Character-based PCFG Induction for Modeling the Syntactic Acquisition of Morphologically Rich Languages

Lifeng Jin  
Tencent AI Lab  
lifengjin@tencent.com

Byung-Doh Oh  
Department of Linguistics  
The Ohio State University  
oh.531@osu.edu

William Schuler  
Department of Linguistics  
The Ohio State University  
schuler.77@osu.edu

Abstract

Unsupervised PCFG induction models, which build syntactic structures from raw text, can be used to evaluate the extent to which syntactic knowledge can be acquired from distributional information alone. However, many state-of-the-art PCFG induction models are word-based, meaning that they cannot directly inspect functional affixes, which may provide crucial information for syntactic acquisition in child learners. This work first introduces a neural PCFG induction model that allows a clean ablation of the influence of subword information in grammar induction. Experiments on child-directed speech demonstrate first that the incorporation of subword information results in more accurate grammars with categories that word-based induction models have difficulty finding, and second that this effect is amplified in morphologically richer languages that rely on functional affixes from a very young age (Mintz, 2013; Haryu and Kajikawa, 2016), which may provide crucial information for syntactic acquisition (Dye et al., 2019). This would presumably make word-based PCFG induction models less appropriate for modeling the acquisition of morphologically rich languages, in which most information about syntactic units and relations is expressed at the subword level (Tsarfaty et al., 2010).

In order to address this issue, this work first defines a character-based model and a minimally-manipulated word-based counterpart for neural PCFG induction. This formulation allows a clean ablation of subword information, and therefore allows its influence on grammar induction to be studied. Experiments using child-directed speech demonstrate that the incorporation of subword information results in more accurate grammars with coherent syntactic categories that the word-based induction model has difficulty finding. Additionally, this effect is found to be amplified in morphologically richer languages that rely on functional affixes to express grammatical relations. A subsequent evaluation on multilingual treebanks shows that the model with subword information achieves state-of-the-art results on many languages, further supporting a distributional model of syntactic acquisition.

1 Introduction

Unsupervised PCFG induction models (Johnson et al., 2007; Jin et al., 2018b) induce grammars from raw text and use those induced grammars to build linguistically meaningful hierarchical structures for sentences. Recent work on PCFG induction has adopted Bayesian or neural word-based PCFG models (Jin et al., 2018b; Kim et al., 2019; Zhu et al., 2020), which have proven to be accurate at discovering syntactic structures solely from word sequences. These results indicate that a human-like grammar can be learned from distributional data (Harris, 1954; Saffran et al., 1996; Aslin and Newport, 2014), providing some evidence against the poverty of the stimulus argument (Chomsky, 1965, 1980) in language acquisition.

However, despite their high performance, these word-based PCFG induction models are not fully suitable as computational-level models of syntactic acquisition. Specifically, they treat words as symbols and do not have direct access to subword information. In contrast, child language learners are known to be sensitive to word-internal functional affixes from a very young age (Mintz, 2013; Haryu and Kajikawa, 2016), which may provide crucial information for syntactic acquisition (Dye et al., 2019). This would presumably make word-based PCFG induction models less appropriate for modeling the acquisition of morphologically rich languages, in which most information about syntactic units and relations is expressed at the subword level (Tsarfaty et al., 2010).

In order to address this issue, this work first defines a character-based model and a minimally-manipulated word-based counterpart for neural PCFG induction. This formulation allows a clean ablation of subword information, and therefore allows its influence on grammar induction to be studied. Experiments using child-directed speech demonstrate that the incorporation of subword information results in more accurate grammars with coherent syntactic categories that the word-based induction model has difficulty finding. Additionally, this effect is found to be amplified in morphologically richer languages that rely on functional affixes to express grammatical relations. Finally, an evaluation on multilingual treebanks shows that the model with subword information achieves state-of-the-art induction results on many languages, providing further evidence for a distributional model of syntactic acquisition.

