How the Constraint Space Structure Enables Learning

Jane Chandlee, Rémi Eyraud, Jeffrey Heinz, Adam Jardine, Jonathan Rawski
Haverford College, Aix-Marseille Université, Stony Brook University, Rutgers University

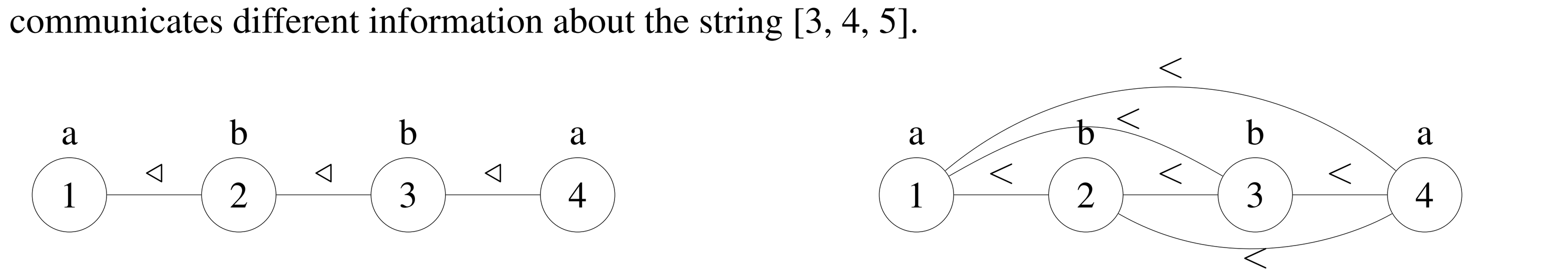
Overview

How is the hypothesis space of a phonological learner structured, and how does the learner use this structure to generalize from examples? Recent work on model-theoretic phonology shows that particular phonological representations and relations play a large role in learning properties of well-formed structures. Here we:

- Extend current model-theoretic accounts of phonology to unconventional string models incorporating featural information;
- show how features structure the hypothesis space into ideals and filters;
- DDescribe a non-statistical, non-enumerative learning algorithm that provably learns the most general constraints over features consistent with the data.
- Its efficiency and integration with statistical models is focus of current research.

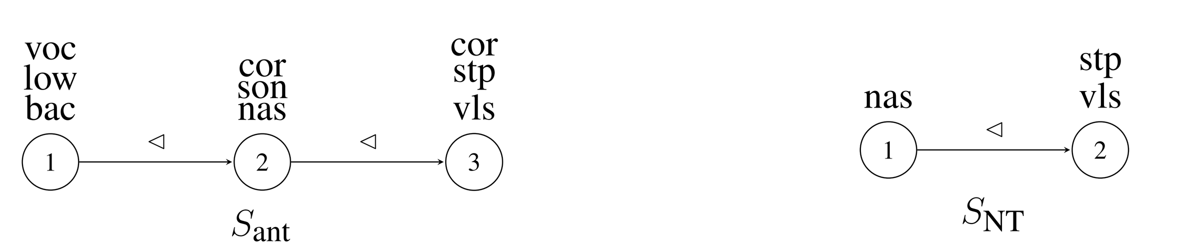
Model-Theoretic Phonological Representations

A model theory for words defines a class of *relational structures*, each of which can be interpreted as a possible word, or as "part" of some possible word [1, 2]. Below are two representations of the string *abba*, with the immediate successor relation ($\triangleleft$), or with the general precedence relation ($<$). Each communicates different information about the string [3, 4, 5].



Unconventional String Models [6, 7]

We can augment word models to include featural information by allowing domain elements to have multiple labeling relations. This allowes substrictures of those models to represent featural constraints.



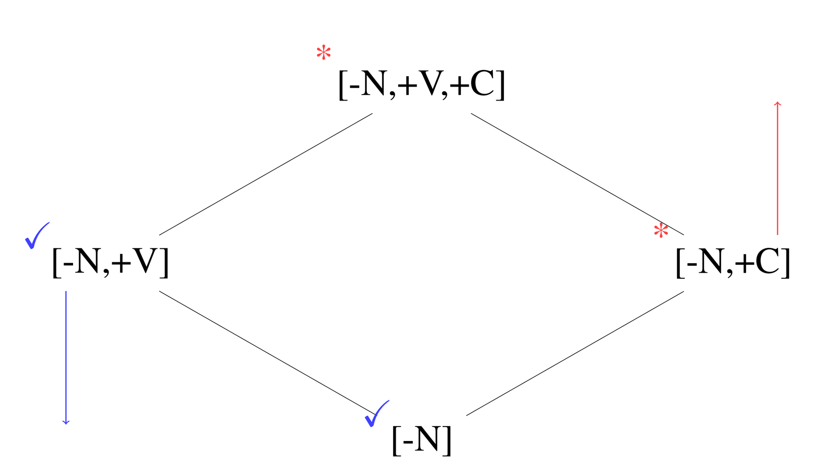
PROBLEM: Features are a central phonological generalization, but they exponentially increase the number of hypotheses learners can entertain when selecting constraint grammars. [8].

