Memory-Bounded Left-Corner Unsupervised Grammar Induction on Child-Directed Input

Cory Shain[1], William Bryce[2], Lifeng Jin[1], Victoria Krakovna[3], Finale Doshi-Velez[4], Timothy Miller[5,6], William Schuler[1], and Lane Schwartz[2]

[1]Dept of Linguistics, The Ohio State University
[2]Dept of Linguistics, University of Illinois at Urbana-Champaign
[3]Dept of Statistics, Harvard University
[4]School of Engineering & Applied Sciences, Harvard University
[5]Boston Children's Hospital
[6]Harvard Medical School

14 Dec. 2016, *COLING 2016*

Modeling syntax acquisition with unsupervised parsing

- Unsupervised grammar induction = inferring syntax from raw text
- Important for:
    - NLP in resource-poor languages
    - Syntactic acquisition modeling

Modeling syntax acquisition with unsupervised parsing

- Unsupervised grammar induction = inferring syntax from raw text
- Important for:
  - NLP in resource-poor languages
  - Syntactic acquisition modeling

Modeling syntax acquisition with unsupervised parsing

- Unsupervised grammar induction = inferring syntax from raw text
- Important for:
  - NLP in resource-poor languages
  - Syntactic acquisition modeling

Modeling syntax acquisition with unsupervised parsing

- Unsupervised grammar induction = inferring syntax from raw text
- Important for:
  - NLP in resource-poor languages
  - Syntactic acquisition modeling

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:
  - Left-corner parsing (Johnson-Laird 1983; Abney and Johnson 1991; Gibson 1991; Resnik 1992; Stabler 1994; Lewis and Vasishth 2005)
  - Constraints on working memory (Miller 1956; Cowan 2001; McElree 2001; Van Dyke and Johns 2012)

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:
  - Left-corner parsing (Johnson-Laird 1983; Abney and Johnson 1991; Gibson 1991; Resnik 1992; Stabler 1994; Lewis and Vasishth 2005)

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:
  - Left-corner parsing (Johnson-Laird 1983; Abney and Johnson 1991; Gibson 1991; Resnik 1992; Stabler 1994; Lewis and Vasishth 2005)
  - Constraints on working memory (Miller 1956; Cowan 2001; McElree 2001; Van Dyke and Johns 2012)

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:
  - Left-corner parsing (Johnson-Laird 1983; Abney and Johnson 1991; Gibson 1991; Resnik 1992; Stabler 1994; Lewis and Vasishth 2005)
  - Constraints on working memory (Miller 1956; Cowan 2001; McElree 2001; Van Dyke and Johns 2012)

Modeling syntax acquisition with unsupervised parsing

- Existing unsupervised parsing systems:
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)
- However, these do not implement:
  - Left-corner parsing (Johnson-Laird 1983; Abney and Johnson 1991; Gibson 1991; Resnik 1992; Stabler 1994; Lewis and Vasishth 2005)
  - Constraints on working memory (Miller 1956; Cowan 2001; McElree 2001; Van Dyke and Johns 2012)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much [illegible] by a human-like learner?)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much syntactic structure is distributionally detectable by a human-like learner?)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much syntactic structure is distributionally [illegible] by a human-like learner?)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much syntactic structure is distributionally detectible by a human-like learner?)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much syntactic structure is distributionally detectible by a human-like learner?)

The UHHMM as a syntax acquisition model

- This work:
  - Unsupervised hierarchical hidden Markov model (UHHMM) parser
    - Left-corner parsing strategy
    - Limited working memory
- Learns from distributional statistics (no world knowledge or reference)
  - Useful for NLP (only textual input needed)
  - Interesting for cognitive modeling (how much syntactic structure is distributionally detectible by a human-like learner?)

The UHHMM as a syntax acquisition model

- We evaluate our learner on a corpus of child-directed input.
- Results beat or closely match those of competing systems.
- **Conclusion:** Much syntactic structure is distributionally detectible.

The UHHMM as a syntax acquisition model

- We evaluate our learner on a corpus of child-directed input.
- Results beat or closely match those of competing systems.
- **Conclusion:** Much syntactic structure is distributionally detectible.

The UHHMM as a syntax acquisition model

- We evaluate our learner on a corpus of child-directed input.
- Results beat or closely match those of competing systems.
- **Conclusion:** Much syntactic structure is distributionally detectible.