2 Related Work

Early work in unsupervised PCFG induction from raw text (Johnson et al., 2007; Liang et al., 2009;  

1Code used in this work is available at https://github.com/lifengjin/charInduction.
Tu, 2012) was not as successful as models of unsupervised constituency parsing (Segier, 2007; Pont-vert et al., 2011). However, recent work from unsupervised parsing (Shen et al., 2019; Wang et al., 2019; Drozdov et al., 2019, 2020) and grammar induction (Jin et al., 2018a, 2019; Kim et al., 2019; Zhu et al., 2020; Jin and Schuler, 2020; Li et al., 2020) shows much improvement over previous results with grammars learned solely from raw text, indicating that statistical regularities relevant to syntactic acquisition can be found in word collocations. For example, Kim et al. (2019) propose a word-based neural PCFG induction model for accurate grammar induction on English. Zhu et al. (2020) further extend this compound PCFG induction model to jointly induce lexical dependencies using a lexicalized PCFG. Jin et al. (2019) augment a PCFG induction model to use contextualized word embeddings with subword information to allow morphological cues to influence induction. Unfortunately, the ELMo embeddings (Peters et al., 2018) used in that work are trained with a large number of tokens, which limits the value of this approach for investigating child-like syntactic acquisition.

In an attempt to answer a similar question, there has also been recent interest in evaluating whether recurrent neural networks like LSTMs can learn grammar-like representations from word sequence alone. Although some initial results seemed promising (Linzen et al., 2016; Gulordava et al., 2018), later studies have shown that some of these results may not necessarily generalize to languages other than English (Ravfogel et al., 2018; Davis and van Schijndel, 2020), yielding an inconsistent picture. Moreover, studies in this line of research try to model specific syntactic phenomena such as subject-verb agreement, filler-gap dependencies, and auxiliary inversion, and therefore have a slightly different focus from grammar induction. Complementary to this approach are studies that aim to reconstruct syntactic representations from contextualized word representations (Tenney et al., 2019; Hewitt and Manning, 2019). Again, however, the use of neural language models assumes access to much more data than the input available to the typical child (Hart and Risley, 1995).

3 Models

Experiments described in this paper use two variants of a neural PCFG induction model, which differ minimally in how terminal expansion is modeled. A Chomsky normal form PCFG with $C$ non-terminal categories is first factored into two separate parts: binary-branching nonterminal expansion rule probabilities, and unary-branching terminal expansion rule probabilities (Jin et al., 2019). Given a tree as a set $\tau$ of nodes $\eta$ undergoing nonterminal expansions $c_{\eta} \rightarrow c_{\eta 1} c_{\eta 2}$ (where $\eta \in \{1,2\}$ is a Gorn address specifying a path of left or right branches from the root), and a set $\tau'$ of nodes $\eta$ undergoing terminal expansions $c_{\eta} \rightarrow w_{\eta}$ (where $w_{\eta}$ is the word at node $\eta$), the marginal probability of a sentence $\sigma$ can be computed as:

$$P(\sigma) = \prod_{\tau,\tau' \in } P(c_{\eta} \rightarrow c_{\eta 1} c_{\eta 2}) \cdot \prod_{\eta \in \tau'} P(c_{\eta} \rightarrow w_{\eta}),$$

which is also the objective function to maximize for all proposed models in this work.

To allow the terminal expansion model to be separated out from the rest of the grammar induction model and make the character- and word-based models differ only by the terminal expansion model, we first define a set of Bernoulli distributions that distribute probability mass between these two sets of rules:

$$P(\text{Term} | c_{\eta}) = \text{softmax}(N(E \delta_{c_{\eta}})), \quad (2)$$

where $c_{\eta}$ is a nonterminal category, $\delta_{c_{\eta}}$ is a Kronecker delta function – a vector with value one at index $c_{\eta}$ and zeros everywhere else – and $E \delta_{c_{\eta}}$ is a category vector input for the Bernoulli distribution of $c_{\eta}$ with $E \in \mathbb{R}^{d \times C}$, a matrix of nonterminal category embeddings of size $d$. $N$ is an arbitrary neural network, which in our implementation is a multilayered residual network (Kim et al., 2019). The residual network consists of $B$ architecturally identical residual blocks. For an input vector $x_{b-1,c_{\eta}}$, each residual block $b$ performs the following computation:

$$x_{b,c_{\eta}} = \text{ReLU}(W_b x_{b-1,c_{\eta}} + b_b) + b'_{b} + x_{b-1,c_{\eta}}. \quad (3)$$

There are two fully connected layers before and after the residual blocks:

$$x_{0,c_{\eta}} = \text{ReLU}(W_0 E \delta_{c_{\eta}} + b_0), \quad (4)$$

$$s_{c_{\eta}} = \text{ReLU}(W_{\text{soft}} x_{B,c_{\eta}} + b_{\text{soft}}). \quad (5)$$

These rules include the expansion rules generating the top node in the tree.
All $\mathbf{W}$’s and $\mathbf{b}$’s are model parameters, and $s_{c_n}$ are the logits for the final softmax in Equation 2. $B$ is set to 2 in all models (Kim et al., 2019). This formulation naturally allows the model to learn from the input data how to allocate preterminal and other nonterminal categories, thereby making it more appropriate for multilingual settings (e.g. agglutinative languages that have many word types may need more preterminal categories in comparison to other nonterminal categories).

