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

Byung-Doh Oh

Department of Linguistics
The Ohio State University

April 26, 2022
OSU Linguistics Colloquiumfest

Roadmap

Modeling task: Unsupervised PCFG induction

Model description: *NeuralWord* and *NeuralChar* (Jin, 2020)

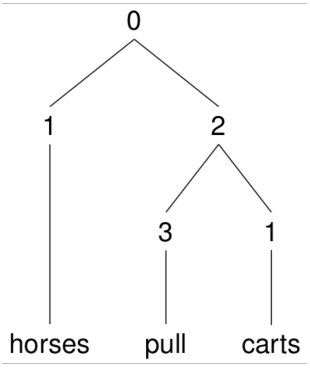
Experiment 1: Evaluation on child-directed speech

Experiment 2: Evaluation on multilingual treebanks

Conclusion and future directions

Modeling task: Unsupervised PCFG induction

Unsupervised PCFG induction

horses pull carts → 

Unsupervised PCFG induction

Nonterminal expansion probabilities:

$P(0 \rightarrow 1\ 2)$

$P(2 \rightarrow 3\ 1)$

...

Terminal expansion probabilities:

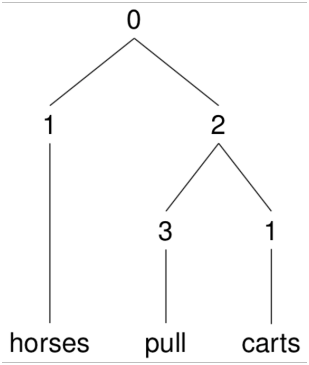
$P(1 \rightarrow \text{horses})$

$P(3 \rightarrow \text{pull})$

$P(1 \rightarrow \text{carts})$

...

$\rightarrow$



Unsupervised PCFG induction

Shows the extent to which grammars can be learned from distributional data alone

Recent neural approaches fairly successful (Kim et al., 2019; Yang et al., 2021; Zhu et al., 2020)

However, word-based PCFGs cannot inspect word affixes

Terminal expansion probabilities:

$P(1 \rightarrow \text{horses})$

$P(3 \rightarrow \text{pull})$

$P(1 \rightarrow \text{carts})$

Unsupervised PCFG induction

Child language learners are sensitive to functional affixes (Dye et al., 2019; Haryu & Kajikawa, 2016; Mintz, 2013)

Word-based models are less appropriate for morphologically rich languages (Tsarfaty et al., 2010)

This work presents

- Character-/word-based counterpart models for neural PCFG induction (Jin, 2020)
- Experiments on child-directed speech corpora and multilingual treebanks

Model description: *NeuralWord* and *NeuralChar* (Jin, 2020)

Model description

Objective function: Marginal probability of sentence σ

- Maximize probabilities of all parse trees that are consistent with observed sentence σ

$$\mathsf{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \to c_{\eta 1}\ c_{\eta 2}} \mathsf{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \to w_\eta} \mathsf{P}(c_\eta \to w_\eta)$$

- Model has access to C category vectors $\mathbf{v}_c$, which are used for probability calculation

"Split" model: Nonterminal or terminal expansion?

- A given category c_η can undergo both nonterminal and terminal expansion

$$\mathsf{P}(\text{Term} \mid c_\eta) = \underset{\{0,1\}}{\mathsf{SoftMax}}(\mathsf{ResNet}_{\text{split}}(\mathbf{v}_{c_\eta}))$$

Model description

Objective function: Marginal probability of sentence σ

- Maximize probabilities of all parse trees that are consistent with observed sentence σ

$$\mathrm{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \to c_{\eta 1}\ c_{\eta 2}} \mathrm{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \to w_\eta} \mathrm{P}(c_\eta \to w_\eta)$$

- Model has access to C category vectors $\mathbf{v}_c$, which are used for probability calculation

