricist* view assumes that there is no language-specific component and that language acquisition is the product of abstractions from empirical input by means of general cognitive capabilities. Despite the apparent opposition between these two views, the essence of the debate lies often in the relative contribution of prior knowledge and linguistic experience (cf. Lidz et al. 2003; Clark and Lappin 2011; Ambridge & Lieven 2011). Following the nativist view, the linguistic evidence is so hopelessly underdetermined that innate components are necessary. This Argument from the Poverty of the Stimulus can be phrased as follows (see Pullum & Scholz 2002 for a detailed discussion):

(i) Children acquire a certain linguistic phenomenon
(ii) The linguistic input does not give enough evidence for acquiring the phenomenon
(iii) There has to be an innate component for the phenomenon

In this paper we will falsify step (ii) for a large number of linguistic phenomena that have been considered "parade cases" of innate constraints (Crain 1991; Adger 2003; Crain and Thornton 2006). We will show that even if a linguistic phenomenon is *not* in a child's input, it can be learned by an 'ideal' learner from a tiny fraction of child-directed utterances, namely by combining fragments from these utterances using the Adam corpus in the Childes database (MacWhinney 2000).

Previous work on empirically solving nativist puzzles, focused on a relatively small set of phenomena such as *auxiliary fronting* (Reali & Christiansen 2005; Clark and Eyraud 2006) and *Anaphoric One* (Foraker et al. 2009). Some of the proposed solutions were based on linear models, such as trigram models (Reali & Christiansen 2005), though Kam et al. (2008) showed that the success of these models depend on accidental English facts. Other empiricist approaches have taken the notion of structural dependency together with a

Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss, pages 10–18,
Avignon, France, April 24 2012.

combination operation as minimal requirements (e.g. Bod 2009), which overcomes the problems raised by Kam et al. (2008). Yet, it remains an open question which of the many other syntactic phenomena in the nativist literature can be acquired by such a general learning method on the basis of child-directed speech.

In this paper we will deal with a much larger set of problems than used before in empiricist computational models. These problems are well-known in the generativist literature (e.g. Ross 1967; Adger 2003; Borsley 2004) and are related to *wh-questions, relative clause formation, topicalization, extraposition* and *left dislocation*. It turns out that these hard cases can be learned by a simple unsupervised grammar induction algorithm that returns the sentence with the best-ranked derivation for a particular phenomenon, using only a very small fraction of the input a child receives.

2 Methodology

Our methodology is very simple: by means of an induced Tree-Substitution Grammar or TSG (see Bod 2009 for an in-depth study), we compute from the alternative sentences of a syntactic phenomenon reported in the generativist literature -- see below -- the sentence with the *best-ranked shortest derivation* (see Section 3) according to the unsupervised TSG. Next, we check whether this sentence corresponds with the grammatical sentence.

For example, given a typical nativist problem like auxiliary fronting, the question is: how do we choose the correct sentence from among the alternatives (0) to (2):

(0) Is the boy who is eating hungry?
(1) *Is the boy who eating is hungry?
(2) *Is the boy who is eating is hungry?

According to Adger (2003), Crain (1991) and others, this phenomenon is regulated by an innate principle. In our empiricist approach, instead, we parse all three sentences by our TSG. Next, the sentence with the best-ranked shortest derivation is compared with the grammatical expression.

Ideally, rather than selecting from given sentences, we would like to have a model that starts with a certain meaning representation for which next the best sentence is generated. In the absence of such a semantic component, we let our model select directly from the set of possible sentences as they are provided in the literature as alternatives, where we will mostly focus on the classical work by Ross (1967) supplemented by the more recent work of Adger (2003) and Borsley (2004). In section 9 we will discuss the shortcomings of our approach and suggest some improvements for future research.

3 Grammar induction with TSG: the best-ranked k-shortest derivation

For our induced grammar, we use the formalism of Tree-Substitution Grammar. This formalism has recently generated considerable interest in the field of grammar induction (e.g. Bod 2006; O'Donnell et al. 2009; Post and Gildea 2009; Cohn et al. 2010). As noted by Cohn et al. (2010) and others, this formalism has a number of advantages. For example, *its productive units (elementary trees of arbitrary size) allow for both structural and lexical sensitivity* (see Bod et al. 2003), while grammars in this formalism are still efficiently learnable from a corpus of sentences in cubic time and space.

