Learning and the typology of word order:

A model of the Final-over-Final condition

Shay Hucklebridge

University of Massachusetts Amherst

NELS 51, November 6th 2020

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.
- Some syntactic patterns may be more 'learnable' than their alternatives,

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.
- Some syntactic patterns may be more 'learnable' than their alternatives,
  - It may be difficult to acquire a stable grammar for harder patterns given the limited time and data a human learner has.

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.
- Some syntactic patterns may be more 'learnable' than their alternatives,
  - It may be difficult to acquire a stable grammar for harder patterns given the limited time and data a human learner has.
  - This may encourage languages to shift away from harder patterns, causing them to be cross-linguistically rare.

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.
- Some syntactic patterns may be more 'learnable' than their alternatives,
  - It may be difficult to acquire a stable grammar for harder patterns given the limited time and data a human learner has.
  - This may encourage languages to shift away from harder patterns, causing them to be cross-linguistically rare.
- The focus here is on the typology of word order

Learnability and typology

- This talk uses computational modeling to investigate the potential influence of learnability on typology.
- Some syntactic patterns may be more 'learnable' than their alternatives,
  - It may be difficult to acquire a stable grammar for harder patterns given the limited time and data a human learner has.
  - This may encourage languages to shift away from harder patterns, causing them to be cross-linguistically rare.
- The focus here is on the typology of word order
  - Particular focus on the learnability of languages containing structures that violate the **Final-over-Final Condition (FOFC)** (Holmberg, 2000; Biberauer et al., 2014; Sheehan et al., 2017)

The Final-Over-Final Condition

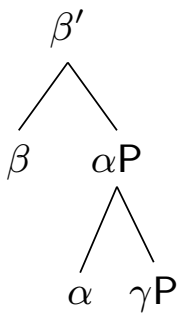
The FOFC:

A head-final phrase may not immediately dominate a head-initial phrase if both are in the same extended projection. (Biberauer et al., 2014)

The Final-Over-Final Condition

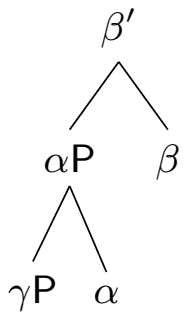
- From Biberauer et al. (2014, 171):

(1) a. **Head-initial** (harmonic)

i. 

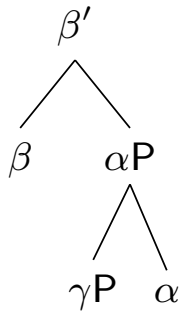
ii. Example: *Aux Verb Object*

b. **Head-final** (harmonic)

i. 

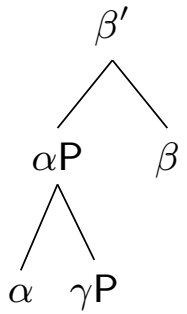
ii. Example: *Object Verb Aux*

c. **Initial-over-final** (disharmonic)

i. 

ii. Example: *Aux Object Verb*

d. **Final-over-initial** (*FOFC)

i. 

ii. Example: **Verb Object Aux*

The learning task

- Learning task conducted here using an Expectation-Driven Learner (Jarosz, 2015; Nazarov and Jarosz, 2017)

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

The learning task

- Learning task conducted here using an Expectation-Driven Learner (Jarosz, 2015; Nazarov and Jarosz, 2017)
  - The EDL is a domain-general statistical learner for parameter systems.

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

The learning task

- Learning task conducted here using an Expectation-Driven Learner (Jarosz, 2015; Nazarov and Jarosz, 2017)
  - The EDL is a domain-general statistical learner for parameter systems.
  - When the EDL is presented with a training token, it samples a setting for each parameter and compares the output of that sample language to the target token.
  - If this results in a match, parameter settings responsible for the match are rewarded. If a parameter setting contributes to a mismatch, it is penalized.
  - Blame is assigned proportionately to each parameter setting's contribution to the match/mismatch (computed using Bayes' rule).

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

The learning task

- The learner was trained on languages consisting of ordered {*Auxiliary, Verb, Object*}, {*Verb, Object*}, and {*Auxiliary, Object*} tokens[1]

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

The learning task

- The learner was trained on languages consisting of ordered {*Auxiliary, Verb, Object*}, {*Verb, Object*}, and {*Auxiliary, Object*} tokens[1]
- The learner was only exposed to strings, and had no access to information about each datum's syntactic structure

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

The learning task

- The learner was trained on languages consisting of ordered {*Auxiliary, Verb, Object*}, {*Verb, Object*}, and {*Auxiliary, Object*} tokens[1]
- The learner was only exposed to strings, and had no access to information about each datum's syntactic structure
- No explicit bias for or against a particular word order was built into the parameter system, or the learner. All four word orders were included in the typology.