Plan

Plan

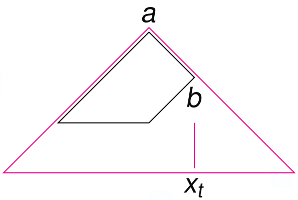
Left-corner parsing

- Maintains a store of derivation fragments $a/b, a'/b', \ldots$, each consisting of active category a lacking awaited category b.
- Incrementally assembles trees by forking/joining fragments.

Left-corner parsing

- Maintains a store of derivation fragments $a/b, a'/b', \ldots,$ each consisting of active category a lacking awaited category b.
- Incrementally assembles trees by forking/joining fragments.

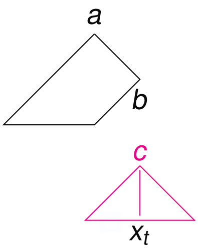
Left-corner parsing: Fork decision



No-fork (shift + match): Word satisfies b. a is complete.

$$\frac{a/b \quad x_t}{a} \; b \to x_t . \tag{–F}$$

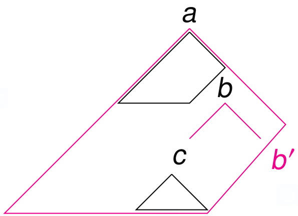
Left-corner parsing: Fork decision



Yes-fork (shift): Word does not satisfy b, fork off new complete category c.

$$\frac{a/b \quad x_t}{a/b \quad c} \; b \xrightarrow{+} c \;\ldots\;; \quad c \rightarrow x_t. \tag{+F}$$

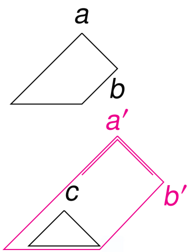
Left-corner parsing: Join decision



Yes-join (predict + match): Complete category c satisfies b while predicting b'. Store updates from $\langle \ldots, a/b, c \rangle$ to $\langle \ldots, a/b' \rangle$.

$$\frac{a/b \quad c}{a/b'} \; b \rightarrow c\ b'. \tag{+J}$$

Left-corner parsing: Join decision



No-join (predict): Complete category c does not satisfy b. Predict new a' and b' from c. Store updates from $\langle \ldots, a/b, c \rangle$ to $\langle \ldots, a/b, a'/b' \rangle$.

$$\frac{a/b \quad c}{a/b \quad a'/b'} \; b \xrightarrow{+} a' \ldots \; ; \quad a' \rightarrow c \; b'. \qquad (\text{–J})$$

Left-corner parsing

- Four possible outcomes:
  - **+F+J:** Yes-fork and yes-join, no change in depth
  - **–F–J:** No-fork and no-join, no change in depth
  - **+F–J:** Yes-fork and no-join, depth increments
  - **–F+J:** No-fork and yes-join, depth decrements

Left-corner parsing

- Four possible outcomes:
  - **+F+J:** Yes-fork and yes-join, no change in depth
  - **–F–J:** No-fork and no-join, no change in depth
  - **+F–J:** Yes-fork and no-join, depth increments
  - **–F+J:** No-fork and yes-join, depth decrements

Left-corner parsing

- Four possible outcomes:
  - **+F+J:** Yes-fork and yes-join, no change in depth
  - **–F–J:** No-fork and no-join, no change in depth
  - **+F–J:** Yes-fork and no-join, depth increments
  - **–F+J:** No-fork and yes-join, depth decrements

Left-corner parsing

- Four possible outcomes:
  - **+F+J:** Yes-fork and yes-join, no change in depth
  - **–F–J:** No-fork and no-join, no change in depth
  - **+F–J:** Yes-fork and no-join, depth increments
  - **–F+J:** No-fork and yes-join, depth decrements

Left-corner parsing

- Four possible outcomes:
  - **+F+J:** Yes-fork and yes-join, no change in depth
  - **–F–J:** No-fork and no-join, no change in depth
  - **+F–J:** Yes-fork and no-join, depth increments
  - **–F+J:** No-fork and yes-join, depth decrements

Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

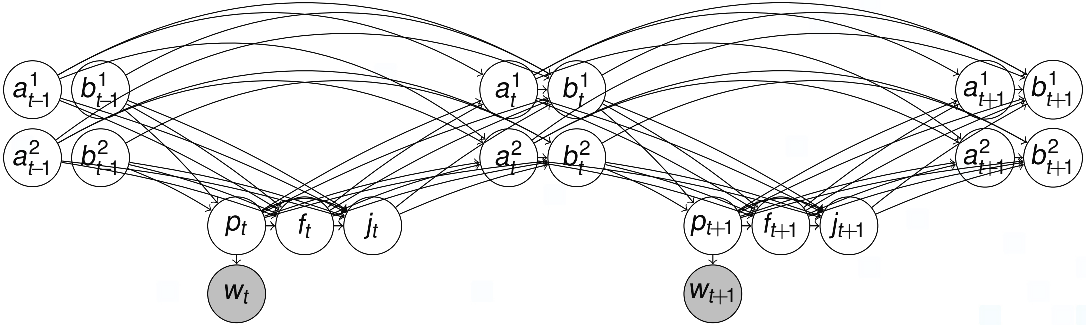
Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

Unsupervised sequence modeling of left-corner parsing

- A left-corner parser can be implemented as an unsupervised probabilistic sequence model using hidden random variables at every time step for:
  - *Active* categories A
  - *Awaited* categories B
  - *Preterminal* or part-of-speech (POS) tags P
  - Binary switching variables F and J
- There is also an observed random variable W over *Words*.

Unsupervised sequence modeling of left-corner parsing



Graphical representation of probabilistic left-corner parsing model across two time steps, with $D = 2$.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Unsupervised sequence modeling of left-corner parsing

- Model trained with batch Gibbs sampling (Beal, Ghahramani, and Rasmussen 2002; Van Gael et al. 2008)
  - Calculate posteriors in a forward pass
  - Sample parse in a backward pass
  - Resample models at each iteration
- Non-parametric (infinite) version described in paper. Parametric learner used in these experiments.
- Parses extracted from a single iteration after convergence.

Plan

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.

    Children have more severe memory limits than adults (Gathercole et al. 2004).

    Greater depth is rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.
    - Children have more severe memory limits than adults (Gathercole 1998).
    - Greater depths rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.
    - Children have more severe memory limits than adults (Gathercole 1998).
    - Greater depths rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.
    - Children have more severe memory limits than adults (Gathercole 1998).
    - Greater depths rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.
    - Children have more severe memory limits than adults (Gathercole 1998).
    - Greater depths rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Experimental setup

- Experimental conditions designed to mimic conditions of early language learning:
  - **Child-directed input:** Child-directed utterances from the Eve corpus of Brown (1973), distributed with CHILDES (MacWhinney 2000).
  - **Limited depth:** Depth was limited to 2.
    - Children have more severe memory limits than adults (Gathercole 1998).
    - Greater depths rarely needed for child-directed utterances.
  - **Small hypothesis space (Newport 1990):** 4 active categories, 4 awaited categories, 8 parts of speech.

Accuracy evaluation methods

- **Gold standard:** Hand-corrected PTB-style trees for Eve (Pearl and Sprouse 2013)
- **Competitors:**
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erk 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)

Accuracy evaluation methods

- **Gold standard:** Hand-corrected PTB-style trees for Eve (Pearl and Sprouse 2013)
- **Competitors:**
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)

Accuracy evaluation methods

- **Gold standard:** Hand-corrected PTB-style trees for Eve (Pearl and Sprouse 2013)
- **Competitors:**
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)

Accuracy evaluation methods

- **Gold standard:** Hand-corrected PTB-style trees for Eve (Pearl and Sprouse 2013)
- **Competitors:**
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)

Accuracy evaluation methods

- **Gold standard:** Hand-corrected PTB-style trees for Eve (Pearl and Sprouse 2013)
- **Competitors:**
  - CCL (Seginer 2007)
  - UPPARSE (Ponvert, Baldridge, and Erik 2011)
  - BMMM+DMV (Christodoulopoulos, Goldwater, and Steedman 2012)