As an example, figure 1 gives two TSG derivations and parse trees for the sentence *She saw the dress with the telescope*. Note that the first derivation corresponds to the shortest derivation, as it consists of only two elementary trees.

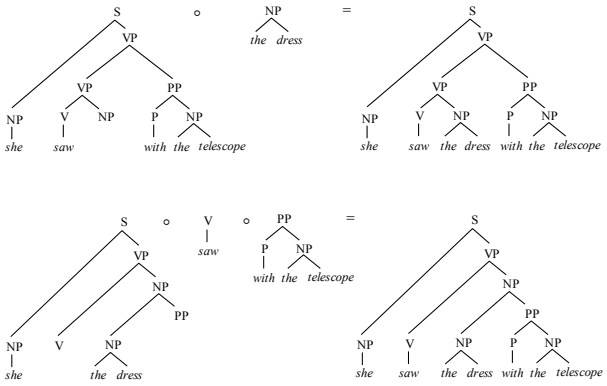

Figure 1. Two TSG derivations, resulting in different parse trees, for the sentence *She saw the dress with the telescope*

Our induction algorithm is similar to Bod (2006) where first, all binary trees are assigned to a set of sentences, and next, the relative frequencies of the subtrees in the binary trees (using a PCFG

reduction, see below) are used to compute the most probable trees. While we will use Bod's method of assigning all binary trees to a set of sentences, we will not compute the most probable tree or sentence. Instead we compute the k-shortest derivations for each sentence after which the sum of ranks of the subtrees in the k derivations determines the *best-ranked shortest derivation* (Bod 2000). This last step is important, since the shortest derivation alone is known to perform poorly (Bansal and Klein 2011). In Zollmann and Sima'an (2005) it is shown that training by means of shortest derivations corresponds to maximum likelihood training in the limit if the corpus grows to infinity.

Our approach to focus on the k shortest derivation rather than the most probable tree or most probable sentence is partly motivated by our different task: it is well-known that the probability of a sentence decreases exponentially with sentence length. This is problematic since, when choosing among alternative sentences, the longest sentence may be (the most) grammatical. Instead, by focusing on the (k-) *shortest derivations* this problem can – at least partly – be overcome.

From an abstract level, our grammar induction algorithm works as follows (see also Zollmann and Sima'an 2005). Given a corpus of sentences:

1. Divide the corpus into a 50% Extraction Corpus (EC) and a 50% Held out Corpus (HC).
2. Assign all unlabeled binary trees to the sentences in EC and store them in a parse forest.
3. Convert the subtrees from the parse forests into a compact PCFG reduction (Goodman 2003).
4. Compute the k-shortest derivations for the sentences in HC using the PCFG reduction.
5. Compute the *best-ranked derivation* for each sentence by the sum of the ranks of the subtrees (where the most frequent subtrees get rank 1, next most frequent subtrees get rank 2, etc., thus the best-ranked derivation is the one with the lowest total score).
6. Use the subtrees in the trees generated by the best-ranked derivations to form the TSG (following Zollmann & Sima'an 2005).

The learning algorithm above does not induce POS-tags. In fact, in our experiments below we test directly on POS-strings. This makes sense because the nativist constraints are also defined on categories of words, and not on specific sentences. Of course, future work should also parse directly with word strings instead of with POS strings (for which unsupervised POS-taggers may be used).

Rather than using the (exponentially many) subtrees from the binary trees to construct our TSG, we convert them into a more compact homomorphic PCFG. We employ Goodman's reduction method where each node in a tree is converted into exactly 8 PCFG rules (Goodman 2003). This PCFG reduction is linear in the number of nodes in the corpus (Goodman 2003, pp. 130-133).

The k-shortest derivations can be computed by Viterbi by assigning each elementary tree equal probability (Bod 2000). We follow the third algorithm in Huang and Chiang (2005), where first a traditional Viterbi-chart is created, which enumerates in an efficient way all possible subderivations. Next, the algorithm starts at the root node and recursively looks for the k-best derivations, where we used k = 100. In addition, we employed the size reduction technique developed in Teichmann (2011) for U-DOP/TSG.