Binary-branching nonterminal expansion rule probabilities for a nonterminal category $c_\eta$ are defined as:

$$
P(c_\eta \rightarrow c_{\eta 1} c_{\eta 2}) = P(\text{Term}=0 \mid c_\eta),$$
$$P(c_\eta \rightarrow c_{\eta 1} c_{\eta 2} \mid c_\eta, \text{Term}=0). \tag{6}$$

The binary-branching nonterminal expansion distribution is defined for all models as:

$$
P(c_\eta \rightarrow c_{\eta 1} c_{\eta 2} \mid c_\eta, \text{Term}=0) = \text{softmax}(W_{\text{non}} \mathbf{E} \delta_{c_\eta} + b_{\text{non}}), \tag{7}$$

where $W_{\text{non}} \in \mathbb{R}^{C \times d}$ and $b_{\text{non}} \in \mathbb{R}^d$ are parameters of the model, and $\mathbf{E}$ is the embedding matrix for all hypothesized nonterminal categories.

The lexical unary-expansion rule probabilities for a preterminal category $c_\eta$ are defined as:

$$
P(c_\eta \rightarrow w_\eta) = P(\text{Term}=1 \mid c_\eta),$$
$$P(c_\eta \rightarrow w_\eta \mid c_\eta, \text{Term}=1), \tag{8}$$

where $w_\eta$ is the generated word token at node $\eta$. The representation of the word $w_\eta$ is usually symbolic in PCFG induction models. In order to probe the effect of character-based models in incorporating subword information to grammar induction, models in this work differ in how words are represented and in turn how the terminal expansion models are defined. The word-based models (NeuralWord) use symbolic representation of words, whereas the character-based models (NeuralChar) use character sequences as observations for the terminal expansion models. They differ minimally by the terminal expansion models and the input they take, which allows a clean manipulation of subword information while holding all other model components fixed.

**Lexical Expansion Models**

The word-based lexical expansion model is a multilayered residual network which takes as input a category embedding and generates a distribution score for all words in the vocabulary:

$$
P(c_\eta \rightarrow w_\eta \mid c_\eta, \text{Term}=1) = \text{softmax}(N'(\mathbf{E} \delta_{c_\eta})). \tag{9}$$

The function $N'$ is a four-layer residual neural network with two residual blocks similar to the one used in generating nonterminal expansion probabilities, except that the output dimension of this network is the size of the vocabulary.

For the character-based induction model, the terminal expansion probability is factored into the product of a locally-normalized sequence model generating all the characters in the word from left to right:

$$
P(c_\eta \rightarrow w_\eta \mid c_\eta, \text{Term}=1) = \prod_{l \in \{l_1, \ldots, l_n\}} P(l_i \mid c_\eta, l_1, \ldots, l_{i-1}), \tag{10}$$

where the character or letter sequence $l_1, \ldots, l_n$ comprises the word $w_\eta$ with $L$ as the character vocabulary. The sequence model is a multilayered long short-term memory network (LSTM) with $B$ layers. The character generation distribution from the LSTM for each letter $l_i$ is:

$$
P(l_i \mid c_\eta, l_1, \ldots, l_{i-1}) = \text{softmax}(W_{\text{char}} \mathbf{h}_{i,b,c_\eta} + b_{\text{char}}), \tag{11}$$

where $W_{\text{char}} \in \mathbb{R}^{[L] \times h}$, $b_{\text{char}} \in \mathbb{R}^{[L]}$, and $h$ is the size of the hidden and cell states. The hidden and cell states of the LSTM are calculated by:

$$
\mathbf{h}_{i,b,c_\eta}, \mathbf{c}_{i,b,c_\eta} = \text{LSTM}(\mathbf{h}_{i,b-1,c_\eta}, \mathbf{h}_{i-1,b,c_\eta}, \mathbf{c}_{i-1,b,c_\eta}), \tag{12}
$$