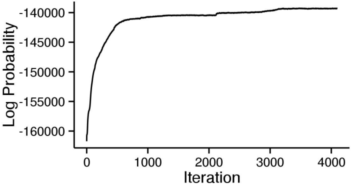
Plan

Results: Comparison to other systems

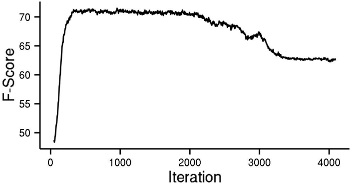
	P	R	F_1
UPPARSE	60.50	51.96	55.90
CCL	64.70	53.47	58.55
BMMM+DMV	63.63	**64.02**	**63.82**
UHHMM	**68.83**	57.18	62.47
Random baseline (UHHMM 1st iter)	51.69	38.75	44.30

Unlabeled bracketing accuracy by system on Eve.

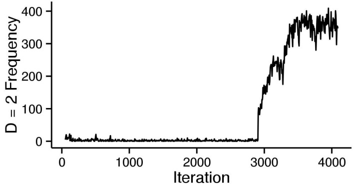
Results: UHHMM timecourse of acquisition



Log probability increases



F-score decreases late



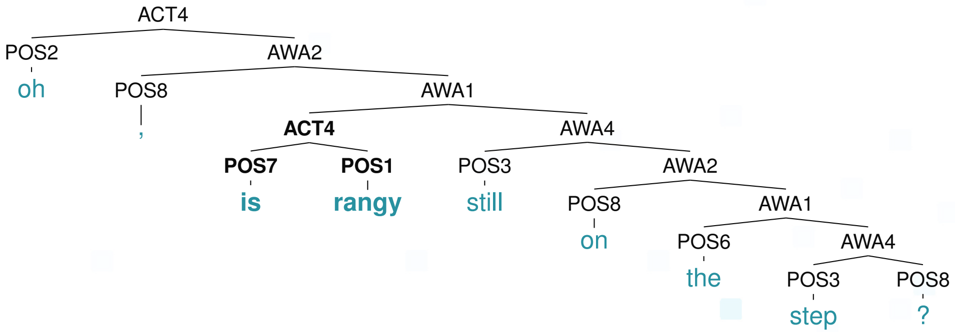
Depth 2 frequency increases late

Results: UHHMM uses of depth 2

- Many uses of depth 2 are linguistically well-motivated.

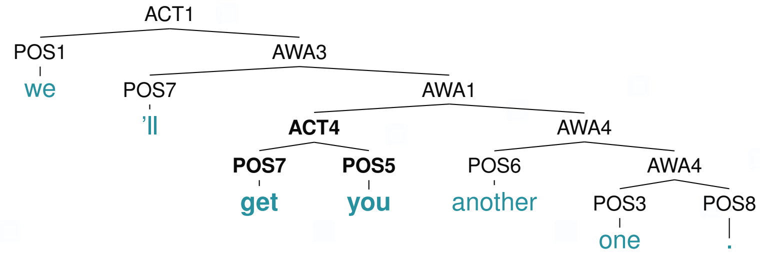
Results: UHHMM uses of depth 2

Subject-auxiliary inversion: (c.f. Chomsky 1968)



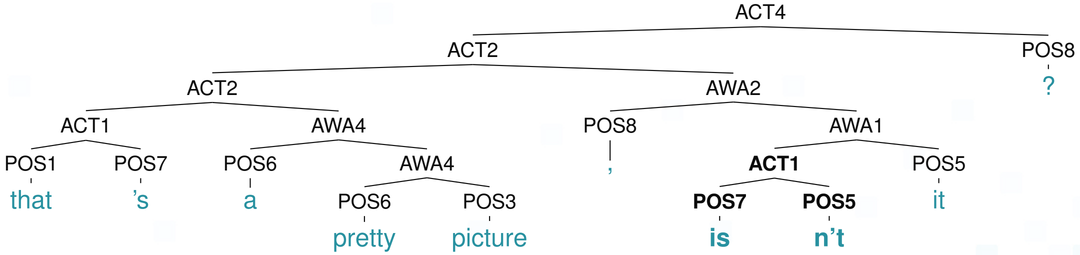
Results: UHHMM uses of depth 2

Ditransitive:



Results: UHHMM uses of depth 2

Contraction:



Results: UHHMM uses of depth 2

- All of these structures have flat representations in gold standard, so these insights are not reflected in our accuracy scores.

