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 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.[1] 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;

[1] Code used in this work is available at https://github.com/lifengjin/charInduction.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4367–4378
November 7–11, 2021.

Tu, 2012) was not as successful as models of unsupervised constituency parsing (Seginer, 2007; Ponvert 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 compound 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 nonterminal categories is first factored into two separate parts: binary-branching nonterminal expansion rule[2] probabilities, and unary-branching terminal expansion rule probabilities (Jin et al., 2019). Given a tree as a set τ of nodes η 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 τ' of nodes η undergoing terminal expansions $c_\eta \rightarrow w_\eta$ (where w_η is the word at node η), the marginal probability of a sentence σ can be computed as:

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

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:

$$\mathsf{P}(\text{Term} \mid c_\eta) = \underset{\{0,1\}}{\text{softmax}}(\text{N}(\mathbf{E}\,\delta_{c_\eta})), \tag{2}$$

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

$$\mathbf{x}_{b,c_\eta} = \text{ReLU}(\mathbf{W}'_b\ \text{ReLU}(\mathbf{W}_b\,\mathbf{x}_{b-1,c_\eta} + \mathbf{b}_b) + \mathbf{b}'_b) + \mathbf{x}_{b-1,c_\eta}. \tag{3}$$

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

$$\mathbf{x}_{0,c_\eta} = \text{ReLU}(\mathbf{W}_0\,\mathbf{E}\,\delta_{c_\eta} + \mathbf{b}_0), \tag{4}$$

$$s_{c_\eta} = \text{ReLU}(\mathbf{W}_{\text{soft}}\,\mathbf{x}_{B,c_\eta} + \mathbf{b}_{\text{soft}}). \tag{5}$$

[2]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_η} 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_η are defined as:

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

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

$$\begin{aligned}\mathsf{P}(c_\eta \rightarrow c_{\eta 1}\ c_{\eta 2} \mid c_\eta, \text{Term=0}) = \\ \underset{c_{\eta 1}, c_{\eta 2}}{\text{softmax}}(\mathbf{W}_{\text{nont}}\, \mathbf{E}\, \delta_{c_\eta} + \mathbf{b}_{\text{nont}}),\end{aligned} \tag{7}$$

where $\mathbf{W}_{\text{nont}} \in \mathbb{R}^{C^2 \times d}$ and $\mathbf{b}_{\text{nont}} \in \mathbb{R}^{C^2}$ 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_η are defined as:

$$\begin{aligned}\mathsf{P}(c_\eta \rightarrow w_\eta) = \mathsf{P}(\text{Term=1} \mid c_\eta) \cdot \\ \mathsf{P}(c_\eta \rightarrow w_\eta \mid c_\eta, \text{Term=1}),\end{aligned} \tag{8}$$

where w_η is the generated word token at node η. The representation of the word w_η 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:

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

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:

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

where the character or letter sequence $l_1, \ldots, l_n$ comprises the word w_η 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 is:

$$\begin{aligned}\mathsf{P}(l_i \mid c_\eta, l_1, \ldots, l_{i-1}) = \\ \underset{l_i}{\text{softmax}}(\mathbf{W}_{\text{char}}\, \mathbf{h}_{i,B,c_\eta} + \mathbf{b}_{\text{char}}),\end{aligned} \tag{11}$$

where $\mathbf{W}_{\text{char}} \in \mathbb{R}^{|L| \times h}$, $\mathbf{b}_{\text{char}} \in \mathbb{R}^{|L|}$, 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_η:

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

where $\mathbf{W}_{b,\text{term}} \in \mathbb{R}^{h \times d}$ and $\mathbf{b}_{b,\text{term}} \in \mathbb{R}^h$ 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

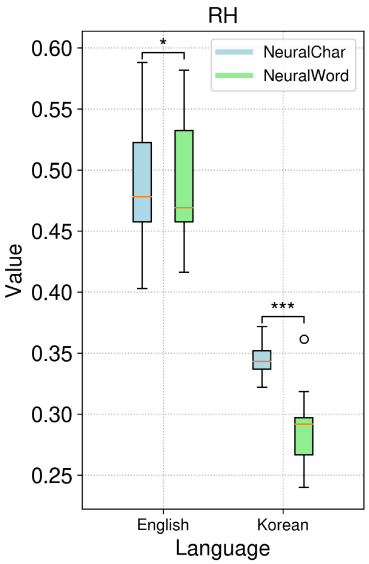

Figure 1: Box plots showing the RH scores from ten runs of the NeuralChar and NeuralWord models trained on the English (Eve) and Korean (Jong) corpora from CHILDES. Statistical significance was determined by a paired permutation test at the sentence level (*: $p < 0.05$, ***: $p < 0.001$).