[1] The training data did not include any languages with variable word order – this is left as a potential further extension of this project

Syntactic assumptions

- Data used in the learning tasks was generated by a 4-parameter system that conformed to the following assumptions:

Syntactic assumptions

- Data used in the learning tasks was generated by a 4-parameter system that conformed to the following assumptions:
- **Headedness:**
  - Harmonic structures may be underlyingly head-final or head-initial (contra Kayne (1994))
  - Disharmonic structures must be derived through movement.

Syntactic assumptions

- Data used in the learning tasks was generated by a 4-parameter system that conformed to the following assumptions:
- **Headedness:**
  - Harmonic structures may be underlyingly head-final or head-initial (contra Kayne (1994))
  - Disharmonic structures must be derived through movement.
- **Movement:**
  - No rightward movement (or rightward specifiers)
  - Movement obeys anti-locality (Grohmann, 2003, 2011; Abels, 2003), so the complement of a head may not move into its specifier.

Parameters

- The training data was generated by a simplified four-parameter system that operated over the hidden structure in 2:

  (2) Unordered structure:

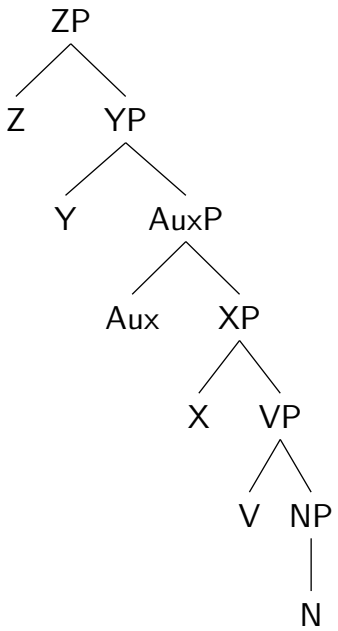
  $\{_{ZP}$ Z, $\{_{YP}$ Y, $\{_{AuxP}$ Aux, $\{_{XP}$ X, $\{_{VP}$ V, $\{_{NP}$ N$\}\}\}\}\}$

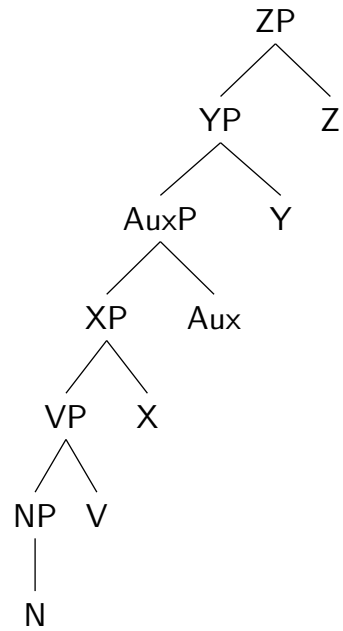
Parameters

- **Parameter 1 (P1)**

  $\textsc{Headedness}\ (\mathrm{VP})$: When set to 0, all heads in the extended projection of VP are linearized to the left. When set to 1, heads are linearized to the right.

(3) a. $\textsc{Headedness}(\mathrm{VP}) = 0$ b. $\textsc{Headedness}(\mathrm{VP}) = 1$



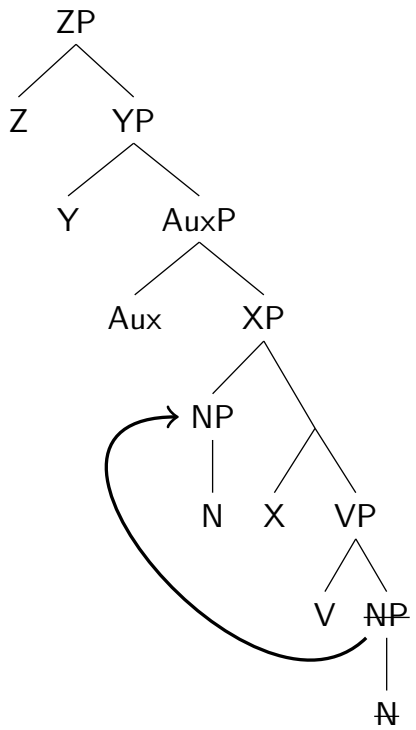


Parameters

- **Parameter 2 (P2)**

  NP-SpecXP: When set to 0, no movement occurs. When set to 1, NP moves to the specifier of XP.