where $b$ is the current layer index. $\mathbf{h}_{i,b-1,c_\eta}$ is the input from the LSTM one layer below, and $\mathbf{h}_{i-1,b,c_\eta}$ and $\mathbf{c}_{i-1,b,c_\eta}$ are the input from the previous time step. $\mathbf{h}_{i,b,c_\eta}$ and $\mathbf{c}_{i,b,c_\eta}$ are the current cell and hidden states of the LSTM. Finally, the initial hidden state of the $b^{th}$ layer of the LSTM depends on the nonterminal category $c_\eta$:

$$
\mathbf{h}_{0,b,c_\eta} = \text{ReLU}(W_{b,\text{term}} \mathbf{E} \delta_{c_\eta} + b_{b,\text{term}}), \tag{13}
$$

where $W_{b,\text{term}} \in \mathbb{R}^{h \times d}$ and $b_{b,\text{term}} \in \mathbb{R}^{h}$ are trainable parameters.

4 Experiment 1: Evaluation on Child-directed Speech

This work first explores the influence of subword information in grammar induction by comparing the
performance of the NeuralChar and NeuralWord models on child-directed speech corpora, which represent authentic linguistic input that child learners are exposed to.

4.1 Evaluation Metric: Recall-Homogeneity

It has been argued that the ability to predict constituent labels that are consistent with linguistic annotation is a crucial part of a grammar and therefore should be accounted for in evaluation (Jin et al., 2019). This work uses Recall-Homogeneity (RH, Jin et al., 2021) as a labeled evaluation metric, which is calculated by multiplying unlabeled recall of bracketed spans in the predicted trees with the homogeneity score (Rosenberg and Hirschberg, 2007) of the predicted labels of the matching spans. Similarly to Recall-V-Measure (RVM, Jin et al., 2019), this metric is made insensitive to the branching factor of the grammar through the use of unlabeled recall. However, the use of homogeneity rather than V-measure assumes that the annotators’ decision to suppress the annotation of in-depth information such as case or subcategorization in category labels is motivated by expediency rather than linguistic theory. Therefore, RH does not penalize induced grammars for the use of categories to make more fine-grained distinctions, to the extent that it does not interfere with the homogeneity of predictions on other attested categories. In this work, unary branches in both gold and predicted trees are removed and only the top category of any unary chain is used for evaluation.3

4.2 Procedures

The NeuralChar and NeuralWord models were evaluated on English and Korean child-directed speech corpora from CHILDES (MacWhinney, 2000). The English corpus comes from the Eve section (Brown, 1973), which contains transcriptions of interactions between Eve and her caregivers at ages from 1 year 6 months to 2 years 3 months. Only the sentences that were uttered by caregivers were kept in the data, which resulted in a set of 14,251 sentences with a mean sentence length of 5.6 words. Penn Treebank-style syntactic annotations for these child-directed utterances are provided by Pearl and Sprouse (2013). The Korean corpus is from the Jong section (Ryu et al., 2015), which contains transcriptions of interactions between Jong and his caregivers recorded at ages from 1 year 3 months to 3 years 5 months. There were 28,620 sentences that were uttered by caregivers, which had a mean sentence length of 5.0 words. As there are no gold syntactic annotations available for this dataset, a subset of 150 sentences was annotated in order to evaluate the two models. This was done by first automatically generating silver parses using the state-of-the-art supervised parser (Kitaev et al., 2019) and subsequently correcting them according to the annotation scheme of Choi (2013).4

The NeuralChar and NeuralWord models were trained on these two corpora ten times with different random seeds, using 90 nonterminal categories and other hyperparameters tuned on the Brown Corpus portion of the Penn Treebank (see Appendix A).5 Following previous work (Seginer, 2007), punctuation marks were left in the input data

---

3Homogeneity is positively correlated with the number of categories used, achieving a perfect score when each constituent is assigned a unique category. However, a grammar with tens of thousands of categories, or even a few hundred categories, is beyond the capacity of current grammar induction models.

4The annotated Korean data is available at https://github.com/lifengjin/charInduction.

5The models were trained on the full set for both corpora. However, while the models were evaluated on the full set for the Eve section, they were evaluated only on the subset of 150 sentences with manually-corrected syntactic annotations for the Jong section.
Table 1: Gold part-of-speech tags and representative characterizations of the eight most frequent preterminal categories induced by the NeuralChar model (NC, top) and the NeuralWord model (NW, bottom). The gold tags that account for more than 5% of each induced category are reported. The gold tags consist of morphological tags that are delimited by plus symbols. Refer to Appendix B for a brief description of the morphological tags.