Plan

Conclusion

- We presented a new grammar induction system (UHHMM) that
  - Models cognitive constraints on human sentence processing and acquisition
  - Achieves results competitive with SOTA raw-text parsers on child-directed input
- This suggests that distributional information can greatly assist syntax acquisition in a human-like language learner, even without access to other important cues (e.g. world knowledge).

Conclusion

- We presented a new grammar induction system (UHHMM) that
  - Models cognitive constraints on human sentence processing and acquisition
  - Achieves results competitive with SOTA raw-text parsers on child-directed input
- This suggests that distributional information can greatly assist syntax acquisition in a human-like language learner, even without access to other important cues (e.g. world knowledge).

Conclusion

- We presented a new grammar induction system (UHHMM) that
  - Models cognitive constraints on human sentence processing and acquisition
  - Achieves results competitive with SOTA raw-text parsers on child-directed input
- This suggests that distributional information can greatly assist syntax acquisition in a human-like language learner, even without access to other important cues (e.g. world knowledge).

Conclusion

- We presented a new grammar induction system (UHHMM) that
  - Models cognitive constraints on human sentence processing and acquisition
  - Achieves results competitive with SOTA raw-text parsers on child-directed input
- This suggests that distributional information can greatly assist syntax acquisition in a human-like language learner, even without access to other important cues (e.g. world knowledge).

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphology)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphology)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Conclusion

- Future plans:
  - Numerous optimizations to facilitate:
    - Larger state spaces
    - Deeper memory stores
    - Non-parametric learning
  - Adding a joint segmentation component in order to:
    - Model joint lexical and syntactic acquisition
    - Exploit word-internal cues (morphemes)
  - Downstream evaluation (e.g. MT)

Thank you!

Github:

https://github.com/tmills/uhhmm/

Acknowledgments:

The authors would like to thank the anonymous reviewers for their comments. This project was sponsored by the Defense Advanced Research Projects Agency award #HR0011-15-2-0022. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References I

Abney, Steven P. and Mark Johnson (1991). “Memory Requirements and Local Ambiguities of Parsing Strategies”. In: *J. Psycholinguistic Research* 20.3, pp. 233–250.

Beal, Matthew J., Zoubin Ghahramani, and Carl E. Rasmussen (2002). “The Infinite Hidden Markov Model”. In: *Machine Learning*. MIT Press, pp. 29–245.

Brown, R. (1973). *A First Language*. Cambridge, MA: Harvard University Press.

Chomsky, Noam (1968). *Language and Mind*. New York: Harcourt, Brace & World.

Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman (2012). “Turning the pipeline into a loop: Iterated unsupervised dependency parsing and PoS induction”. In: *NAACL-HLT Workshop on the Induction of Linguistic Structure*. Montreal, Canada, pp. 96–99.

Cowan, Nelson (2001). “The magical number 4 in short-term memory: A reconsideration of mental storage capacity”. In: *Behavioral and Brain Sciences* 24, pp. 87–185.

Gathercole, Susan E. (1998). “The development of memory”. In: *Journal of Child Psychology and Psychiatry* 39.1, pp. 3–27.

References II

- Gibson, Edward (1991). "A computational theory of human linguistic processing: Memory limitations and processing breakdown". PhD thesis. Carnegie Mellon.
- Johnson-Laird, Philip N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA, USA: Harvard University Press. ISBN: 0-674-56882-6.
- Lewis, Richard L. and Shravan Vasishth (2005). "An activation-based model of sentence processing as skilled memory retrieval". In: *Cognitive Science* 29.3, pp. 375–419.
- MacWhinney, Brian (2000). *The CHILDES project: Tools for analyzing talk*. Third. Mahwah, NJ: Lawrence Elrbaum Associates.
- McElree, Brian (2001). "Working Memory and Focal Attention". In: *Journal of Experimental Psychology, Learning Memory and Cognition* 27.3, pp. 817–835.
- Miller, George A. (1956). "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information". In: *Psychological Review* 63, pp. 81–97.
- Newport, Elissa (1990). "Maturational constraints on language learning". In: *Cognitive Science* 14, pp. 11–28.

References III

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

Ponvert, Elias, Jason Baldridge, and Katrin Erik (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*. Portland, Oregon, pp. 1077–1086.