(4) NP-SpecXP = 1

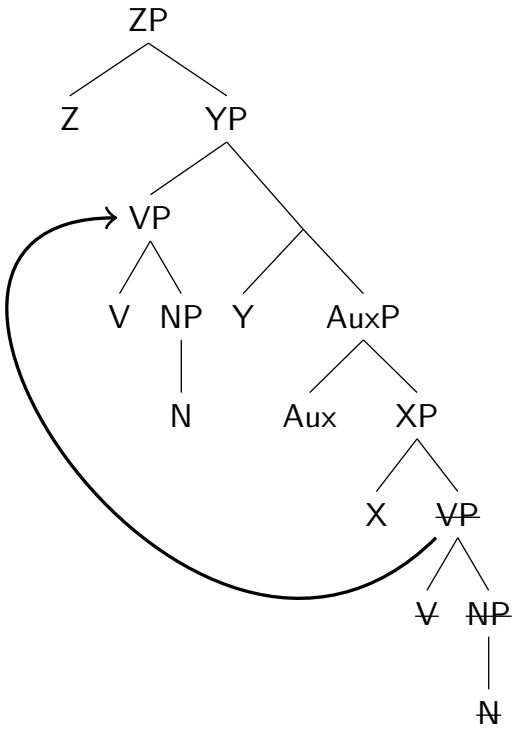


Parameters

- **Parameter 3 (P3)**

  VP-SpecYP: When set to 0, no movement occurs. When set to 1, VP moves to the specifier of YP.

(5) VP-SpecYP = 1

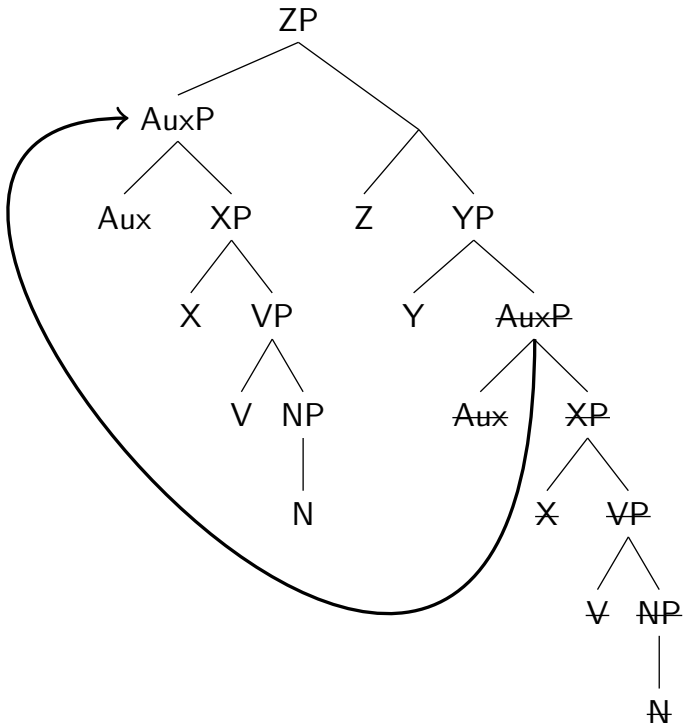


Parameters

- **Parameter 4 (P4)**

  AuxP-SpecZP: When set to 0, no movement occurs. When set to 1, AuxP moves to the specifier of ZP.

(6) AuxP-SpecZP = 1



The typology

- These parameters generate a set of 16 languages.
- Languages are named by their parameters.
- A head-final language with no movement is called '1000' after its parameter settings:

(7) **Harmonic head-final language 1000:**

HEADEDNESS (P1):	1
NP-SPECXP (P2):	0
VP-SPECYP (P3):	0
AUXP-SPECZP (P4):	0

The typology

- The full set of languages predicted by these four parameters, sorted by their surface word order pattern, is given in table 1:

Table 1: Word order patterns and their languages

	Head initial	**Head Final**	**Initial/final**	**Final/initial**	HI+F	HF+F
	0011, 0001 0000	1101, 1100, 1010 1001, 1000	1011, 0111 0101, 0100	1110, 0010	0110	1111
{Aux{O}} {V{O}} {Aux {V {O}}}	Aux-O V-O Aux-V-O	O-Aux O-V O-V-Aux	Aux-O O-V Aux-O-V	O-Aux V-O V-O-Aux	Aux-O V-O V-Aux-O	O-Aux O-V O-Aux-V

The typology

- The full set of languages predicted by these four parameters, sorted by their surface word order pattern, is given in table 1:

Table 1: Word order patterns and their languages

	Head initial	**Head Final**	**Initial/final**	**Final/initial**	HI+F	HF+F
	0011, 0001 0000	1101, 1100, 1010 1001, 1000	1011, 0111 0101, 0100	1110, 0010	0110	1111
{Aux{O}} {V{O}} {Aux {V {O}}}	Aux-O V-O Aux-V-O	O-Aux O-V O-V-Aux	Aux-O O-V Aux-O-V	O-Aux V-O V-O-Aux	Aux-O V-O V-Aux-O	O-Aux O-V O-Aux-V

- Bolded languages are the word orders from (1).

