# Modeling German Word Order Acquisition via Bayesian Inference

**Annika Heuser**
Massachusetts Institute of Technology
Departments of Brain & Cognitive Sciences
and Electrical Engineering & Computer Science
aheuser@mit.edu

**Polina Tsvilodub**
Osnabrück University
Institute of Cognitive Science
ptsvilodub@uos.de

## 1 Introduction

The question of how children acquire the grammatical principles of their native language has been debated in cognitive science and linguistics for over 50 years (e.g., Chomsky, 1965; Xu and Tenenbaum, 2007). Perfors et al. (2011) suggest that children compare various *syntactic hypotheses* in order to determine which is most compatible with the so-called primary linguistic data (cf. Chomsky, 1965; Wexler and Culicover, 1980; Clark and Lappin, 2013; among many others), and that a Bayesian model selection approach might be able to do some of this work. More specifically, they demonstrate that a Bayesian inference system presented with child-directed input would prefer a probabilistic context free grammar (PCFG) that used hierarchical phrase structure over a regular grammar representing linear phrase structure, without any initial prior bias towards one grammar type or another. Therefore, they argued that correct linguistic generalizations can be achieved by a learner equipped with domain-general Bayesian inference capacities but without language-specific innate knowledge.

In their work, the various syntactic hypotheses were represented by grammars that were all able to parse the entirety of a pre-processed English child-directed speech corpus. However, a computational model without the ability to compare competing hypotheses that are supported by different subsets of the data is severely limited due to the inevitability of noise. Children are also very likely to hear errors, actual or only perceived (Krentz and Corina, 2008; Friederici et al., 2011). Furthermore, hypotheses concerning typological tendencies might not warrant comparison over fully congruent data sets. The premise of this work is therefore to extend the Bayesian system designed by Perfors et al. (2011) to allow for comparing models supported by different data subsets. We evaluated this system in the context of word order acquisition in German.

To this end, we designed PCFGs representing different word order hypotheses a child might entertain. Many linguists agree that German word order is Subject-Object-Verb (SOV) Verb-Second (V2) (henceforth: SOV+V2) (e.g., Bierwisch, 1963). Yet many simple German sentences also maintain Subject-Verb-Object (SVO) order. Because of the SVO word order of many simple German sentences, a child might initially consider a left-branching grammar only able to account for SVO sentences, only to reject it in order to account for embedded sentences and sentences with auxiliaries.

Note that we do not commit to a specific theoretical position on the plausibility of humans performing verb movement operations, and merely lean on this theory to specify conceivable word order hypotheses, assuming that a PCFG representing SVO word order should be able to parse a decent number of German sentences, but fewer than the SOV+V2 PCFG. Additionally, our PCFGs are comparable in complexity in contrast to complexities differing by orders of magnitude for the grammars used by Perfors et al. (2011). Our PCFGs can thus serve as snapshots of different hypotheses comparable in *a priori* complexity that a child may at some point compare in order to determine the word order of their language, though they by no means exhaust the space of hypotheses that a child may consider.

## 2 Methods

We hand-designed four CFGs to represent potential competing word order hypotheses a child could consider: SOV, SVO, SOV+V2, and SVO+V2. The CFGs were all converted to PCFGs via the Inside-Outside algorithm[1] (Baker, 1979) over the sentences from the German Leo corpus of the

[1]We gratefully acknowledge Tristan Thrush for providing his implementation of the algorithm in Python.



*Proceedings of the Society for Computation in Linguistics (SCiL) 2021*, pages 420-424.
Held on-line February 14-19, 2021

CHILDES database that each CFG could parse (MacWhinney, 2000; Behrens, 2006). The crucial scoring criterion to compare the different types of grammars was the Bayes score following Perfors et al. (2011) which can be considered by a learner equipped with Bayesian inference capacity:

$$P(G,T|D) \propto P(D|G,T) \cdot P(G|T) \cdot P(T) \quad (1)$$

where G = grammar, T = grammar type—i.e., SOV, SVO, etc.—and D = data. So far, the only difference between our experimental setup and that of Perfors et al. (2011) was that the PCFGs were not trained over the same data. Therefore, the likelihoods, $P(D|G,T)$, of the four grammars are not directly comparable, especially because they are each calculated over a different number of sentences. Thus, the more sentences over which the likelihood is calculated, the smaller it will be, because more probabilities are multiplied together for the likelihood calculation. We determined how to reconcile this fact in order to effectively compare the four grammars.