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

[3]Homogeneity 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.

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

[5]The 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.

Induced category	Count	Attested category (relative frequency)	Representative characterization
NC-63	100	sf (1.0)	Full stop
NC-29	73	npd+jxt (0.23), nq (0.12), ncn (0.12), npd+jcs (0.1), npd (0.1), nq+jcs (0.07), ncn+jcs (0.05)	Nouns for child (“Jong”), demonstrative pronoun (“this”)
NC-62	48	sf (1.0)	Question mark
NC-38	25	px+ef (0.32), pvg+ef (0.2), paa+ef (0.2), pvg+ep+ef (0.16)	Sentence-final verbs
NC-16	21	pvg+ecx (0.67), pvg+ecs (0.14), paa+ecc (0.1), paa+ef (0.1)	Verbs preceding auxiliary verbs
NC-2	20	ncn (0.55), ncn+jcj (0.15), ncn+jcs (0.1), pad+ef (0.05), mag (0.05), ncn+jxt (0.05), pvd+ecs (0.05)	Nouns for caretaker (“mom”)
NC-6	20	ii (1.0)	Interjections
NC-7	20	pad+ef (1.0)	Expression of agreement (“That is so”)
NW-55	61	sf (1.0)	Full stop
NW-32	51	ii (0.45), pad+ef (0.2), ncn (0.12), mag (0.08), maj (0.06)	Interjections, expression of agreement (“That is so”)
NW-54	50	sf (1.0)	Question mark
NW-0	46	ncn (0.35), npd+jxt (0.07)	Various nouns
NW-14	39	sf (1.0)	Full stop
NW-10	34	ncn+jcs (0.24), mag (0.15), ncn (0.06), pvg+ecs (0.06), ncn+jxc (0.06), nq (0.06), paa+ecs (0.06)	Difficult to characterize (heterogeneous)
NW-44	34	paa+ef (0.18), pvg+ef (0.15), ncn+jp+ef (0.09), pvg+ep+ef (0.06), mag (0.06), pvg+etm (0.06), pvg+ef+jxf (0.06), paa+ef+jxf (0.06)	Sentence-final verbs
NW-29	30	mag (0.2), ncn+jcs (0.1), ncn (0.1), paa+etm (0.1), npp (0.07), pvg+ecx (0.07), ncn+jxt (0.07)	Difficult to characterize (heterogeneous)

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.

as a proxy for prosodic cues about phrasal boundaries, but were removed afterward for labeled evaluation. Subsequently, in order to examine whether there was a significant difference in the RH measure between the NeuralChar and NeuralWord models, the conventional paired permutation test used in supervised parsing was conducted following Jin et al. (2021). In a paired permutation test, the predicted trees from two induced grammars are randomly permuted in order to calculate an empirical distribution of the difference in the chosen evaluation metric. This empirical distribution calculates the probability of the observed difference due to chance. Since the evaluation metric of interest was RH, the per-sentence-recall and per-sentence-homogeneity of each sentence were first calculated for each model, which were subsequently permuted randomly to calculate an empirical distribution over the difference in RH.

4.3 Results

Figure 1 shows the RH scores from the ten runs of each model trained on the two child-directed speech corpora. On the Korean corpus, the NeuralChar model (mean RH = 0.388) strongly outperformed the NeuralWord model (mean RH = 0.291), with the difference in RH being significant according to a permutation test. The opposite trend was observed on the English corpus, where the NeuralWord model (mean RH = 0.488) performed better than the NeuralChar model (mean RH = 0.487).

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

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

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.[6] 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.

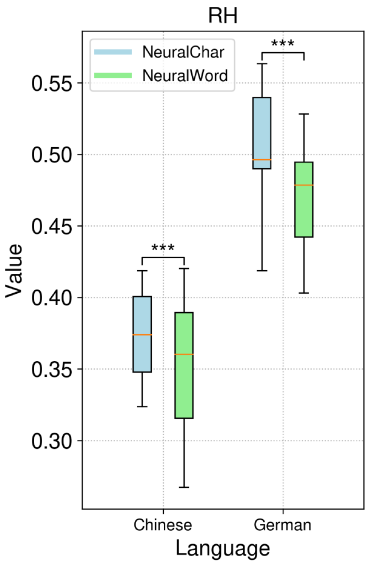

Figure 2: Box plots showing the RH scores from ten runs of the NeuralChar and NeuralWord models trained on the Chinese (Tong) and German (Leo) corpora from CHILDES. Statistical significance was determined by a paired permutation test at the sentence level (***: $p < 0.001$).