The typology

- The full set of languages predicted by these four parameters, sorted by their surface word order pattern, is given in table 1:

Table 1: Word order patterns and their languages

	Head initial	**Head Final**	**Initial/final**	**Final/initial**	HI+F	HF+F
	0011, 0001 0000	1101, 1100, 1010 1001, 1000	1011, 0111 0101, 0100	1110, 0010	0110	1111
{Aux{O}} {V{O}} {Aux {V {O}}}	Aux-O V-O Aux-V-O	O-Aux O-V O-V-Aux	Aux-O O-V Aux-O-V	O-Aux V-O V-O-Aux	Aux-O V-O V-Aux-O	O-Aux O-V O-Aux-V

- Bolded languages are the word orders from (1).
- Note that these parameters generate two additional languages: Head-initial with VP fronting (HI+F), and Head-final with AuxP fronting (HF+F). I will not discuss these in detail here, as they do not impact the results.

Results

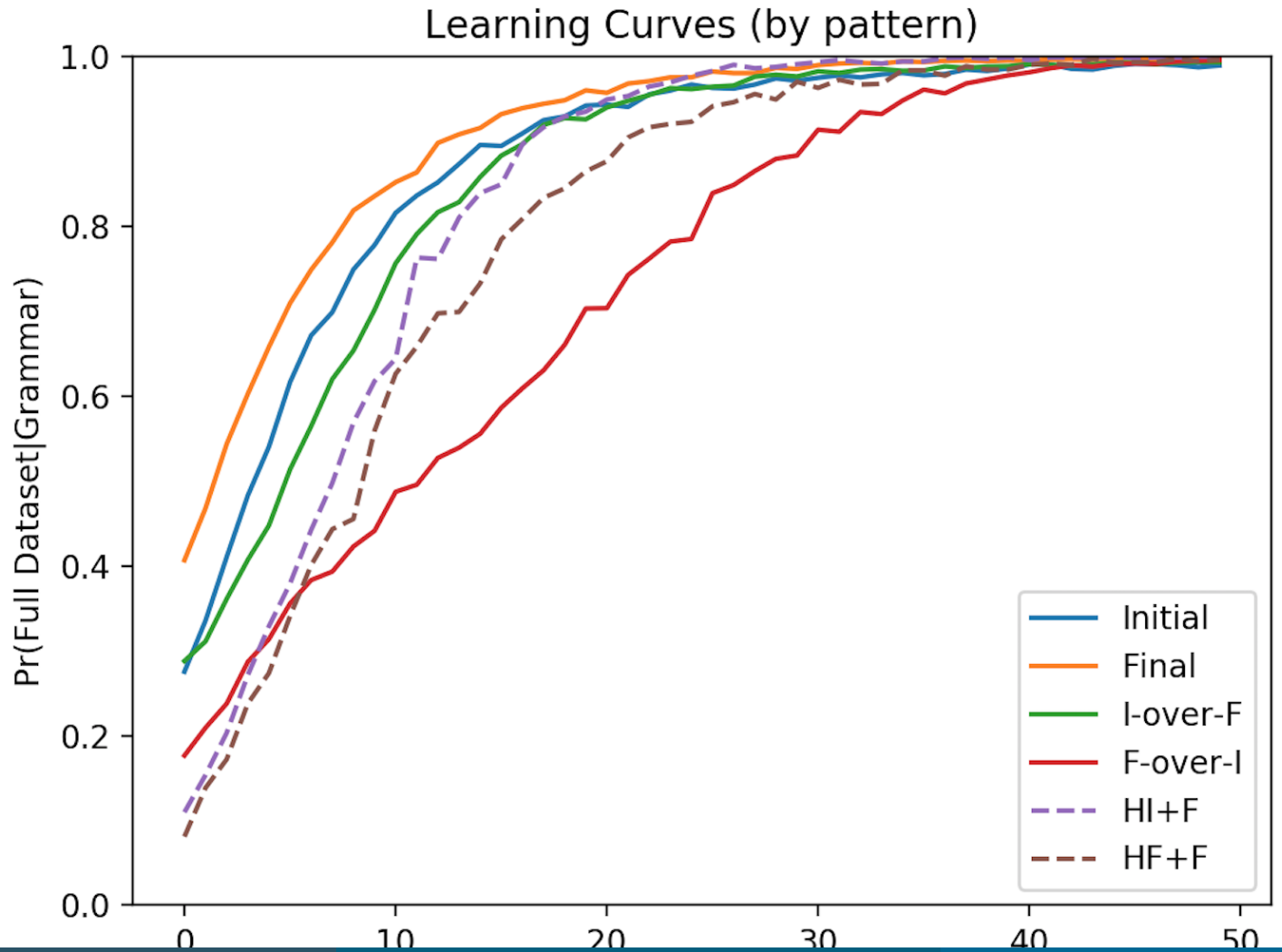
- Since nothing in the data distinguished between weakly-equivalent languages, the learner acquired one parameter setting per word-order pattern.