We used all 12K child-*directed* utterances in the Adam corpus from the Childes database (MacWhinney 2000). These utterances come with POS-tags, which were stripped off the sentences and fed to our TSG induction algorithm. The child-directed sentences were randomly split into 50% EC and 50% HC. The subtrees from EC were used to derive a TSG for the POS-strings from HC. The resulting TSG consisted of 914,744 different subtrees. No smoothing was used. With the methodology explained in Section 2, we used this TSG to test against a number of well-known nativist problems from the literature (Ross 1967; Adger 2003).

It may be important to stress that the Adam corpus is based on only 2 hours of recordings per fortnight. This corresponds to just a tiny fraction of the total number of utterances heard by Adam. Thus our TSG has access only to this very small fraction of Adam's linguistic input, and we do not assume that our model (let alone a child) literally stores all previously heard utterances.

4 The problem of wh-questions

The study of wh-questions or wh-movement is one of oldest in syntactic theory (Ross 1967) and is usually dealt with by a specific set of "island constraints", where islands are constituents out of

which wh-elements cannot move. These constraints are incorporated in the more recent Minimalist framework (Adger 2003, pp. 389ff). Of course, our goal is different from Minimalism (or generative grammar in general). Rather than trying to explain the phenomenon by separate constraints, we try to model them by just one, more general constraint: the best-ranked (k-shortest) derivation. We do not intend to show that the constraints proposed by Ross, Adger and others are incorrect. We want to demonstrate that these constraints can also be modeled by a more *general* principle. Additionally, we intend to show that the phenomena related to wh-questions can be modeled by using only a tiny fraction of child-directed speech.

4.1 Unbounded scope of wh-questions

First of all we must account for the seemingly unbounded scope of wh-movement: wh-questions can have infinitely deep levels of embedding. The puzzle lies in the fact that children only hear constructions of level 1, e.g. (3), but how then is it possible that they can generalize (certainly as adults) this simple construction to more complex ones of levels 2 and 3 (e.g. (4) and (5))?

(3) who did you steal from?
(4) who did he say you stole from?
(5) who did he want her to say you stole from?

The initial nativist answer developed by Ross (1967) was to introduce a transformational rule with variables, and in the more recent Minimalist framework it is explained by a complex interplay between the so-called Phase Impenetrability Constraint and the Feature Checking Requirement (Adger 2003).

Our model proposes instead to build constructions like (4) and (5) by simply using fragments children heard before. When we let our induced TSG parse sentence (3), we obtain the following derivation consisting of 3 subtrees (where the operation 'o' stands for leftmost node substitution of TSG-subtrees). For reasons of space, we represent the unlabeled subtrees by squared brackets, and for reasons of readability we substitute the POS-tags with the words. (As mentioned above we trained and tested only with POS-strings.)

[X [who [X [did X]]] o [X [X from]] o [X [you steal]] =

[X [who [X [did [X [[X [you steal]] from]]]]]

Although this derivation is *not* the shortest one in terms of number of subtrees, it obtained the best ranking (sum of subtree ranks) among the 100-shortest derivations. In fact, the derivation above consists of three highly frequent subtrees with (respective) ranking of 1,153 + 7 + 488 = 1,648. The absolute shortest derivation (k=1) consisted of only one subtree (i.e. the entire tree) but had a ranking of 26,223.

Sentences (4) and (5) could also be parsed by combinations of three subtrees, which in this case were also the shortest derivations. The following is the shortest derivation for (4):

[X [who [X [did he say X]]] o [X [X from]] o [X [you stole]] =

[X [who [X [did he say [X [[X [you stole]] from]]]]]

It is important to note that when looking at the speech produced by Adam himself, he only produced (3) but not (4) and (5) – and neither had he heard these sentences as a whole. It thus turns out that our induced TSG can deal with the presumed unbounded scope of wh-questions on the basis of simple combination of fragments heard before.

4.2 Complex NP constraint

The first constraint-related problem we deal with is the difference in grammaticality between sentences (4), (5) and (6), (7):

(6) *who did you he say stole from?
(7) * who did you he want her to say stole from?