Nonterminal expansion probabilities

$$\mathrm{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) = \mathrm{P}(\mathrm{Term}{=}0 \mid c_\eta) \cdot \underset{c_{\eta 1}, c_{\eta 2}}{\mathrm{SoftMax}}(\mathbf{W}_{\mathrm{nont}}\, \mathbf{v}_{c_\eta})$$

Model description

Objective function: Marginal probability of sentence σ

- Maximize probabilities of all parse trees that are consistent with observed sentence σ

$$\mathrm{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \to c_{\eta 1}\ c_{\eta 2}} \mathrm{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \to w_\eta} \mathrm{P}(c_\eta \to w_\eta)$$

- Model has access to C category vectors $\mathbf{v}_c$, which are used for probability calculation

Character-based terminal expansion probabilities (*NeuralChar*)

$$\mathrm{P}(c_\eta \to w_\eta) = \mathrm{P}(\mathrm{Term}{=}1 \mid c_\eta) \cdot \prod_{l_i \in \{l_1, \ldots, l_n\}} \mathrm{P}(l_i \mid c_\eta, l_1, \ldots, l_{i-1})$$

Model description

Objective function: Marginal probability of sentence σ

- Maximize probabilities of all parse trees that are consistent with observed sentence σ

$$\mathrm{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \to c_{\eta 1}\ c_{\eta 2}} \mathrm{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \to w_\eta} \mathrm{P}(c_\eta \to w_\eta)$$

- Model has access to C category vectors $\mathbf{v}_c$, which are used for probability calculation

Word-based terminal expansion probabilities (*NeuralWord*)

$$\mathrm{P}(c_\eta \to w_\eta) = \mathrm{P}(\mathrm{Term}{=}1 \mid c_\eta) \cdot \underset{w_\eta}{\mathrm{SoftMax}}(\mathrm{ResNet}_{\mathrm{term}}(\mathbf{v}_{c_\eta}))$$

Model description

Objective function: Marginal probability of sentence σ

- Maximize probabilities of all parse trees that are consistent with observed sentence σ

$$\mathsf{P}(\sigma) = \sum_{\tau \text{ for } \sigma} \prod_{\eta \in \tau \text{ s.t. } c_\eta \to c_{\eta 1}\ c_{\eta 2}} \mathsf{P}(c_\eta \to c_{\eta 1}\ c_{\eta 2}) \cdot \prod_{\eta \in \tau \text{ s.t. } c_\eta \to w_\eta} \mathsf{P}(c_\eta \to w_\eta)$$

- Model has access to C category vectors $\mathbf{v}_c$, which are used for probability calculation

- During training, the inside algorithm is used to calculate the marginal probability $\mathsf{P}(\sigma)$
- During evaluation, the CYK algorithm is used to return the most likely parse tree for σ

Experiment 1: Evaluation on child-directed speech

Experiment 1

NeuralChar and *NeuralWord* trained and evaluated on transcriptions of child-directed speech from CHILDES (MacWhinney, 2000)

- English (Brown, 1973): Eve (1;6-2;3)
- Korean (Ryu et al., 2015): Jong (1;3-3;5)

Reference trees: Manually annotated syntactic trees

- English (Brown, 1973): Penn Treebank-style annotations (Pearl & Sprouse, 2013)
- Korean (Ryu et al., 2015): Silver trees generated from a supervised parser (Kitaev et al., 2019) then manually corrected

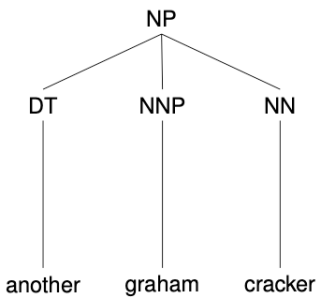
Results from 10 runs using 90 categories

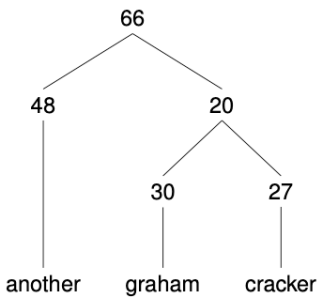
Experiment 1

Evaluation metric: Recall-Homogeneity (RH; Jin et al., 2021)