### 2.1 Data Preparation

Our data consisted of the 257,644 utterances produced by adults in the Leo corpus (Behrens, 2006). We ran the Stanford NLP tagger using the Stuttgart-Tübingen (STTS) tag set on these sentences (Toutanova et al. (2003), see the appendices of Smith (2003) for a list of the tags). Upon closer inspection of the data, we noticed that there were more tagging errors and ungrammatical sentences than we expected. Therefore, we built a pipeline for data preparation. Overall, the pipeline removed filler words that led to many tagging errors, as well as very short sentence fragments unlikely to contribute to learning syntactic information for the task at hand. We excluded 97,009 sentences altogether. Importantly, the data preparation pipeline simplified the task of designing the grammars and reduced the number of ungrammatical sentences from the Leo corpus that they were evaluated on.

### 2.2 Grammar Design

The SOV+V2 grammar was hand-designed as a representation of standard German. We stress that advancing beyond the Perfors et al. (2011) grammars, our grammar was able to parse more complicated syntactic phenomena, such as questions and negations. The terminals of the SOV+V2 CFG consisted of the POS tags of the processed data. Although the SOV+V2 CFG was designed to parse German, it was still only able to parse 48.98% of all POS-tagged sentences in the corpus and only 21.29% of unique POS strings. To determine the extent to which residual errors left in the corpus were to blame for this, as opposed to the shortcomings of the CFG, we randomly sampled 100 POS strings and their corresponding sentences that the SOV+V2 CFG could parse and 100 that it could not. We determined how many of these sentences were grammatical, how many were tagged correctly, and how many of the POS strings could correspond to grammatical German sentences (s. Section 3).

The other three CFGs were derived from the SOV+V2 CFG, such that about 80% of the CFGs were unchanged across the grammar types. As expected, the SOV+V2 CFG parsed the greatest percentage of POS strings from the processed Leo corpus (s. Figure 1). The relative ranking in parsing ability did not change over only the unique POS strings of our processed data.

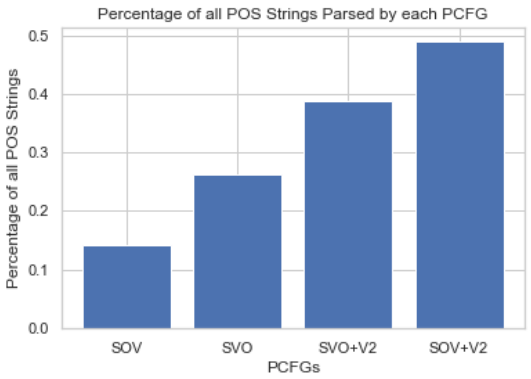


Figure 1: The percentage of POS strings corresponding to sentences from the processed Leo corpus that each PCFG can parse.

### 2.3 Calculating the Likelihood and the Prior

To compute the posterior, we first calculated the likelihood of the data being generated by the individual grammars. We trained a PCFG from each CFG using *all* the POS strings that the CFG could parse from the processed Leo corpus as the training data. Next, we determined the overall likelihood of the data being generated from the particular grammar $G$ of type $T$ by calculating the product of the likelihoods of the $n$ sentences in the data. The likelihood of a particular sentence is determined by summing over the probabilities of all possible parses for a given sentence $s_i$. The greater the value

of $n$, the smaller $P(D|G,T)$ will be, so grammars are penalized for parsing more sentences. Therefore, we determined the mean $P(s_i)$ of each PCFG as an average estimator for the grammar's tightness-of-fit per sentence. Nonetheless, the ranking of the PCFGs' mean $P(s_i)$ was the exact inverse of the ranking in Figure 1. We identified three characteristics of the PCFGs able to parse more sentences that could account for this: (1) greater average sentence length, (2) greater average number of parses per sentence, (3) greater generalizability over different sentence types.

To confirm the role of (1) in a greater likelihood of the data, we compared the mean sentence log-likelihood across PCFGs against the sentence length. We suspected that grammars unable to parse many sentences, such as the SOV PCFG, were only able to parse short sentences, which have greater likelihoods because they require the application of fewer production rules. The correlation coefficient for the PCFGs' mean sentence log likelihood and mean sentence length was indeed -0.9971($p = 0.0029$). We therefore normalized each sentence's log likelihood by the sentence length in order to combat the penalty that the PCFGs able to parse longer sentences incurred.