The question usually posed is: how do children know that they can generalize from what they hear in sentence (3) to sentences (4) and (5) but not to (6) and (7). This phenomenon is dealt with in generative grammar by introducing a specific restriction: the complex NP constraint (see Adger 2003). But we can also solve it by the best-ranked derivation. To do so, we compare sentences with the same level of embedding, i.e. (4) and (6), both of

level 2, and (5) and (7), of level 3. We thus view respectively (4), (6) and (5), (7) as competing expressions.

It turns out that (6) like (4) can be derived by minimally 3 subtrees, but with a worse ranking score. Similarly, (7) can also be derived by minimally 3 subtrees with a worse ranking score than (5). Since we tested on POS-strings, the result holds not only for these sentences of respective levels 2 and 3, but for all sentences of this type. Thus rather than assuming that the complex NP constraint must be innate, it can be modelled by recombining fragments from a fraction of previous utterances on the basis of the best-ranked derivation.

4.3 Left branch condition

The second wh-phenomenon we will look into is known as the Left Branch Condition (Ross 1967; Adger 2003). This condition has to do with the difference in grammaticality between (8) and (9):

(8) which book did you read?
(9) *which did you read book?

When we let our TSG parse these two sentences, we get the respective derivations (8') and (9'), where for reasons of readability we now give the substree-yields only:

(8') [X you read] o [which X] o [book did]

ranking: 608 + 743 + 8,708 = 10,059

(9') [which did X] o [you read book]

ranking: 12,809 + 1 = 12,810

Here we thus have a situation that, when looking at the 100-best derivations, the subtree ranking overrules the shortest derivation: although (9') is shorter than (8'), the rank of (8') nevertheless overrules (9'), leading to the correct alternative. Of course, it has to be seen whether this perhaps coincidentally positive result can be confirmed on other child-directed corpora.

4.4 Subject wh-questions

An issue that is not considered in early work on wh-questions (such as Ross 1967), but covered in the minimalist framework is the phenomenon that arises with subject wh-questions. We have to explain how children know that (10) is the grammatical way of asking the particular question, and (11), (12) and (13) are not.

(10) who kissed Bella
(11) *kissed who Bella
(12) *did who kiss Bella
(13) *who did kiss Bella

When we let our model parse these sentences, we obtain the following four derivations (where we give again only the subtree-yields):

(10') [who X] o [kissed Bella]

ranking: 22 + 6,694 = 6,716

(11') [X Bella] o [kissed who]

ranking: 24 + 6,978 = 7,002

(12') [did X Bella] o [who kiss]

ranking: 4,230 + 8,527 = 12,757

(13') [X kiss Bella] o [who did]

ranking: 4,636 + 2,563 = 7,199

Although all derivations are equally short, the best (= lowest) ranking score prefers the correct alternative.

4.5 Other wh-constraints modelled empirically

Besides the constraints given above, there are various other constraints related to wh-questions. These include:

- Sentential Subject Constraint
- WH-questions in situ
- Embedded WH-questions
- WH-islands
- Superiority
- Coordinate Structure Constraint

All but one of these constraints could be correctly modelled by our TSG, preferring the correct alternative on the basis of the best-ranked derivation and a fraction of a child's input. The only excep-

tion is the Coordinate Structure Constraint, as in (14) and (15):

(14) you love chicken and what?
(15) *what do you love chicken and?

Contrary to the ungrammaticality of (15), our TSG parser assigned the best rank to the derivation of (15). Of course it has to be seen how our TSG would perform on a corpus that is larger than Adam. Moreover, we will see that our TSG can correctly model the Coordinate Structure Constraint for other phenomena, even on the basis of the Adam corpus.

5 The problem of Relative clause formation

A phenomenon closely related to wh-questions is relative clause formation. As in 4.2, generativist/nativist approaches use the same complex NP constraint to distinguish between the grammatical sentence (16) and the ungrammatical sentence (17). The complex NP constraint is in fact believed to be universal.

(16) the vampire who I read a book about is dangerous
(17) *the vampire who I read a book which was about is dangerous

In (16), the 'moved' phrase `the vampire' is taken out of the non-complex NP `a book about <the vampire>; in (17), however, `the vampire' is 'moved' out of the complex NP `a book which was about <the vampire>', which is not allowed.