- Recall: How well does learned grammar G recall attested constituents? (↑)
- Homogeneity: How homogeneous are syntactic categories of learned grammar G? (↑)

Does not penalize a grammar for finer-grained analyses of constituents or parts-of-speech



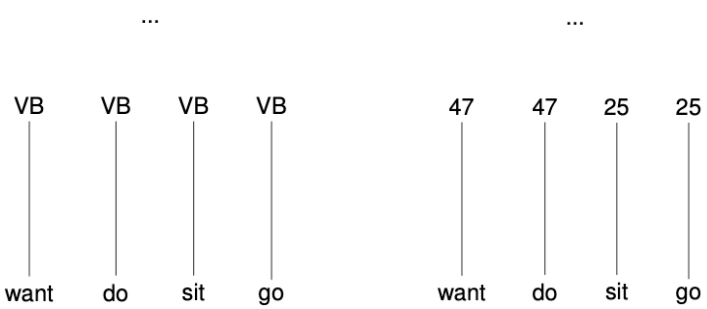


Experiment 1

Evaluation metric: Recall-Homogeneity (RH; Jin et al., 2021)

- Recall: How well does learned grammar G recall attested constituents? (↑)
- Homogeneity: How homogeneous are syntactic categories of learned grammar G? (↑)

Does not penalize a grammar for finer-grained analyses of constituents or parts-of-speech

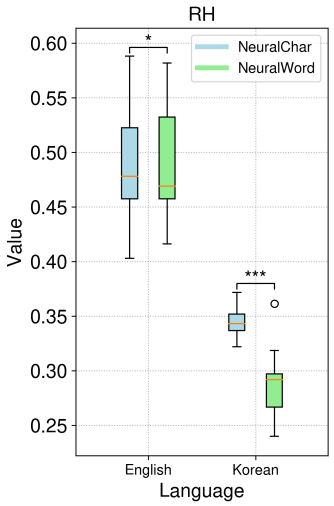


Experiment 1

Statistical testing: Paired permutation test (Demšar, 2006)

- Non-parametric alternative to the *t*-test
- Predicted trees from two models were randomly permuted to calculate an empirical distribution over the difference in RH
- Calculates the probability of the observed difference due to chance

Results


RH
NeuralChar
NeuralWord
*
***
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
Value
English
Korean
Language

Induced preterminal categories

Induced category	Count	Attested category (relative frequency)	Examples
NC-63	100	sf (1.0)	.
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)	이거는, 종현이, 이거, 이게, 아빠, 종현아, 종현이가, 이건
NC-62	48	sf (1.0)	?
NC-38	25	px+ef (0.32), pvg+ef (0.2), paa+ef (0.2), pvg+ep+ef (0.16)	와, 있어, 먹어, 갔었어, 쪘다, 쪘네, 했었어, 놀구요
NC-16	21	pvg+ecx (0.67), pvg+ecs (0.14), paa+ecc (0.1), paa+ef (0.1)	가져, 타러, 보고, 많아요, 알고, 길고, 작아요
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)	엄마, 엄마랑, 엄마가, 그래, 그냥, 엄마는, 그러고
NC-6	20	ii (1.0)	아이구, 아우, 아이고, 아휴, 오, 오오
NC-7	20	pad+ef (1.0)	그렇지, 그래, 그지, 그지요
NW-55	61	sf (1.0)	.
NW-32	51	ii (0.45), pad+ef (0.2), ncn (0.12), mag (0.08), maj (0.06)	그렇지, 어, 짠, 또, 아빠, 자, 엄마, 여기
NW-54	50	sf (1.0)	?
NW-0	46	ncn (0.35), npd+jxt (0.07)	이거는, 이게, 여기, 이, 물, 책, 엄마, 꽃
NW-14	39	sf (1.0)	.
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)	많이, 책도, 목이, 꽃이, 가렸네, 백일, 전신, 살이
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)	적어요, 아빠가, 때야, 쪘다, 쪘네, 나와, 그냥, 목욕했어, 보네
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)	종현이, 너, 다, 작은, 구두는, 진짜, 살, 디게