Feature Ideals and Grammatical Entailments

Let S and T be segments represented as bundles of n-ary features. Then T is an **feature extension** of S for grammar G ($S <_G T$) iff T is the result of inserting one or more n-ary features of G in S.

Feature Ideals: If T is a feature extension of S for G and G generates T, then G generates S.

Feature Filters: If T is a feature extension of S for G and G forbids S, then G forbids T.



Organizing the hypothesis space into sets of ideals and principal filters allows the learner to exploit these grammatical entailments they provide.

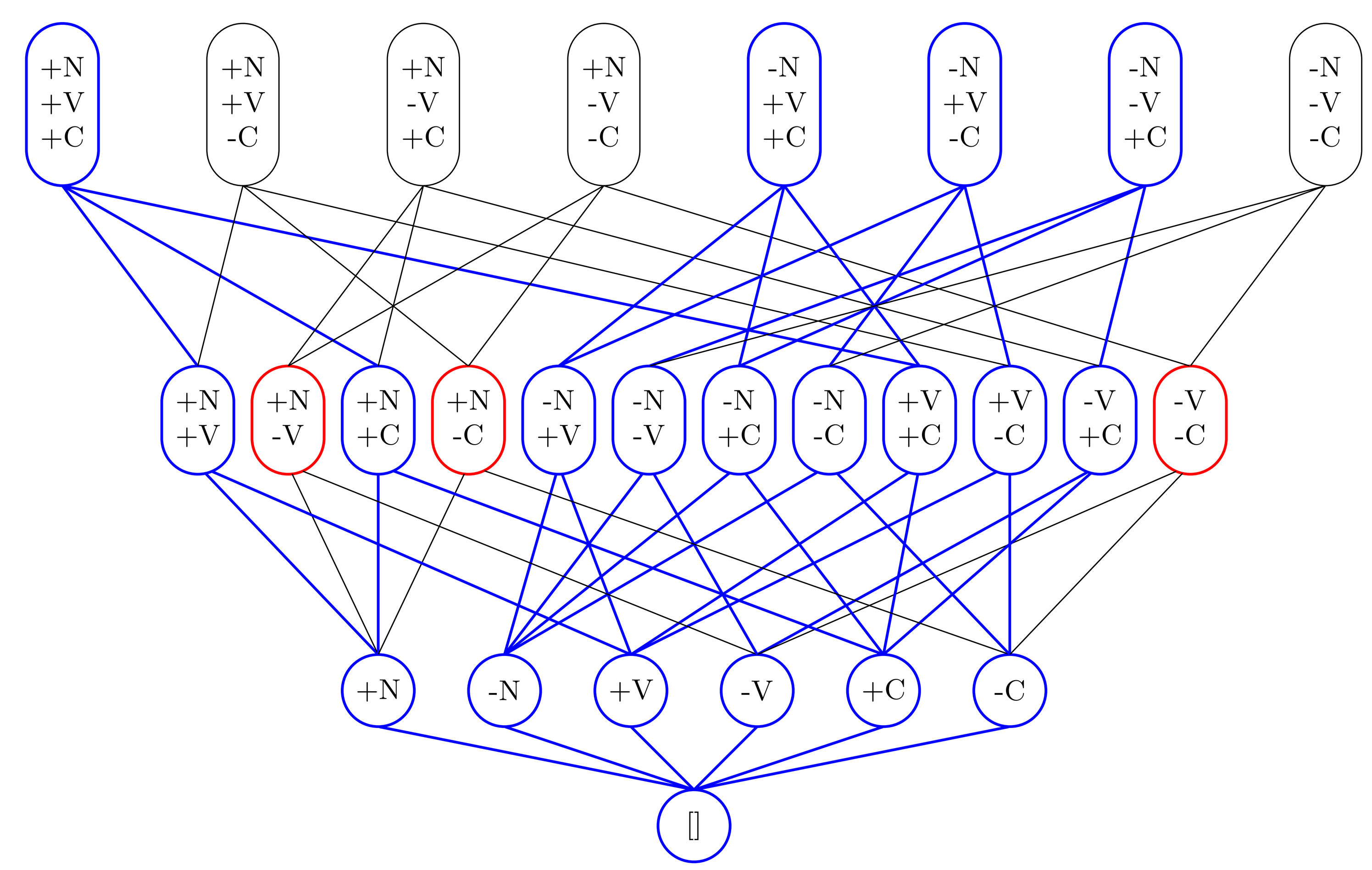
Example: Banning Singular Segments

Suppose the learning data consists of:

- [+N,+V,+C] (voiced nasal consonants),
- [-N,+V,+C] (voiced nonnasal consonants),
- [-N,-V,+C] (voiceless nonnasal consonants)
- [-N,+V,-C] (voiced nonnasal vowels),

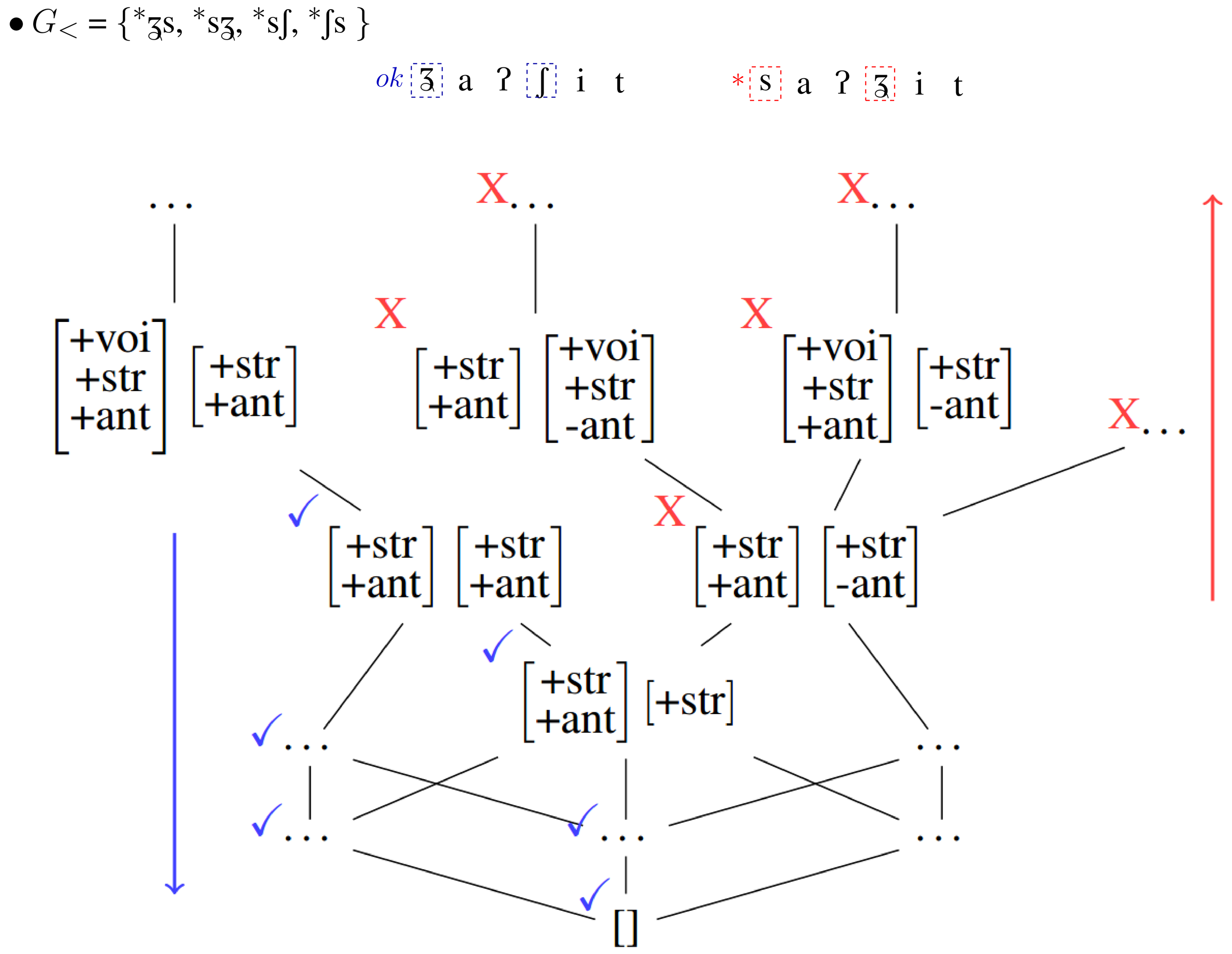
What constraints ought to be posited?

Positing *[+N,-V] (voiceless nasals), *[+N,-C] (nasal vowels), and *[-V,-C] (voiceless vowels) accounts for the absence of the four unobserved feature combinations with fewer constraints.