Figure 2 shows the RH scores from the ten runs of each model trained on the two child-directed

[6]For each model, the run with the highest likelihood from Section 4.2 was analyzed.

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.[7]

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).[8] 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[9] 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

[7] 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.

[8] Refer to Appendix C for basic statistics of the treebanks.

[9] 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.

Models / RH	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018a)	16.5	12.4	23.4	16.8	10.3	14.9	23.5	7.1	6.3	8.1	13.9
Compound (Kim et al., 2019)	21.1	21.2	**36.8**	37.7	**41.4**	23.5	15.2	5.6	**35.1**	15.8	25.3
Compound-v (Kim et al., 2019)	16.9	22.6	35.0	39.9	39.4	29.1	13.1	7.0	33.0	**24.0**	26.0
L-PCFG (Zhu et al., 2020)	24.4	19.4	15.0	18.2	28.3	17.0	30.1	10.2	17.4	10.2	19.0
NeuralWord (this work)	23.0	20.8	29.7	29.8	33.8	21.6	29.8	11.7	22.0	15.1	23.7
Flow (Jin et al., 2019)	25.4	18.7	21.6	25.3	29.7	25.4	24.4	15.0	31.0	—	24.1
NeuralChar (this work)	**29.1**	**23.9**	33.4	**40.7**	39.3	**29.5**	**40.2**	**16.3**	21.0	12.8	**28.5**

Models / F1	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018a)	35.3	36.6	50.6	39.6	36.4	45.4	36.2	26.5	43.2	**42.7**	39.3
Compound (Kim et al., 2019)	32.4	34.2	**51.7**	48.2	**49.7**	40.5	22.9	19.1	**50.1**	34.3	38.3
Compound-v (Kim et al., 2019)	27.6	37.4	50.9	49.6	47.9	**49.2**	21.6	20.7	47.2	38.3	39.1
L-PCFG (Zhu et al., 2020)	**45.0**	**46.2**	36.2	34.4	46.8	38.4	45.2	30.0	32.1	27.3	38.2
NeuralWord (this work)	36.9	41.3	44.4	41.5	44.4	40.0	42.4	23.3	35.2	37.5	38.7
Flow (Jin et al., 2019)	35.3	38.1	38.6	40.3	38.0	45.0	33.8	34.4	47.1	—	39.0
NeuralChar (this work)	42.0	44.9	49.9	**51.5**	47.7	48.6	**55.9**	**34.6**	33.1	28.7	**43.7**

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

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a Treebank for French. In *Treebanks: Building and using parsed corpora*, pages 165–187. Springer Netherlands, Dordrecht, Netherlands.

Butler Alastair, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. 2018. The Keyaki Treebank Parsed Corpus.

Richard N. Aslin and Elissa L. Newport. 2014. Distributional language learning: Mechanisms and models of category formation. *Language Learning*, 64(s2):86–105.

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

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Jinho D. Choi. 2013. Preparing Korean data for the shared task on parsing morphologically rich languages. Technical Report 1309.1649, arXiv.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Noam Chomsky. 1980. On cognitive structures and their development: A reply to Piaget. In *Language and learning: The debate between Jean Piaget and Noam Chomsky*, pages 751–755. Harvard University Press, Cambridge, MA.

Forrest Davis and Marten van Schijndel. 2020. Recurrent neural network language models always learn English-like relative clause attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990.

Xiangjun Deng, Virginia Yip, Brian Macwhinney, Stephen Matthews, Mai Ziyin, Zhong Jing, and Hannah Lam. 2018. A Multimedia Corpus of Child Mandarin: The Tong Corpus. *The Journal of Chinese Linguistics*, 46(1):69–92.

Andrew Drozdov, Subendhu Rongali, Yi-Pei Chen, Tim O'Gorman, Mohit Iyyer, and Andrew McCallum. 2020. Unsupervised parsing with S-DIORA: Single tree encoding for deep inside-outside recursive autoencoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4832–4845.