Results

- Since nothing in the data distinguished between weakly-equivalent languages, the learner acquired one parameter setting per word-order pattern.
- EDL reached over 95% accuracy on all word-order patterns by 50 passes through the data (no pattern was unlearnable).
    - This was using a learning rate of 0.1. Averages were taken across 40 reps.

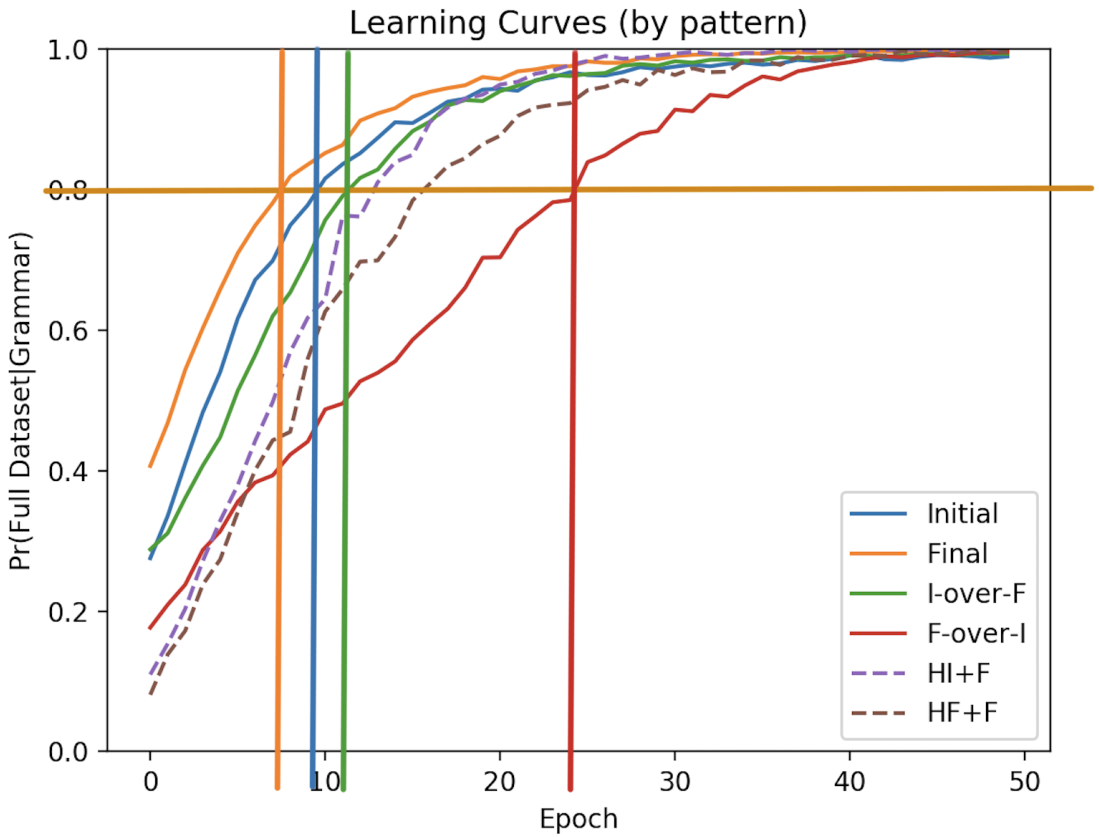
Results

- Learning curved by word order pattern averaged over 50 passes through the data:



Results

- We can see the relative learnability clearly by looking at (approximately) how long it takes the learner to reach 80% accuracy on a pattern:



Head-final: 80% by **7** passes through the data

Head-initial: 80% by **9** passes through the data

Initial/final: 80% by **11** passes through the data

Final/initial: 80% by **20+** passes through the data

Head-initial+VP fronting and Head-final+AuxP fronting took about 13 and 15 passes through the data, respectively, to reach 80%.

Discussion

- Why are *FOFC languages so hard?

[2] Individual parameter curves for each word-order pattern can be seen in the appendix.

Discussion

- Why are *FOFC languages so hard?
  - All four parameters had to be correctly set for the learner to achieve accuracy.

[2] Individual parameter curves for each word-order pattern can be seen in the appendix.

Discussion

- Why are *FOFC languages so hard?
  - All four parameters had to be correctly set for the learner to achieve accuracy.
    - The harmonic languages and initial-over-final languages only required at max. 3 parameters to be set before the setting of the remaining parameter ceased to matter.[2]

[2] Individual parameter curves for each word-order pattern can be seen in the appendix.