Yet our TSG could also predict the correct alternative by means of the best ranked derivation alone, by respectively derivations (16') and (17'):

(16') [the vampire X is dangerous] o [who I read X] o [a book about]

ranking: 1,585,992 + 123,195 + 5,719 = 1,714,906

(17') [the vampire X is dangerous] o [who I read X] o [a book which X] o [was about]

ranking: 1,585,992 + 123,195 + 184,665 + 12,745 = 1,906,597

Besides the complex NP constraint, the phenomenon of relative clause formation also uses most other constraints related to wh-questions: Left branch condition, Sentential Subject Constraint and Coordinate Structure Constraint. All these constraints could be modelled with the best-ranked derivation – this time including Coordinate structures (as e.g. (18) and (19)) that were unsuccessfully predicted by our TSG for wh-questions.

(18) Bella loves vampires and werewolves who are unstable
(19) *werewolves who Bella loves vampires and are unstable

6 The problem of Extraposition from NP

A problematic case for many nativist approaches is the so-called "Extraposition from NP" problem for which only ad hoc solutions exist. None of the constraints previously mentioned can explain (20) and (21):

(20) that Jacob picked Bella up who loves Edward is possible
(21) *that Jacob picked Bella up is possible who loves Edward

As Ross (1967), Borsley (2004) and others note, the Complex NP Constraint cannot explain (20) and (21), because it applies to elements of a sentence dominated by an NP, and here the moved constituent `who loves Edward' is a sentence dominated by an NP. Therefore, an additional concept needs to be introduced: `upward boundedness', where a rule is said to be upward bounded if elements moved by that rule cannot be moved over the boundaries of the first sentence above the elements being operated on (Ross 1967; Borsley 2004).

Thus additional machinery is needed to explain the phenomenon of Extraposition from NP. Instead, our notion of best ranked derivation needs no additional machinery and can do the job, as shown by derivations (20') and (21'):

(20') [X is possible] o [that Jacob picked X] o [Bella up X] o [who loves Edward]

ranking: 175 + 465,494 + 149,372 + 465,494 = 1,080,535

(21') [X is possible X] o [that Jacob picked X] o [Bella up] o [who loves Edward]

ranking: 3,257 + 465,494 + 176,910 + 465,494 = 1,111,155

7 The problem of Topicalization

Also the phenomenon of Topicalization is supposed to follow the Complex NP constraint, Left branch condition, Sentential Subject Constraint and Coordinate Structure Constraint, all of which can again be modelled by the best ranked derivation. For example, the topicalization in (22) is fine but in (23) it is not.

(22) Stephenie's book I read
(23) * Stephenie's I read book

Our TSG predicts the correct alternative by means of the best ranked derivation:

(22') [X I read] o [Stephenie's book]
ranking: 608 + 2,784 = 3,392

(23') [Stephenie's X book] o [I read]
ranking: 3,139 + 488 = 3,627

8 The problem of Left dislocation

The phenomenon of Left dislocation provides a particular challenge to nativist approaches since it shows that there are grammatical sentences that do not obey the Coordinate Structure Constraint (see Adger 2003; Borsley 2004). A restriction that is mentioned but not explained by Ross (1967), is the fact that in Left dislocation the moved constituent must be moved to the left of the main clause. Instead, movement merely to the left of a subordinate clause results in an ungrammatical sentence. For example, (24) is grammatical, because `Edward' is moved to the left of the main clause. Sentence (25), on the other hand, is ungrammatical, because `Edward' is only moved to the left of the subordinate clause `that you love <Edward>'.

(24) Edward, that you love him is obvious
(25) *that Edward, you love him is obvious

Our TSG has no problem in distinguishing between these two alternatives, as is shown below:

(24') [Edward X is obvious] o [that you love him]
ranking: 590,659 + 57,785 = 648,444

(25') [that X is obvious] o [Edward you love him]
ranking: 876,625 + 415,940 = 1,292,565

9 Discussion and conclusion

We have shown that an unsupervised TSG can capture virtually all phenomena related to wh-questions in a simple and uniform way. Furthermore, we have shown that our model can be extended to cover other phenomena, even phenomena that fall out of the scope of the traditional nativist account. Hence, for at least these phenomena, Arguments from Poverty of Stimulus can no longer be invoked. That is, step (ii) in Section 1 where it is claimed that children cannot learn the phenomenon on the basis of input alone, is refuted.