Andrew Drozdov, Patrick Verga, Mohit Yadav, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised latent tree induction with deep inside-outside recursive auto-encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1129–1141.

Cristina Dye, Yarden Kedar, and Barbara Lust. 2019. From lexical to functional categories: New foundations for the study of language development. *First Language*, 39(1):9–32.

Micha Elsner and Cory Shain. 2017. Speech segmentation with a neural encoder model of working memory. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1080.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Na-Rae Han, Shijong Ryu, Sook-Hee Chae, Seung-yun Yang, Seunghun Lee, and Martha Palmer. 2006. Korean Treebank Annotations Version 2.0.

Zellig Harris. 1954. Distributional structure. In *The structure of language: Readings in the philosophy of language*, volume 10, pages 33–49. Prentice-Hall, Hoboken, NJ.

Betty Hart and Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Brookes Publishing, Baltimore, MD.

Etsuko Haryu and Sachiyo Kajikawa. 2016. Use of bound morphemes (noun particles) in word segmentation by Japanese-learning infants. *Journal of Memory and Language*, 88:18–27.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018a. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2721–2731.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2018b. Unsupervised grammar induction with depth-bounded PCFG. *Transactions of the Association for Computational Linguistics*, 6:211–224.

Lifeng Jin, Finale Doshi-Velez, Timothy Miller, Lane Schwartz, and William Schuler. 2019. Unsupervised learning of PCFGs with normalizing flow. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2442–2452.

Lifeng Jin and William Schuler. 2020. Grounded PCFG induction with images. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 396–408.

Lifeng Jin, Lane Schwartz, Finale Doshi-Velez, Timothy Miller, and William Schuler. 2021. Depth-bounded statistical PCFG induction as a model of human grammar acquisition. *Computational Linguistics*, 47(1):181–216.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov chain Monte Carlo. *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 139–146.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505.

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. 2020. An empirical comparison of unsupervised constituency parsing methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3278–3283.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. *NEMLAR Conference on Arabic Language Resources and Tools*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Toben H. Mintz. 2013. The segmentation of sub-lexical morphemes in English-learning 15-month olds. *Frontiers in Psychology*, 4(24):1–12.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 182–185.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Ju-Yeon Ryu, Kaoru Horie, and Yasuhiro Shirai. 2015. Acquisition of the Korean imperfective aspect markers –ko iss– and –a iss– by Japanese learners: A multiple-factor account. *Language Learning*, 65(4):791–823.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

Cory Shain and Micha Elsner. 2020. Acquiring language from speech by learning to remember and predict. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 195–214.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron C. Courville. 2019. Ordered Neurons: Integrating tree structures into recurrent neural networks. In *7th International Conference on Learning Representations*.

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a tree-bank of modern Hebrew text. *Traitment Automatique des Langues*, 42(2):1–34.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations*.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL): What, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12.

Kewei Tu. 2012. *Unsupervised learning of probabilistic grammars*. Ph.D. thesis, Iowa State University.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070.

Marcin Woliński, Elżbieta Hajnicz, and Tomasz Bartosiak. 2018. A new version of the składnica treebank of Polish harmonised with the walenty valency dictionary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021. Video-aided unsupervised grammar induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1513–1524.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

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).

Tag label	Tag description
ncn	Non-predicative common noun
npd	Demonstrative pronoun
npp	Personal pronoun
nq	Proper noun
jcj	Conjunctive case particle
jcs	Subjective case particle
jp	Predicative marker
jxc	Common auxiliary
jxf	Final auxiliary
jxt	Topical auxiliary
paa	Attributive adjective
pad	Demonstrative adjective
pvd	Demonstrative verb
pvg	General verb
px	Auxiliary verb
ecc	Coordinate conjunction EM
ecs	Subordinate conjunction EM
ecx	Auxiliary conjunction EM
ef	Final EM
ep	Pre-final EM
etm	Adnominalizing EM
mag	General adverb
maj	Conjunctive adverb
ii	Interjection
sf	Sentence-final punctuation

Table 4: Description of the morphological tags used in the Korean annotation scheme of Choi (2013). EM: ending marker.

C Statistics of Treebanks

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

Language	# sentences	# word types
Arabic	12754	30810
Mandarin	14907	27386
English	45407	40300
French	15965	23342
German	19396	45346
Hebrew	6189	15249
Japanese	34675	39333
Korean	9686	42899
Polish	13022	35798
Vietnamese	9553	12277

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