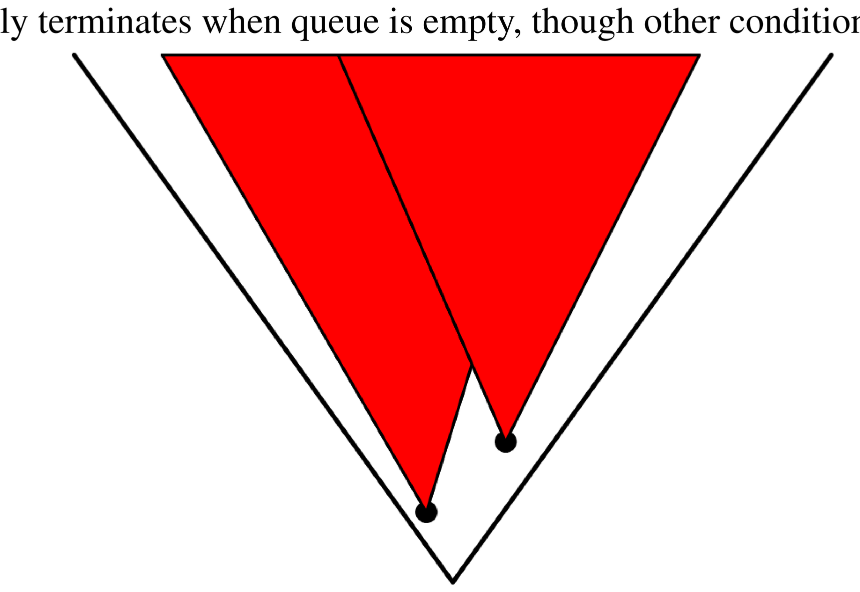
Example: Aari Long distance sibilant harmony

- In Aari, all sibilants agree in anteriority.
  - (1) baʔse 'he brought'
  - (2) ʒaʔʃit 'I arrived'
- $G_<$ = {*ʒs, *sʒ, *sʃ, *ʃs}



Bottom-Up Learning Algorithm

- Data is batch input
- Traverses the constraint space bottom-up, considering constraints from most general to most specific, beginning with the empy constraint
- For each structure S (black dots), checks whether S is present in the data
- If not, S is added to the constraint grammar, and more specific variants are skipped (red cones are filters)
- If so, algorithm adds to the queue of constraints to consider next constraints which are one degree more specific (adds one feature or one domain element) and do not belong to any red cones.
- Algorithm naturally terminates when queue is empty, though other conditions may be provided.



Learning Guarantees [9]

Given a finite positive data sample, the bottom-up learner finds a constraint grammar G such that:

1. the largest forbidden substructure is of size k
2. G is consistent, i.e. it covers the data:
   - $D \subseteq L(G)$
3. $L(G)$ is the smallest language in $\mathcal{L}$ which covers the data
   - for all $L \in \mathcal{L}$ where $D \subseteq L$, $L(G) \subseteq L$
4. G includes structures S that are restrictions of structures S' included in other grammars G' that also satisfy (1,2,3)
   - for all $S' \in G'$, there exists $S \in G$ such that $S \sqsubseteq S'$.

Statistics and Structure

- Structured Hypothesis Spaces allow for correct generalization
- Hayes and Wilson are right to have a generality relation in their MaxEnt Learner, but why not use the ordering it gives?
- What is the efficiency tradeoff between statistics and structure?
- Is there a constraint learner which can allow this structure?

References

[1] Herbert B. Enderton. *A Mathematical Introduction to Logic*. Academic Press, 2nd edition, 2001.

[2] Leonid Libkin. *Elements of Finite Model Theory*. Springer, 2004.

[3] Jeffrey Heinz. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623–661, 2010.

[4] Jeffrey Heinz. String extension learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 897–906, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[5] James Rogers, Jeffrey Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert, and Sean Wibel. Cognitive and sub-regular complexity. In Glyn Morrill and Mark-Jan Nederhof, editors, *Formal Grammar*, volume 8036 of *Lecture Notes in Computer Science*, pages 90–108. Springer, 2013.

[6] Kristina Strother-Garcia, Jeffrey Heinz, and Hyun Jin Hwangbo. Using model theory for grammatical inference: a case study from phonology. In Sicco Verwer, Menno van Zaanen, and Rick Smetsers, editors, *Proceedings of The 13th International Conference on Grammatical Inference*, volume 57 of *JMLR: Workshop and Conference Proceedings*, pages 66–78, October 2016.

[7] Mai Ha Vu, Ashkan Zehfroosh, Kristina Strother-Garcia, Michael Sebok, Jeffrey Heinz, and Herbert G. Tanner. Statistical relational learning with unconventional string models. *Frontiers in Robotics and AI*, 5(76):1–26, July 2018.

[8] Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440, 2008.

[9] Luc De Raedt. *Logical and Relational Learning*. Springer-Verlag Berlin Heidelberg, 2008.

[10] Jeffrey Heinz, Anna Kasprzik, and Timo Kötzing. Learning with lattice-structured hypothesis spaces. *Theoretical Computer Science*, 457:111–127, October 2012.