Phenomenon	**Succesful?**
Subject Auxiliary Fronting	yes
WH-Questions	
Unbounded Scope	yes
Complex NP Constraint	yes
Coordinate Structure Constraint	no
Left Branch Condition	yes
Subject WH-questions	yes
WH in situ	yes
Superiority	yes
Extended Superiority	yes
Embedded WH-questions	yes
WH-islands	yes
Relative Clause Formation	
Complex NP Constraint	yes
Coordinate Structure Constraint	yes
Sentential Subject Constraint	yes
Left Branch Condition	yes
Extraposition from NP	yes
Topicalization	
Complex NP Constraint	yes
Coordinate Structure Constraint	yes
Sentential Subject Constraint	yes
Left Branch Condition	yes
Left Dislocation	
Coordinate Structure Constraint	yes
Restriction	yes

Table 1. Overview of empiricist solutions to nativist problems tested so far (using as input the child-directed sentences in the Adam corpus of the Childes database), and whether they were successful.

Table 1 gives an overview of all phenomena we have tested so far with our model, and whether they can be successfully explained by the best-ranked k-shortest derivation (not all of these phenomena could be explicitly dealt with in the current paper).

Previous empiricist computational models that dealt with learning linguistic phenomena typically focused on auxiliary fronting (and sometimes on a couple of other problems – see Clark and Eyraud 2006). MacWhinney (2004) also describes ways to model some other language phenomena empirically, but this has not resulted into a computational framework. To the best of our knowledge, ours is the first empiricist computational model that also deals with the problems of wh-questions, relative clause formation, topicalization, extraposition from NP and left dislocation.

Many other computational models of language learning focus either on inducing syntactic structure (e.g. Klein and Manning 2005), or on evaluating which sentences can be generated by a model with which precision and recall (e.g. Bannard et al. 2009; Waterfall et al. 2010). Yet that work leaves the presumed 'hard cases' from the generativist literature untouched. This may be explained by the fact that most empiricist models do not deal with the concept of (absolute) grammaticality, which is a central concept in the generativist framework. It may therefore seem that the two opposing approaches are incommensurable. But this is only partly so: most empiricist models do have an implicit notion of relative grammaticality or some other ranking method for sentences and their structures. In some cases, like our model, the top-ranking can simply be equated with the notion of grammaticality. In this way empiricist and generativist models *can* be evaluated on the same problems.

There remains a question what our unsupervised TSG then exactly explains. It may be quite successful in refuting step (ii) in the Argument from the Poverty of the Stimulus, but it does not really explain where the preferences of children come from. Actually it only explains that these preferences come from child-directed input provided by caregivers. Thus the next question is: where do the caregivers get their preferences from? From *their* caregivers -- ad infinitum? It is exactly the goal of generative grammar to try to answer these questions. But as we have shown in this paper, these answers are motivated by an argument that does not hold. Thus our work should be seen as (1) a refutation of this argument (of the Poverty of the Stimulus) and (2) an alternative approach that can model all the hard phenomena on the basis of just one principle (the best-ranked derivation). The question where the preferences may eventually come from, should be answered within the field of language evolution.

While our TSG could successfully learn a number of linguistic phenomena, it still has shortcomings. We already explained that we have only tested on part of speech strings. While this is not essentially different from how the nativist approach defines their constraints (i.e. on categories and functions of words, not on specific words themselves), we believe that any final model should be tested on word strings. Moreover, we have tested only on English. There is a major question how our approach performs on other languages, for example, with rich morphology.

So far, our model only ranks alternative sentences (for a certain phenomenon). Ideally, we would want to test a system that produces for a given meaning to be conveyed the various possible sentences ordered in terms of their rankings, from which the top-ranked sentence is taken. In the absence of a semantic component in our model, we could only test the already given alternative sentences and assess whether our model could predict the correct one.

Despite these problems, our main result is that with just a tiny fraction of a child's input the correct sentence can be predicted by an unsupervised TSG for virtually all phenomena related to wh-questions as well as for a number of other phenomena that even fall out of the scope of the traditional generativist account.

Finally it should be noted that our result is not in contrast with all generativist work. For example, in Hauser et al. (2002), it was proposed that the core language faculty comprises just recursive tree structure and nothing else. The work presented in this paper may be the first to show that one general grammar induction algorithm makes language learning possible for a much wider set of phenomena than has previously been endeavored.