Induced grammatical rules

Induced rule	Count	Representative characterization	Examples
11 → 43 63	84	Attachment of full stop after declaratives	아이구 이뻐라 + ., 이리로 와 엄마랑 보자 + ., 아우 이쁘다 우리 종현이네 + .
11 → 3 62	42	Attachment of question mark after questions	이게 누구에요 + ?, 목욕하는 거지 종현이가 + ?, 이거는 누구야 + ?
43 → 76 53	20	Attachment of noun before imperatives	물 + 해 봐, 어 책 + 가져 와, 엄마 구두하고 종현이 구두 + 찾아 봐
43 → 76 43	13	Attachment of two declarative utterances	아이구 이뻐라 + 우리 쫑꼬가 이렇게 뚱뚱했어요, 이거는 짧고요 + 이거는 길어요
45 → 29 34	12	Left attachment of nouns	종현이 + 한 살, 이게 + 종현이 백일, 이게 + 언젠가
3 → 76 3	11	Attachment of adverb before questions	또 + 이거 누구야, 또 + 이거는, 또 + 또
3 → 29 89	10	Attachment of noun before question verbs	이게 + 누구에요, 이거는 + 누구야, 이거 + 누구야
53 → 59 75	9	Right attachment of imperative verbs	해 + 봐, 와 + 봐, 찾아 + 봐

Frequent rules from *NeuralWord* not very interpretable, other than those for punctuation

Replication using silver data

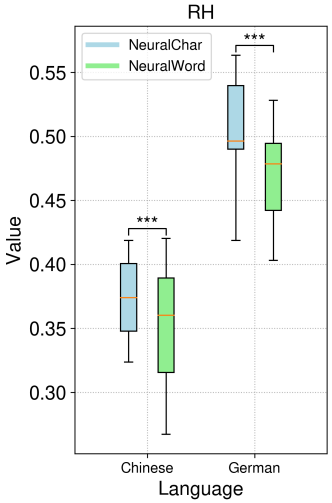
NeuralChar and *NeuralWord* trained and evaluated on transcriptions of child-directed speech from CHILDES (MacWhinney, 2000)

- Chinese (Deng et al., 2018): Tong (1;0-4;5)
- German (Behrens, 2006): Leo (1;11-4;11)

Reference trees: Automatically generated syntactic trees (Kitaev et al., 2019)

Recall-Homogeneity from 10 runs using 90 categories

Results



Colors: 红色, 蓝色, 黄色, 橘黄色; Languages: 英语, 汉语, 日语

Experiment 2: Evaluation on multilingual treebanks

Experiment 2

NeuralChar and *NeuralWord* trained and evaluated on 10 newswire constituency treebanks

- 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), Vietnamese (Nguyen et al., 2009)

Recall-Homogeneity and F1 scores from 3 runs using 90 categories

Compared against recent PCFG induction systems

Results

Models / RH	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	**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., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	42.0	44.9	49.9	**51.5**	47.7	48.6	**55.9**	**34.6**	33.1	28.7	**43.7**

Results

Models / RH	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	**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., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	42.0	44.9	49.9	**51.5**	47.7	48.6	**55.9**	**34.6**	33.1	28.7	**43.7**

Results

Models / RH	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	**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., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	42.0	44.9	49.9	**51.5**	47.7	48.6	**55.9**	**34.6**	33.1	28.7	**43.7**

Japanese: 土工 が, 足音 が, 姿 が; Korean: 르노가, 의지가, 군대가

Results

Models / RH	Individual languages										Average
	Ar	Zh	En	Fr	De	He	Ja	Ko	Pl	Vi	
DIMI (Jin et al., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	**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., 2018)	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 (Jin, 2020)	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 (Jin, 2020)	42.0	44.9	49.9	**51.5**	47.7	48.6	**55.9**	**34.6**	33.1	28.7	**43.7**

Conclusion and future directions

Conclusion

Neural models for unsupervised PCFG induction (Jin, 2020)

- Allows clean manipulation of terminal expansion model

Subword information leads to more accurate grammars on child-directed speech

- Bigger impact on morphologically richer languages

Strong induction results on multilingual treebanks, especially on labeled evaluation

Further support for a distributional model of syntactic acquisition

Future directions

Incorporating token-specific contextualized word representations (e.g. Devlin et al., 2018)

Modeling language acquisition with more realistic input

- Ideally starting with acoustic signals à la Shain and Elsner (2020)
- Maybe a more reasonable middle ground: syllable-level input
- Visually grounded approaches (Jin & Schuler, 2020; Zhang et al., 2021)

Thank you for listening!

Thanks to Clippers for constructive feedback
...and to my committee members for their kind understanding
...and most importantly to William Schuler and Lifeng Jin for their patient mentorship

Supplementary slides

Model hyperparameters

Hyperparameter	Value
Number of categories (both)	90
Size of category embeddings (both)	128
Size of LSTM cell/hidden states (*NeuralChar*)	512
Optimizer	Adam (Kingma & Ba, 2015)
Learning rate	0.001
Gradient clipping threshold	5.0
Batch size	2 sentences
Maximum sentence length	40 words

Morphological tags in the Korean annotation scheme (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

Statistics of training data

Language	# sentences	# word types
English (Eve)	14251	1958
Korean (Jong)	28620	16079
Chinese (Tong)	19541	3036
German (Leo)	20000	7545
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

References I

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. *Treebanks: Building and using parsed corpora* (pp. 165–187). Springer Netherlands. https://doi.org/10.1007/978-94-010-0201-1_10

Alastair, B., Yoshimoto, K., Hiyama, S., Horn, S. W., Nagasaki, I., & Kubota, A. (2018). The Keyaki Treebank Parsed Corpus. http://www.compling.jp/keyaki/

Behrens, H. (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes*, *21*(1-3), 2–24. https://doi.org/10.1080/01690960400001721

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Choi, J. D. (2013). *Preparing Korean data for the shared task on parsing morphologically rich languages* (tech. rep. No. 1309.1649). arXiv. http://arxiv.org/abs/1309.1649

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Deng, X., Yip, V., Macwhinney, B., Matthews, S., Ziyin, M., Jing, Z., & Lam, H. (2018). A multimedia corpus of child Mandarin: The Tong Corpus. *The Journal of Chinese Linguistics*, *46*(1), 69–92. https://doi.org/10.1353/jcl.2018.0002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/arXiv:1811.03600v2

References II

Dye, C., Kedar, Y., & Lust, B. (2019). From lexical to functional categories: New foundations for the study of language development. *First Language*, *39*(1), 9–32. https://doi.org/10.1177/0142723718809175

Han, N.-R., Ryu, S., Chae, S.-H., Yang, S.-y., Lee, S., & Palmer, M. (2006). Korean Treebank Annotations Version 2.0.. https://doi.org/10.35111/02nk-p662

Haryu, E., & Kajikawa, S. (2016). Use of bound morphemes (noun particles) in word segmentation by Japanese-learning infants. *Journal of Memory and Language*, *88*, 18–27. https://doi.org/https://doi.org/10.1016/j.jml.2015.11.007

Jin, L. (2020). *Computational Modeling of Syntax Acquisition with Cognitive Constraints* (Doctoral dissertation). The Ohio State University.