<table>
<thead>
<tr>
<th>Induced category</th>
<th>Count</th>
<th>Attested category (relative frequency)</th>
<th>Representative characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC-63</td>
<td>100</td>
<td>sf (1.0)</td>
<td>Full stop</td>
</tr>
<tr>
<td>NC-29</td>
<td>73</td>
<td>npd+jxt (0.23), nq (0.12), ncn (0.12), npd (0.1), nq+jcs (0.07), ncn+jvcs (0.05)</td>
<td>Nouns for child (“Jong”), demonstrative pronoun (“this”)</td>
</tr>
<tr>
<td>NC-62</td>
<td>48</td>
<td>sf (1.0)</td>
<td>Question mark</td>
</tr>
<tr>
<td>NC-38</td>
<td>25</td>
<td>pxvref (0.32), pvgsref (0.2), paavref (0.2), pvgsrepref (0.16)</td>
<td>Sentence-final verbs</td>
</tr>
<tr>
<td>NC-16</td>
<td>21</td>
<td>pvgsreccc (0.67), pvgsreccs (0.14), paavreccs (0.1), paavref (0.1)</td>
<td>Verbs preceding auxiliary verbs</td>
</tr>
<tr>
<td>NC-20</td>
<td>20</td>
<td>ncn (0.55), ncn+jvcs (0.15), ncn+jvcs (0.1), padvref (0.05), mag (0.05), ncn+jvcs (0.05), pvdrsvcs (0.05)</td>
<td>Nouns for caretaker (“mom”)</td>
</tr>
<tr>
<td>NC-6</td>
<td>20</td>
<td>ii (1.0)</td>
<td>Interjections</td>
</tr>
<tr>
<td>NC-7</td>
<td>20</td>
<td>padvref (1.0)</td>
<td>Expression of agreement (“That is so”)</td>
</tr>
<tr>
<td>NW-55</td>
<td>61</td>
<td>sf (1.0)</td>
<td>Full stop</td>
</tr>
<tr>
<td>NW-32</td>
<td>51</td>
<td>ii (0.45), padvref (0.2), ncn (0.12), mag (0.08), maj (0.06)</td>
<td>Interjections, expression of agreement (“That is so”)</td>
</tr>
<tr>
<td>NW-54</td>
<td>50</td>
<td>sf (1.0)</td>
<td>Question mark</td>
</tr>
<tr>
<td>NW-0</td>
<td>46</td>
<td>ncn (0.35), npd+jxt (0.07)</td>
<td>Various nouns</td>
</tr>
<tr>
<td>NW-14</td>
<td>39</td>
<td>sf (1.0)</td>
<td>Full-stop</td>
</tr>
<tr>
<td>NW-10</td>
<td>34</td>
<td>ncn+jvcs (0.24), mag (0.15), ncn (0.06), pvgsreccs (0.06), ncn+jvcs (0.06), nq (0.06), paavreccs (0.06)</td>
<td>Difficult to characterize (heterogeneous)</td>
</tr>
<tr>
<td>NW-44</td>
<td>34</td>
<td>paavref (0.18), pvgsref (0.15), ncn+jvcs (0.09), pvgsrepref (0.06), mag (0.06), pvgsreccs (0.06), pvgsreccs (0.06), pvgsreccs (0.06), paavref+jvcs (0.06)</td>
<td>Sentence-final verbs</td>
</tr>
<tr>
<td>NW-29</td>
<td>30</td>
<td>mag (0.2), ncn+jvcs (0.1), ncn (0.1), pvgsreccs (0.07), ncn+jvcs (0.07)</td>
<td>Difficult to characterize (heterogeneous)</td>
</tr>
</tbody>
</table>

Table 2: Representative characterizations of the eight most frequent nonterminal expansion rules induced by the NeuralChar model on Korean child-directed speech.