Discussion

- Why are *FOFC languages so hard?
  - All four parameters had to be correctly set for the learner to achieve accuracy.
    - The harmonic languages and initial-over-final languages only required at max. 3 parameters to be set before the setting of the remaining parameter ceased to matter.[2]
  - There were only two possible final-over-initial languages; 1110 and 0010.

[2] Individual parameter curves for each word-order pattern can be seen in the appendix.

Discussion

- Why are *FOFC languages so hard?
  - All four parameters had to be correctly set for the learner to achieve accuracy.
    - The harmonic languages and initial-over-final languages only required at max. 3 parameters to be set before the setting of the remaining parameter ceased to matter.[2]
  - There were only two possible final-over-initial languages; 1110 and 0010.
    - The learner was pulled in two different directions for parameters 1 & 2, with no way to decide between settings of 1 or 0

[2] Individual parameter curves for each word-order pattern can be seen in the appendix.

Discussion

- Why is the final-over-initial pattern more difficult than the other disharmonic patterns?

Discussion

- Why is the final-over-initial pattern more difficult than the other disharmonic patterns?
- The challenge presented to learning by final-over-initial patterns is reducible to the asymmetry between leftward and rightward movement.

Discussion

- Why is the final-over-initial pattern more difficult than the other disharmonic patterns?
- The challenge presented to learning by final-over-initial patterns is reducible to the asymmetry between leftward and rightward movement.
    - String-vacuous leftward movements create a number of weakly-equivalent initial-over-final languages

Discussion

- Why is the final-over-initial pattern more difficult than the other disharmonic patterns?
- The challenge presented to learning by final-over-initial patterns is reducible to the asymmetry between leftward and rightward movement.
  - String-vacuous leftward movements create a number of weakly-equivalent initial-over-final languages
  - This smooths out the parameter space and guides the learner to a single setting.

Discussion

- Why is the final-over-initial pattern more difficult than the other disharmonic patterns?
- The challenge presented to learning by final-over-initial patterns is reducible to the asymmetry between leftward and rightward movement.
  - String-vacuous leftward movements create a number of weakly-equivalent initial-over-final languages
  - This smooths out the parameter space and guides the learner to a single setting.
  - Leftward movement does not create weakly-equivalent languages for the final-over-initial pattern, and so the learner gets stuck between 1110 and 0010.

Take-aways

- **What this result demonstrates:**
  - Under simple and familiar assumptions about a syntax, the word orders that violate the FOFC are learnable (by EDL), but at a significant delay.
- **What it means:**
  - Learnability may be the source of the rarity of FOFC violations – languages may shift away from these difficult-to-learn patterns.
  - It may also account for apparent exceptions to the FOFC (e.g. in Bhatt and Dayal (2007); Erlewine (2017); Sheehan et al. (2017)) as *FOFC languages were never completely unlearnable.

References

Abels, K. (2003). *Successive cyclicity, anti-locality, and adposition stranding*. PhD thesis, University of Connecticut Storrs, CT.

Bhatt, R. and Dayal, V. (2007). Rightward scrambling as rightward remnant movement. *Linguistic Inquiry*, 38(2):287–301.

Biberauer, T., Holmberg, A., and Roberts, I. (2014). A syntactic universal and its consequences. *Linguistic Inquiry*, 45(2):169–225.

Erlewine, M. Y. (2017). Low sentence-final particles in Mandarin Chinese and the Final-over-Final Constraint. *Journal of East Asian Linguistics*, 26(1):37–75.

Grohmann, K. K. (2003). Successive cyclicity under (anti-) local considerations. *Syntax*, 6(3):260–312.

Grohmann, K. K. (2011). Anti-locality: Too-close relations in grammar. *The Oxford handbook of linguistic minimalism*, pages 260–290.

Holmberg, A. (2000). Deriving OV order in Finnish. *The Derivation of VO and OV*, pages 123–152.

Jarosz, G. (2015). Expectation driven learning of phonology. *Ms., University of Massachusetts Amherst*.