If empiricist models want to compete with generativist models, they should compete in the same arena, with the same phenomena.

References

D. Adger, 2003. *Core syntax: A minimalist approach*. Oxford University Press, 2003.

B. Ambridge and E. Lieven, 2011). *Child Language Acquisition. Contrasting Theoretical Approaches*. Cambridge University Press.

M. Bansal and D. Klein, 2011. The Surprising Variance in Shortest-Derivation Parsing, *Proceedings ACL-HLT 2011*.

C. Bannard, E. Lieven and M. Tomasello, 2009. Modeling Children's Early Grammatical Knowledge, *Proceedings of the National Academy of Sciences*, 106, 17284-89.

R. Bod, R. Scha and K. Sima'an (eds.), 2003. *Data-Oriented Parsing*, CSLI Publications/University of Chicago Press.

R. Bod, 2006. An all-subtrees approach to unsupervised parsing. *Proceedings ACL-COLING*.

R. Bod, 2009. From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. *Cognitive Science*, 33(5), 752-793.

R. Borsley, 2004. *Syntactic Theory: A Unified Approach*, Oxford University Press.

A. Clark and R. Eyraud, 2006. Learning Auxiliary Fronting with Grammatical Inference. *Proceedings CONLL 2006.*

A. Clark and S. Lappin, 2011. *Linguistic Nativism and the Poverty of the Stimulus*, Wiley-Blackwell.

T. Cohn, P. Blunsom, and S. Goldwater, 2010. Inducing Tree-Substitution Grammars, *Journal of Machine Learning Research*, JMLR 11, 3053-3096.

S. Crain, 1991. Language acquisition in the absence of experience. *Behavorial and Brain Sciences*, 14, 597-612.

S. Crain and R. Thornton. Acquisition of syntax and semantics, 2006. In M. Traxler and M. Gernsbacher, editors, *Handbook of Psycholinguistics*. Elsevier.

S. Foraker, T. Regier, N. Khetarpal, A. Perfors, and J. Tenenbaum, 2009. Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One. *Cognitive Science*, 33, 287-300.

J. Goodman, 2003. Efficient parsing of DOP with PCFG-reductions. In R. Bod, R. Scha & K. Sima'an (Eds.), *Data-oriented parsing*, 125–146. CSLI Pubs.

M. Hauser, N. Chomsky and T. Fitch, 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298, 1569–1579.

L . Huang and D. Chiang, 2005. Better k-best parsing. In *Proceedings IWPT 2005*, pp. 53–64.

X. Kam, L. Stoyneshka, L. Tornyova, J. Fodor and W. Sakas, 2008. Bigrams and the Richness of the Stimulus. *Cognitive Science*, *32*, 771-787.

D. Klein and C. Manning, 2005 Natural language grammar induction with a generative constituent-context model. *Pattern Recognition*, 38, 1407–1419.

J. Lidz, S. Waxman and J. Freedman, 2003. What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, 89, B65–B73

B. MacWhinney, 2000. *The CHILDES project: Tools for analyzing talk*. Mawah, NJ: Erlbaum

B. MacWhinney, 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 3, 883- 914.

T. O'Donnell, N. Goodman, and J. Tenenbaum, 2009. Fragment grammar: Exploring reuse in hierarchical generative processes. *Technical Report MIT-CSAIL-TR-2009-013*, MIT.

M. Post and D. Gildea, 2009. Bayesian learning of a tree substitution grammar. In *Proceedings of the ACL-IJCNLP 2009.*

G. Pullum and B. Scholz, 2002. Empirical assessment of stimulus povery arguments. *The Linguist Review, 19(2002)*, 9-50.

F. Reali and M. Christiansen, 2005. Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007-1028.

J. Ross, 1967. *Constraints on variables in syntax*. PhD thesis, Massachusetts Institute of Technology.

C. Teichmann, 2011. Reducing the size of the representation for the uDOP-estimate. *Proceedings EMNLP 2011*.

H. Waterfall, B. Sandbank, L. Onnis, and S. Edelman, 2010. An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language*, 37, 671-703.

A. Zollmann and K. Sima'an. 2005. A Consistent and Efficient Estimator for Data-Oriented Parsing. In *Journal of Automata, Languages and Combinatorics*, 10 (2005), 367-388.