Resnik, Philip (1992). “Left-Corner Parsing and Psychological Plausibility”. In: *Proceedings of COLING*. Nantes, France, pp. 191–197.

Seginer, Yoav (2007). “Fast Unsupervised Incremental Parsing”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 384–391.

Stabler, Edward (1994). “The finite connectivity of linguistic structure”. In: *Perspectives on Sentence Processing*. Lawrence Erlbaum, pp. 303–336.

References IV

Van Dyke, Julie A. and Clinton L. Johns (2012). “Memory interference as a determinant of language comprehension”. In: *Language and Linguistics Compass* 6.4, pp. 193–211. ISSN: 15378276. DOI: 10.1016/j.pestbp.2011.02.012.Investigations. arXiv: NIHMS150003.

Van Gael, Jurgen et al. (2008). “Beam sampling for the infinite hidden Markov model”. In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 1088–1095.

Plan

Introduction

Left-corner parsing via unsupervised sequence modeling

Experimental setup

Results

Conclusion

Appendix

Appendix: Joint conditional probability

Variable	Meaning
t	position in the sequence
w_t	observed word at position t
D	depth of the memory store at position t
$q_t^{1..D}$	stack of derivation fragments at t
a_t^d	active category at position t and depth $1 \leq d \leq D$
b_t^d	awaited category at position t and depth $1 \leq d \leq D$
f_t	fork decision at position t
j_t	join decision at position t
θ	state x state transition matrix

Table 1: Variable definitions used in defining model probabilities.

Appendix: Joint conditional probability

$$
\begin{aligned}
\mathrm{P}(q_t^{1..D}\ w_t \mid q_{1..t-1}^{1..D}\ w_{1..t-1}) &= \mathrm{P}(q_t^{1..D}\ w_t \mid q_{t-1}^{1..D}) && (1)\\
&\stackrel{\text{def}}{=} \mathrm{P}(p_t\ w_t\ f_t\ j_t\ a_t^{1..D}\ b_t^{1..D} \mid q_{t-1}^{1..D}) && (2)\\
&= \mathrm{P}_{\theta_P}(p_t \mid q_{t-1}^{1..D}) \cdot \\
&\quad\ \mathrm{P}_{\theta_W}(w_t \mid q_{t-1}^{1..D}\ p_t) \cdot \\
&\quad\ \mathrm{P}_{\theta_F}(f_t \mid q_{t-1}^{1..D}\ p_t\ w_t) \cdot \\
&\quad\ \mathrm{P}_{\theta_J}(j_t \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t) \cdot \\
&\quad\ \mathrm{P}_{\theta_A}(a_t^{1..D} \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t\ j_t) \cdot \\
&\quad\ \mathrm{P}_{\theta_B}(b_t^{1..D} \mid q_{t-1}^{1..D}\ p_t\ w_t\ f_t\ j_t\ a_t^{1..D}) && (3)
\end{aligned}
$$

Appendix: Part-of-speech model

$$\mathsf{P}_{\theta_P}(p_t \mid q_{t-1}^{1..D}) \stackrel{\text{def}}{=} \mathsf{P}_{\theta_P}(p_t \mid d\; b_{t-1}^{d}); \quad d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \tag{4}$$

Appendix: Lexical model

$$\mathrm{P}_{\theta_W}(w_t \mid q_{t-1}^{1..D}\; p_t) \stackrel{\text{def}}{=} \mathrm{P}_{\theta_W}(w_t \mid p_t) \tag{5}$$

Appendix: Fork model

$$\mathsf{P}_{\theta_F}(f_t \mid q^{1..D}_{t-1}\ p_t\ w_t) \stackrel{\text{def}}{=} \mathsf{P}_{\theta_F}(f_t \mid d\ b^{d}_{t-1}\ p_t); \quad d = \max_{d'}\{q^{d'}_{t-1} \neq q_{\bot}\} \tag{6}$$

Appendix: Join model

$$
\mathrm{P}_{\theta_J}(j_t \mid q_{t-1}^{1:D}\ f_t\ p_t\ w_t) \overset{\text{def}}{=} \begin{cases} \mathrm{P}_{\theta_J}(j_t \mid d\ a_{t-1}^{d}\ b_{t-1}^{d-1});\quad d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 0 \\ \mathrm{P}_{\theta_J}(j_t \mid d\ p_t\ b_{t-1}^{d});\quad d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} & \text{if } f_t = 1 \end{cases} \tag{7}
$$