<table>
<thead>
<tr>
<th>Induced rule</th>
<th>Count</th>
<th>Representative characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 → 43 63</td>
<td>84</td>
<td>Attachment of full stop after declaratives</td>
</tr>
<tr>
<td>11 → 3 62</td>
<td>42</td>
<td>Attachment of question mark after questions</td>
</tr>
<tr>
<td>43 → 76 53</td>
<td>20</td>
<td>Attachment of noun before imperatives</td>
</tr>
<tr>
<td>43 → 76 43</td>
<td>13</td>
<td>Attachment of two declarative utterances</td>
</tr>
<tr>
<td>45 → 29 34</td>
<td>12</td>
<td>Left attachment of nouns</td>
</tr>
<tr>
<td>3 → 76 3</td>
<td>11</td>
<td>Attachment of adverb before questions</td>
</tr>
<tr>
<td>3 → 29 89</td>
<td>10</td>
<td>Attachment of noun before question verbs</td>
</tr>
<tr>
<td>53 → 59 75</td>
<td>9</td>
<td>Right attachment of imperative verbs</td>
</tr>
</tbody>
</table>

This indicates that the subword information leveraged by the NeuralChar model results in notably more accurate grammars on an agglutinative language like Korean. In contrast, subword information does not seem to help the induction of a mostly analytic language like English, in which grammatical relationships are primarily conveyed by word order.

### 4.4 Analysis of Categories and Rules Induced by Character-based Model

Subsequently, in order to further analyze how the character-based terminal expansion model helps grammar induction from Korean child-directed speech, the preterminal categories and nonterminal expansion rules induced by the NeuralChar model were compared to those induced by the Neural-
Word model. As Korean is an agglutinative language that marks grammatical information primarily through suffixation, it was hypothesized that the subword information leveraged by the NeuralChar model would result in preterminal categories that are more homogeneous with regard to part-of-speech.

The eight most frequent preterminal categories induced by the NeuralChar model and the NeuralWord model are presented in Table 1. Most notably, the categories used by the NeuralChar model seem to form linguistically coherent clusters; categories NC-29 and NC-2 correspond mostly to nouns (tags starting with ‘n’) while categories NC-38 and NC-16 correspond mostly to predicates (verbs and adjectives; tags starting with ‘p’). In contrast, such systematic categorization was not observed in the categories induced by the NeuralWord model. Many of the frequent categories were heterogeneous with regard to the attested gold part-of-speech; categories like NW-32, NW-10, and NW-29 corresponded to both nouns and verbs, as well as adverbs (tags starting with ‘m’). These results indicate that leveraging subword information allows PCFG induction models to learn syntactic categories that are more linguistically coherent.

The lack of such coherence in categories induced by the NeuralWord model further shows that word order information alone is insufficient to accurately induce syntactic categories from a language with relatively free word order like Korean.

The induced nonterminal expansion rules in Table 2 show that the NeuralChar model learns distinct nonterminal categories corresponding to declaratives (Category 43), questions (Category 3), and imperatives (Category 53). Even though the nonterminal category for questions could be learned trivially based on the presence of a question mark, the finding that the NeuralChar model learns the distinction between declaratives and imperatives is noteworthy, as the two sentence types differ only by the suffix on the sentence-final verb. On the contrary, the frequent rules induced by the NeuralWord model were not very interpretable, other than those that were used to attach punctuation marks.

### 4.5 Replication Using Silver Data

Finally, to examine whether a similar relationship between morphological typology and induction performance is observed in languages other than English and Korean, the NeuralChar and NeuralWord models were also evaluated on Mandarin Chinese and German child-directed speech corpora from CHILDES. The Chinese corpus consists of 19,541 caregiver utterances from the Tong section (Deng et al., 2018) with a mean sentence length of 5.7 words, which were recorded at ages from 1 year 0 months and 4 years 5 months. The German corpus contains 20,000 child-directed utterances randomly sampled from the Leo section (Behrens, 2006), as the original corpus contained many duplicate utterances in interactions between Leo and his caregivers between ages 1 year 11 months and 4 years 11 months. The sampled dataset had a mean sentence length of 6.7 words. However, since gold syntactic annotations were not available for these corpora, the reference trees were automatically generated using the Kitaev et al. (2019) parser. Following similar procedures as Section 4.2, the NeuralChar and NeuralWord models were trained ten times on each corpus using different random seeds. Subsequently, the RH score of each run was calculated, and a paired permutation test was conducted to determine the statistical significance of the difference in model performance.