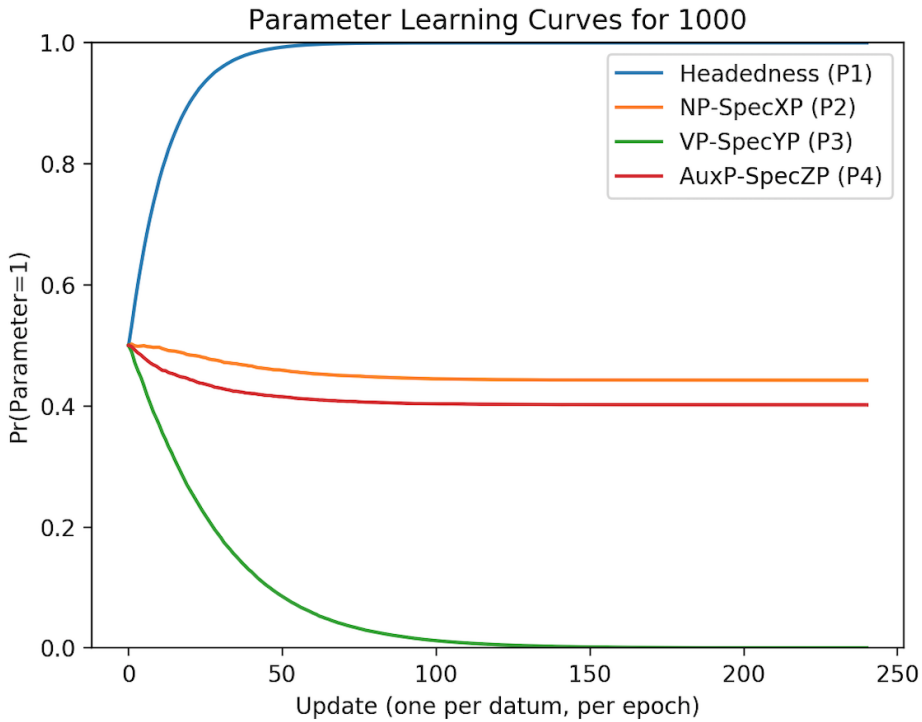
Kayne, R. S. (1994). *The antisymmetry of syntax*. Number 25. Mit Press.

Nazarov, A. and Jarosz, G. (2017). Learning parametric stress without domain-specific mechanisms. In *Proceedings of the annual meetings on phonology*, volume 4.

Sheehan, M., Biberauer, T., Roberts, I., and Holmberg, A. (2017). *The Final-over-Final Condition: A syntactic universal*, volume 76. MIT Press.

Appendix:

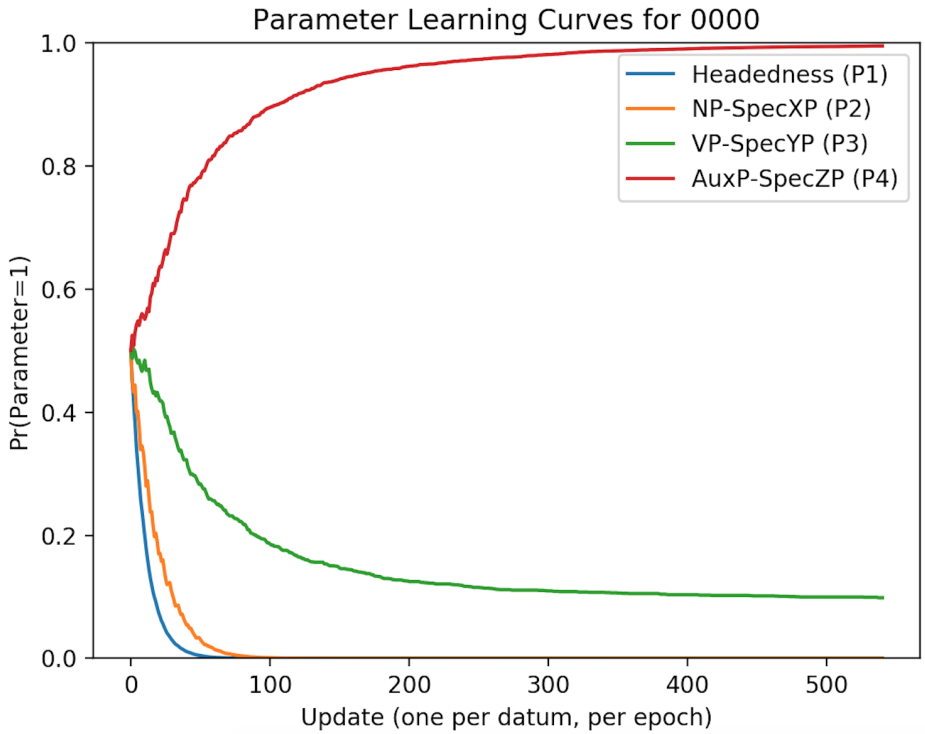
- Learning curves for individual parameters for the head-final harmonic languages (80 passes through the data):



- Once P1 (headedness) and P3 (VP-SpecYP) are correctly set, the settings of P2 & P4 no longer matter, as the learner as already achieved 100% accuracy, and so P2 & P4 plateau (aren't set).