Appendix: Active category model

$$\mathsf{P}_{\theta_A}(a_t^{1..D} \mid q_{t-1}^{1..D}\ f_t\ p_t\ w_t\ j_t) \stackrel{\text{def}}{=}$$
$$\begin{cases} [\![a_t^{1..d-2} = a_{t-1}^{1..d-2}]\!] \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] & \cdot\, [\![a_t^{d+0..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \quad \text{if } f_t = 0, j_t = 1 \\ [\![a_t^{1..d-1} = a_{t-1}^{1..d-1}]\!] \cdot \mathsf{P}_{\theta_A}(a_t^d \mid d\ b_{t-1}^{d-1}\ a_{t-1}^d) & \cdot\, [\![a_t^{d+1..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \quad \text{if } f_t = 0, j_t = 0 \\ [\![a_t^{1..d-1} = a_{t-1}^{1..d-1}]\!] \cdot [\![a_t^{d} = a_{t-1}^{d}]\!] & \cdot\, [\![a_t^{d+1..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \quad \text{if } f_t = 1, j_t = 1 \\ [\![a_t^{1..d-0} = a_{t-1}^{1..d-0}]\!] \cdot \mathsf{P}_{\theta_A}(a_t^{d+1} \mid d\ b_{t-1}^{d}\ p_t) & \cdot\, [\![a_t^{d+2..D} = a_\perp]\!];\ \ d = \max_{d'}\{q_{t-1}^{d'} \neq q_\perp\} \quad \text{if } f_t = 1, j_t = 0 \end{cases} \tag{8}$$

Appendix: Awaited category model

$$P_{\theta_B}(b_t^{1..D} \mid q_{t-1}^{1..D}\ f_t\ p_t\ w_t\ j_t\ a_t^{1..D}) \stackrel{\text{def}}{=}$$

$$\begin{cases} [\![b_t^{1..d-2} = b_{t-1}^{1..d-2}]\!] \cdot P_{\theta_B}(b_t^{d-1} \mid d\ b_{t-1}^{d-1}\ a_{t-1}^{d}) \cdot [\![b_t^{d+0..D} = b_\bot]\!]; \quad d = \max_{d'} \{ q_{t-1}^{d'} \neq q_\bot \} & \text{if } f_t = 0, j_t = 1 \\ [\![b_t^{1..d-1} = b_{t-1}^{1..d-1}]\!] \cdot P_{\theta_B}(b_t^{d} \mid d\ a_t^{d}\ a_{t-1}^{d}) \cdot [\![b_t^{d+1..D} = b_\bot]\!]; \quad d = \max_{d'} \{ q_{t-1}^{d'} \neq q_\bot \} & \text{if } f_t = 0, j_t = 0 \\ [\![b_t^{1..d-1} = b_{t-1}^{1..d-1}]\!] \cdot P_{\theta_B}(b_t^{d} \mid d\ b_{t-1}^{d}\ p_t) \cdot [\![b_t^{d+1..D} = b_\bot]\!]; \quad d = \max_{d'} \{ q_{t-1}^{d'} \neq q_\bot \} & \text{if } f_t = 1, j_t = 1 \\ [\![b_t^{1..d-0} = b_{t-1}^{1..d-0}]\!] \cdot P_{\theta_B}(b_t^{d+1} \mid d\ a_t^{d+1}\ p_t) \cdot [\![b_t^{d+2..D} = b_\bot]\!]; \quad d = \max_{d'} \{ q_{t-1}^{d'} \neq q_\bot \} & \text{if } f_t = 1, j_t = 0 \end{cases} \tag{9}$$