Figure 2 shows the RH scores from the ten runs of each model trained on the two child-directed
speech corpora. The NeuralChar model (mean RH = 0.502) outperformed the NeuralWord model (mean RH = 0.470) on the German corpus, with the difference in RH being significant according to a permutation test. The same trend was observed on the Chinese corpus, on which the NeuralChar model (mean RH = 0.372) performed better than the NeuralWord model (mean RH = 0.351). The caveat of using silver data to evaluate model performance notwithstanding, a similar relationship between morphological typology and induction performance is observed on German and Chinese; the increase in performance from the NeuralChar model seems to be more salient on the morphologically richer German than on Chinese.\footnote{Although Mandarin Chinese is very analytic in the sense that it has almost no inflectional affixes, the character-based model seems to have helped induction by identifying one-character-long derivational morphemes on compound nouns.}

5 Experiment 2: Evaluation on Multilingual Treebanks

Subsequently, in order to determine if the evaluated models are competitive with state-of-the-art grammar induction models, we present experiments on multilingual newswire treebanks to test the performance of the NeuralChar and NeuralWord models using both labeled evaluation (RH) and standard unlabeled evaluation (bracketing F1). These are evaluated on newswire to compare against models developed on predominantly newswire data. The results show that the NeuralChar model outperforms these models in grammar induction on most languages. In addition, comparison of model performance on different languages hints at the inductive biases embodied by each of these systems.

5.1 Procedures

The NeuralChar and NeuralWord models are evaluated on ten constituency treebanks, including Arabic (Maamouri et al., 2004), Chinese (Xia et al., 2000), English (Marcus et al., 1993), French (Abeillé et al., 2003), German (Skut et al., 1998), Hebrew (Sima’an et al., 2001), Japanese (Alastair et al., 2018), Korean (Han et al., 2006), Polish (Woliński et al., 2018) and Vietnamese (Nguyen et al., 2009).\footnote{Refer to Appendix C for basic statistics of the treebanks.} Compared induction models include a pure word-based Bayesian PCFG model (DIMI, Jin et al., 2018a); a PCFG induction model that generates independently trained character-based word vectors (Flow, Jin et al., 2019); word-based neural models Compound and Compound-v (Kim et al., 2019), which differ in that Compound-v induces sentence-specific grammars, as well as its extension with lexical dependencies (L-PCFG, Zhu et al., 2020). At least three random initial seeds are used for each model and each language, and the average performance of grammars with the highest likelihoods are reported. Hyperparameters for the NeuralChar and NeuralWord models are reported in Appendix A, and those for other baseline models followed reported values in their respective papers.

5.2 Results

Table 3 reports the labeled evaluation results for all models\footnote{Evaluation of the Flow model (Jin et al., 2019) on Vietnamese was not available due to its use of pretrained ELMo embeddings for individual syllables rather than words.} as well as unlabeled evaluation results using F1 for reference. The labeled and unlabeled evaluation scores generally correlate with each other very strongly. First, the NeuralChar model performs very well across the majority of evaluated languages, showing that the character-based lexical expansion model is able to capture regularities in subwords that help grammatical categorization and constituency boundary detection. In addition, all models generally perform better on languages with a small number of high-frequency function words such as German and English, than on languages with a large number of high-frequency function words and affixes such as Chinese and Korean. On languages that primarily use marker affixes (e.g. Arabic, Japanese, and Korean), the results show that the character-based models are able to use information from these affixes for grammar induction, resulting in higher performance compared to word-based induction models.

However, for all models, it still seems much easier to extract statistical information from words than from affixes, as demonstrated by the stark contrast in the performance of the NeuralChar model on Japanese and Korean. Japanese and Korean are both agglutinative languages with similar syntax, but according to the Japanese annotation guidelines, all case markers are separated from their stems and are treated as separate function words. In contrast, according to the Korean annotation guidelines, case markers are treated like suffixes and are left unseparated from the word stem. This difference leads to 25% of the tokens in the Japanese dataset being function words but leaves only 2% of function words in the Korean dataset. This can partially
Table 3: Average labeled Recall-Homogeneity and unlabeled F1 scores from various unsupervised grammar induction models on multilingual treebanks. The upper group of models are word-based models, and the lower group of models have access to subword information. The language codes are: Ar: Arabic; Zh: Chinese; En: English; Fr: French; De: German; He: Hebrew; Ja: Japanese; Ko: Korean; Pl: Polish; Vi: Vietnamese.

explain the results that all models perform much better on Japanese than on Korean.

There are indications of another inductive bias at work in the Compound models. One main architectural difference between NeuralWord and Compound is that Compound distinguishes preterminal (lexical) tags from other nonterminal (phrasal) tags while NeuralWord does not. This means that given the same number of syntactic categories, the NeuralWord models have a larger search space of possible grammars than Compound. This distinction between lexical-phrasal categories seems to benefit induction on Indo-European languages, as can be seen by the higher performance of Compound models. Nonetheless, this distinction does not seem to be helpful on other languages. We leave investigations of the exact nature of this inductive bias to future work, although we conjecture that it could be related to the obligatory use of determiners in Indo-European languages.