In addition, we suspected that longer sentences are more likely to be ambiguous, and therefore have more than one parse. We confirmed that this was the case: the SOV PCFG, which had the highest likelihood, had the lowest average number of parses per sentence (1.33 parses/sentence vs 1.59 for SVO, 1.50 for SVO+V2, and 1.45 for SOV+V2). Therefore, (2) might be partially responsible for PCFGs able to parse more sentences having a lower likelihood; however, it is clearly not as predictive of the PCFGs' data likelihoods as mean sentence length. For this reason, we did not correct for the potential effect of the number of parses.

Lastly, the ranking of PCFG parsing ability presented in Figure 1 was the same over just the unique POS strings of our data, which we deem to reasonably approximate a grammar's generalization to distinct grammatical phenomena. Yet, the more sentence types a grammar can parse, the lower is the probability of a single sentence parse (Perfors et al., 2011). Therefore, we inferred that PCFGs able to parse more sentences were penalized for their ability to generalize to a greater number of sentence types. To counter this, we weighted the mean sentence log likelihood, normalized by sentence length, by the percentage of unique sentence types that the PCFG can parse.

The second component of the posterior of each grammar is the prior probability of the particular grammar given a specific type, $P(G|T)$ (s. Eq. 1). The prior of each grammar was calculated by following the generative process underlying the selection of that grammar from the space of grammars of that type, formalized in terms of a meta-grammar, generally preferring simpler grammars which require making less design decisions (cf. Perfors et al., 2011; Feldman et al., 1969). Refer to Perfors et al. (2011) for details on the implementation. Crucially, the grammar types are all equally likely, so the uniform $P(T)$ component of the prior can be dropped from the proportion.

## 3 Results

### 3.1 Corpus Analysis

First, we assessed the critical assumption that the system's input resembled that of a child through our statistical analysis of the SOV+V2 CFG's ability to distinguish grammatical and ungrammatical German POS strings. We used this analysis to determine the approximate percentage of errors in the processed Leo corpus.

We found that only 11% of the sampled POS strings that the SOV+V2 CFG could not parse were actually representative of grammatical German sentences. Similarly, only 12% of the sentences that the CFG could parse should not have been parseable because they were not representative of grammatical sentences. From the sample error proportions, we calculated the error proportion upper bound for the POS strings of the processed data that the SOV+V2 CFG could and could not parse to be 0.20 and 0.19 respectively ($p < 0.05$). These upper bounds are particularly noteworthy given that the SOV+V2 CFG can only parse 21.29% of the unique POS strings in the processed corpus. By calculating the total number of ungrammatical POS strings in the entire processed Leo corpus, we conclude that errors in the data (tagging or otherwise) were responsible for the majority of unique POS stings that could not be parsed. The result of multiplying our upper-bounds and their complements by the number of unique POS strings that the SOV+V2 CFG could and could not parse is shown in Table 1. By summing over Table 1's rightmost column, we approximated the total number of unique ungrammatical POS strings in the processed

Table 1: Proportion of Grammatical/Ungrammatical POS Strings in the Processed Leo Corpus

| | Grammatical POS String | Ungrammatical POS String |
|---|---|---|
| Parsed by SOV+V2 CFG | 0.80(15,916) = 12,732 | 0.20(15,916) = 3,183 |
| Not Parsed by SOV+V2 CFG | 0.19(58,842) = 11,180 | 0.81(58,842) = 47,662 |

Leo corpus to be 50,845, or 68% of the total number of unique POS strings. With the assumption that each unique ungrammatical POS string only occurs once in the processed Leo corpus, we conservatively estimated that 31.65% (50,854/160,635 total POS strings) of the entire corpus are ungrammatical POS strings. Therefore, we argue that this data might not accurately represent a child's linguistic input because we deem adults very unlikely to produce ungrammatical speech over 30% of the time in their native language. Although our subjective grammaticality judgments may contribute to a somewhat inaccurate estimation of this error proportion, this is likely counteracted by the conservative assumption that unique ungrammatical POS strings only occur once in the entire processed corpus.

### 3.2 Posterior Probabilities of the Grammars

The different grammars were compared using their posterior probabilities given the processed data they could parse, computed via Bayes' rule (s. Eq. 1). The prior probabilities of each of the grammars are reported in Table 2. Rules that were never used during training were discarded so as to compute the prior probability only over the relevant rules. After applying the suggested normalization steps (s. Section 2.3), we multiplied the result by 22,657, the number of sentences parsed by the SOV PCFG, which parsed the smallest subset of the data. We decided to multiply by the minimum number of sentences parsed in order to estimate a lower bound of the likelihood's weight on the posteriors of the grammars compared to the priors. We confirmed the stability of the qualitative results reported: With 95% confidence, the log likelihood ranking of the PCFGs will hold irrespective of the initialization.[2] We realize that the PCFG preference based on the posterior probabilities shown in Table 2 could be achieved by simply picking the PCFG able to parse the greatest percentage of the corpus. However, in different circumstances, perhaps given less error-ridden data, or comparing grammars than can parse approximately the same percentage of a corpus, this may not be the case.

[2]Only two initializations were performed for computational tractability reasons.

Table 2: All in terms of log. From left to right: posterior of grammar $G$ and type $T$, likelihood of data $D$, and prior of $G$ given $T$.

| G | $P(G,T\|D)$ | $P(D\|G,T)$ | $P(G\|T)$ |
|---|---|---|---|
| SOV | -851,854.0 | -849,792.2 | -2,061.8 |
| SVO | -418,402.1 | -416,430.2 | -1,971.9 |
| SVO+V2 | -322,738.5 | -320,747.9 | -1,990.7 |
| SOV+V2 | -223,738.5 | -221,574.6 | -2,163.9 |

## 4 Discussion

With the additional calculations called for by our extended Bayesian model selection scheme to counteract the three compounding effects of parsing a larger subset of the data that we identified, our system prefers the expected SOV+V2 grammar. We recommend that our result be verified with a demonstrably less error-ridden German child-directed speech corpus. This error proportion may be an artifact of the specific corpus we chose, or the result of POS taggers performing worse on languages other than English; however, this demonstrates that corpus quality and POS-tagging accuracy should not be taken for granted. The results of our scheme should also be compared with the preference of strict Bayesian inference over smoothed versions of the same PCFGs augumented to parse the entire dataset; though these require greater computational resources to train.

In sum, we propose a Bayesian model selection scheme as a more flexible model of child language acquisition able to compare hypotheses compatible with different subsets of a child-directed speech corpus. We envision our flexible model selection scheme supporting bilingual language acquisition modeling in future research. We leave the verification of the assumptions that children can represent such grammars and employ Bayesian inference to future research. Nonetheless, our scheme is especially useful when the training data is relatively sparse or contains errors that a grammar should not be able to parse, but would still cause the grammar to be invalidated under strict Bayesian inference.

## References


James K Baker. 1979. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132.

Heike Behrens. 2006. The input–output relationship in first language acquisition. *Language and cognitive processes*, 21(1-3):2–24.

Manfred Bierwisch. 1963. *Grammatik des deutschen Verbs*. Akademie Verlag.

Noam Chomsky. 1965. Aspects of the theory of syntax. *Cambridge, MA: MIT Press*, (1977):71–132.

Alexander Clark and Shalom Lappin. 2013. Complexity in language acquisition. *Topics in cognitive science*, 5(1):89–110.

Jerome A Feldman, James Gips, James J Horning, and Stephen Reder. 1969. Grammatical complexity and inference. Technical report, Stanford Uiversity, CA, Dept. of Computer Science.

Angela D Friederici, Jutta L Mueller, and Regine Oberecker. 2011. Precursors to natural grammar learning: preliminary evidence from 4-month-old infants. *PLoS One*, 6(3):e17920.

Ursula C Krentz and David P Corina. 2008. Preference for language in early infancy: the human language bias is not speech specific. *Developmental Science*, 11(1):1–9.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

George Smith. 2003. A brief introduction to the tiger treebank. *Ms. Universität Potsdam*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, pages 173–180. Association for Computational Linguistics.

Kenneth Wexler and Peter W. Culicover. 1980. *Formal principles of language acquisition*. MIT Press (MA).

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.