Jin, L., Doshi-Velez, F., Miller, T., Schuler, W., & Schwartz, L. (2018). Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2721–2731. https://aclanthology.org/D18-1292

Jin, L., Doshi-Velez, F., Miller, T., Schwartz, L., & Schuler, W. (2019). Unsupervised learning of PCFGs with normalizing flow. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2442–2452. https://aclanthology.org/P19-1234

Jin, L., & Schuler, W. (2020). Grounded PCFG induction with images. *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*, 396–408. https://www.aclweb.org/anthology/2020.aacl-main.42

References III

Jin, L., Schwartz, L., Doshi-Velez, F., Miller, T., & Schuler, W. (2021). Depth-bounded statistical PCFG induction as a model of human grammar acquisition. *Computational Linguistics*, *47*(1), 181–216. https://doi.org/10.1162/coli_a_00399

Kim, Y., Dyer, C., & Rush, A. (2019). Compound probabilistic context-free grammars for grammar induction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2369–2385. https://aclanthology.org/P19-1228

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *ICLR*. https://doi.org/10.1063/1.4902458

Kitaev, N., Cao, S., & Klein, D. (2019). Multilingual constituency parsing with self-attention and pre-training. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3499–3505. https://www.aclweb.org/anthology/P19-1340

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. *NEMLAR Conference on Arabic Language Resources and Tools*. https://dl.acm.org/doi/pdf/10.5555/1621804.1621808

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. https://www.aclweb.org/anthology/J93-2004

References IV

Mintz, T. H. (2013). The segmentation of sub-lexical morphemes in English-learning 15-month olds. *Frontiers in Psychology*, *4*(24), 1–12. https://doi.org/https://doi.org/10.3389/fpsyg.2013.00024

Nguyen, P.-T., Vu, X.-L., Nguyen, T.-M.-H., Nguyen, V.-H., & Le, H.-P. (2009). Building a large syntactically-annotated corpus of Vietnamese. *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, 182–185. https://www.aclweb.org/anthology/W09-3035

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*(1), 23–68. https://doi.org/10.1080/10489223.2012.738742

Ryu, J.-Y., Horie, K., & Shirai, Y. (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. https://doi.org/10.1111/lang.12132

Shain, C., & Elsner, M. (2020). Acquiring language from speech by learning to remember and predict. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 195–214. https://www.aclweb.org/anthology/2020.conll-1.15

Sima'an, K., Itai, A., Winter, Y., Altman, A., & Nativ, N. (2001). Building a tree-bank of modern Hebrew text. *Traitment Automatique des Langues*, *42*(2), 1–34. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.5385&rep=rep1&type=pdf

References V

Skut, W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A Linguistically Interpreted Corpus of German Newspaper Text. *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation.*, 7. http://arxiv.org/abs/cmp-lg/9807008

Tsarfaty, R., Seddah, D., Goldberg, Y., Kübler, S., Candito, M., Foster, J., Versley, Y., Rehbein, I., & Tounsi, L. (2010). Statistical parsing of morphologically rich languages (SPMRL): What, how and whither. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, 1–12. https://aclanthology.org/W10-1401

Woliński, M., Hajnicz, E., & Bartosiak, T. (2018). A new version of the składnica treebank of Polish harmonised with the walenty valency dictionary. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. https://aclanthology.org/L18-1289

Xia, F., Palmer, M., Xue, N., Okurowski, M. E., Kovarik, J., Chiou, F.-D., Huang, S., Kroch, T., & Marcus, M. (2000). Developing guidelines and ensuring consistency for Chinese text annotation. *Proceedings of the Second International Conference on Language Resources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2000/pdf/287.pdf

Yang, S., Zhao, Y., & Tu, K. (2021). PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1487–1498. https://aclanthology.org/2021.naacl-main.117

References VI

Zhang, S., Song, L., Jin, L., Xu, K., Yu, D., & Luo, J. (2021). Video-aided unsupervised grammar induction. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1513–1524. https://aclanthology.org/2021.naacl-main.119

Zhu, H., Bisk, Y., & Neubig, G. (2020). The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, *8*, 647–661. https://aclanthology.org/2020.tacl-1.42