Finally, the average RH and F1 scores show that the NeuralChar model would be the best model to try on a new language if no inductive biases pertaining to grammar induction are known about that language. Furthermore, the average F1 scores are very similar across models compared to the average RH scores, which indicates that models that perform similarly in terms of unlabeled evaluation may produce constituent labels of varying quality.

6 Conclusion and Future Work

This paper presents a character-based model and a minimally-manipulated word-based counterpart for neural PCFG induction. The character-based model allows subword information to influence grammar induction, which is more consistent with the information that child language learners are able to leverage. Experiments and analyses using child-directed speech corpora show that the incorporation of subword information results in more accurate grammars with linguistically homogeneous syntactic categories, with its impact being stronger on morphologically richer languages. Additionally, this model achieves state-of-the-art induction results on many languages, providing further support for a distributional model of syntactic acquisition. Taken together, these results indicate that the proposed character-based model incorporates more realistic input and captures more successful learning outcomes, thus making it a more plausible cognitive model.

While this work addresses a major drawback of word-based PCFG induction models, the proposed model nonetheless relies on symbolic representations of characters, which are abstractions from potentially noisy perceptual input. Future work could explore the extent to which syntactic knowledge can be acquired from lower-level (e.g. phonemic or
acoustic) input alone by including a word segmentation task (Elsner and Shain, 2017; Shain and Elsner, 2020) for the model. Additionally, recent work in unsupervised grammar induction (Jin and Schuler, 2020; Zhang et al., 2021) has shown that incorporating visual information in the form of images and videos helps learn constituents that denote entities or action. Adopting such grounded approaches can help answer questions about the extent to which the visual scene or other contexts contain relevant information, and eventually about the nature of input required for syntactic acquisition.

**Acknowledgments**

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

**Ethical Considerations**

Experiments presented in this work used datasets from previously published research (i.e. child-directed speech data and multilingual treebanks), in which the procedures for data collection and validation are outlined.

**References**


A Hyperparameters and Data Preprocessing

The NeuralWord and NeuralChar models differ minimally in terms of hyperparameters used for training. The following values of each hyperparameter apply to both models if not stated otherwise. The number of nonterminals ($C$) is 90. The learning rate for the Adam optimizer is 0.001, with other hyperparameters following the default values of PyTorch. The size of the category embeddings is 128. For the character-based induction model, the size of the hidden and cell states of the LSTMs is 512. The gradients are renormalized when the norm exceeds 5.0. The size of each training batch is 2 sentences. The data likelihood is calculated for the entire training set every two epochs starting from the second epoch. The NeuralWord models usually take about 2 to 8 epochs to converge, and the NeuralChar models take about 12 to 20 epochs to converge, which is indicated by reaching the highest marginal likelihood. All models were trained on V100 GPUs with 16G memory.

Sentences with 40 or fewer words (including punctuation) were kept in the training data, and all words were lowercased. The first 14 and the last 14 characters were concatenated to form a word if the word had more than 28 characters. No limit was imposed on the vocabulary size for any model.

B Morphological Tags in the Korean Annotation Scheme of Choi (2013)

Table 4 provides a description of the morphological tags used in the Korean annotation scheme of Choi (2013).

C Statistics of Treebanks

Table 5 shows the basic statistics of sentences and words from treebanks used in this work.

<table>
<thead>
<tr>
<th>Language</th>
<th># sentences</th>
<th># word types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>12754</td>
<td>30810</td>
</tr>
<tr>
<td>Mandarin</td>
<td>14907</td>
<td>27386</td>
</tr>
<tr>
<td>English</td>
<td>45407</td>
<td>40300</td>
</tr>
<tr>
<td>French</td>
<td>15965</td>
<td>23342</td>
</tr>
<tr>
<td>German</td>
<td>19396</td>
<td>45346</td>
</tr>
<tr>
<td>Hebrew</td>
<td>6189</td>
<td>15249</td>
</tr>
<tr>
<td>Japanese</td>
<td>34675</td>
<td>39333</td>
</tr>
<tr>
<td>Korean</td>
<td>9686</td>
<td>42899</td>
</tr>
<tr>
<td>Polish</td>
<td>13022</td>
<td>35798</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>9553</td>
<td>12277</td>
</tr>
</tbody>
</table>

Table 5: Statistics of the treebanks used in this work.