Appendix: Graphical model

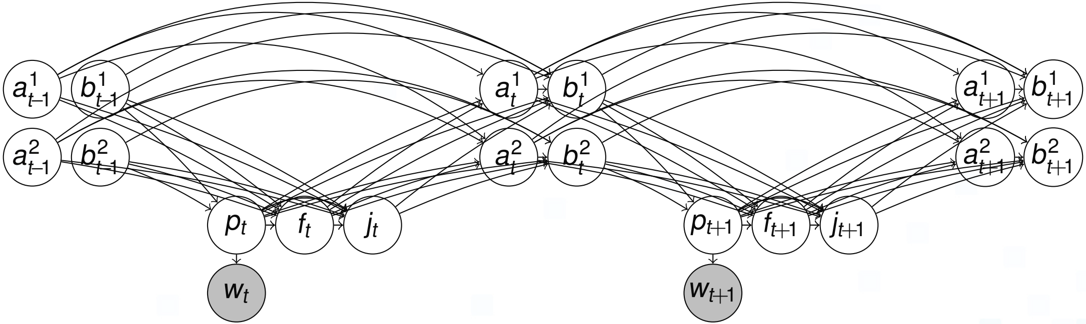


Figure 1: Graphical representation of probabilistic left-corner parsing model expressed in Equations 6–9 across two time steps, with $D = 2$.

Appendix: Punctuation

- Punctuation poses a problem — keep or remove?
  - **Remove:** Doesn't exist in input to human learners.
  - **Keep:** Might be proxy for intonational phrasal cues.
- Punctuation was kept in training data in main result presented above.
- We did an additional UHHMM run trained on data with punctuation removed (2000 iterations).

Appendix: Punctuation

- Punctuation poses a problem — keep or remove?
  - **Remove:** Doesn’t exist in input to human learners.
  - **Keep:** Might be proxy for intonational phrasal cues.
- Punctuation was kept in training data in main result presented above.
- We did an additional UHHMM run trained on data with punctuation removed (2000 iterations).

Appendix: Punctuation

- Punctuation poses a problem — keep or remove?
  - **Remove:** Doesn't exist in input to human learners.
  - **Keep:** Might be proxy for intonational phrasal cues.
- Punctuation was kept in training data in main result presented above.
- We did an additional UHHMM run trained on data with punctuation removed (2000 iterations).

Appendix: Punctuation

- Punctuation poses a problem — keep or remove?
  - **Remove:** Doesn't exist in input to human learners.
  - **Keep:** Might be proxy for intonational phrasal cues.
- Punctuation was kept in training data in main result presented above.
- We did an additional UHHMM run trained on data with punctuation removed (2000 iterations).

Appendix: Punctuation

- Punctuation poses a problem — keep or remove?
  - **Remove:** Doesn't exist in input to human learners.
  - **Keep:** Might be proxy for intonational phrasal cues.
- Punctuation was kept in training data in main result presented above.
- We did an additional UHHMM run trained on data with punctuation removed (2000 iterations).

Appendix: Results (without punctuation)

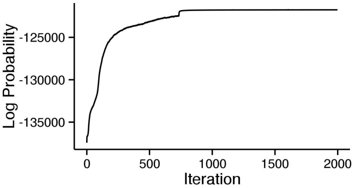


Figure 2: Log Probability (no punc)

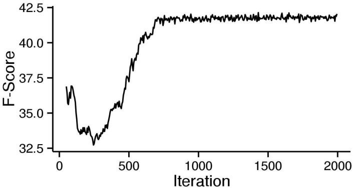


Figure 3: F-Score (no punc)

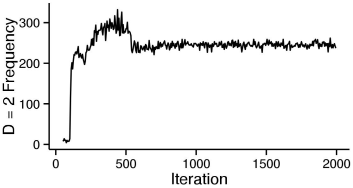


Figure 4: Depth=2 Frequency (no punc)

Appendix: Comparison by system (with and without punctuation)

	With punc			No punc		
	P	R	F1	P	R	F1
UPPARSE	60.50	51.96	55.90	38.17	48.38	42.67
CCL	64.70	53.47	58.55	56.87	47.69	51.88
BMMM+DMV (directed)	62.08	62.51	62.30	61.01	59.24	60.14
BMMM+DMV (undirected)	63.63	**64.02**	**63.82**	61.34	**59.33**	**60.32**
UHHMM-4000, binary	46.68	58.28	51.84	37.62	46.97	41.78
UHHMM-4000, flattened	**68.83**	57.18	62.47	**61.78**	45.52	52.42
Right-branching	68.73	**85.81**	**76.33**	**68.73**	**85.81**	**76.33**

Table 2: Parsing accuracy by system on Eve with and without punctuation (phrasal cues) in the input.