Appendix:

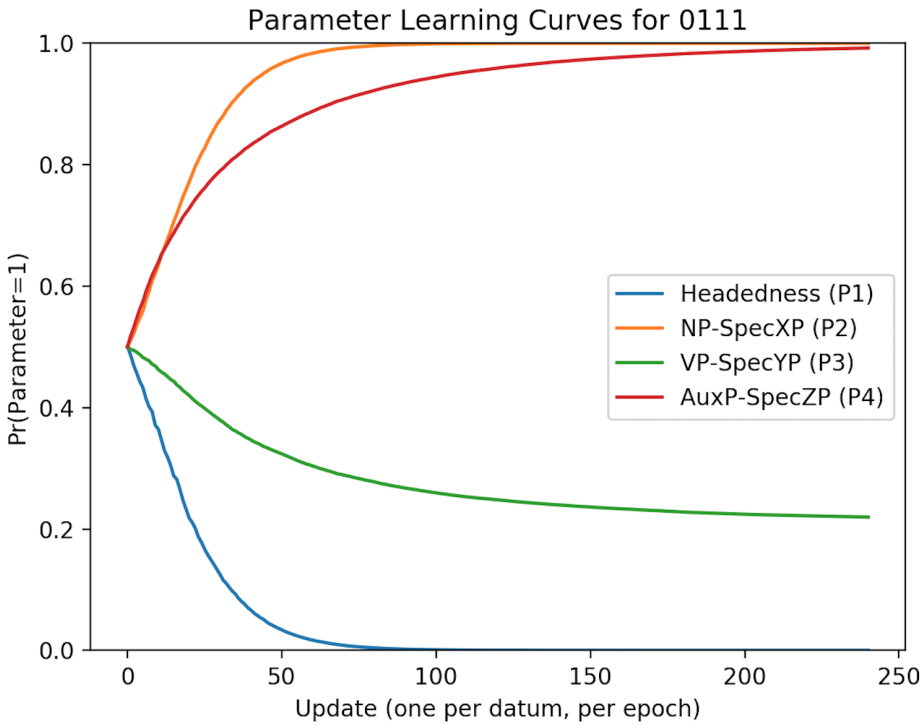
- Learning curves for individual parameters for the head-initial harmonic languages:



- P1 & P2 are set quickly. The learner can then set either P3 or P4, and the remaining setting won't matter. It achieves high accuracy quickly, but takes more updates than other patterns to reach 100% (around 160 passes through the data).

Appendix

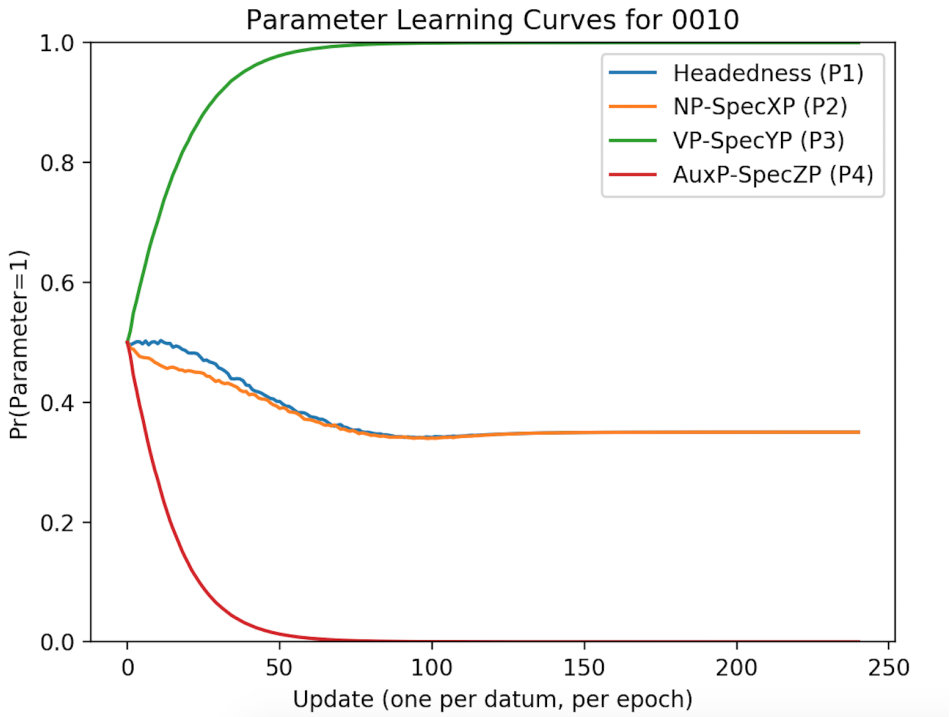
- Learning curves for individual parameters for the initial-over-final disharmonic language (80 passes through the data:



- P1, P2, and P4 must be correctly set. Once this happens, P3 plateaus, as either setting will produce the correct output.

Appendix

- Learning curves for individual parameters for the final-over-initial (*FOFC) disharmonic languages (80 passes through the data):



- P3 & P4 are set quickly, and then the learner cannot consistently decide how to set P1 & P2.

Appendix

- The plateau in the final-over-initial table is not indicative of either setting being equally good, but of the learner picking different settings on different runs. This is evident when looking at individual runs instead of the average: