Evaluating a Phonotactic Learner for MITSL$^2_2$ Languages

Jacob Johnson Aniello De Santo

MeLo Lab, Department of Linguistics, University of Utah ✉: jacob.k.johnson@utah.edu ✉: aniello.desanto@utah.edu

Key Findings

- MITSL2IA (De Santo & Aksënova, 2021) is succesful on several subregular patterns, despite low data
- Aksënova (2020)'s evaluation pipeline is valuable for testing subregular learning algorithms
- Transparent learners could be used to inspect the quality of the data in samples available to learners

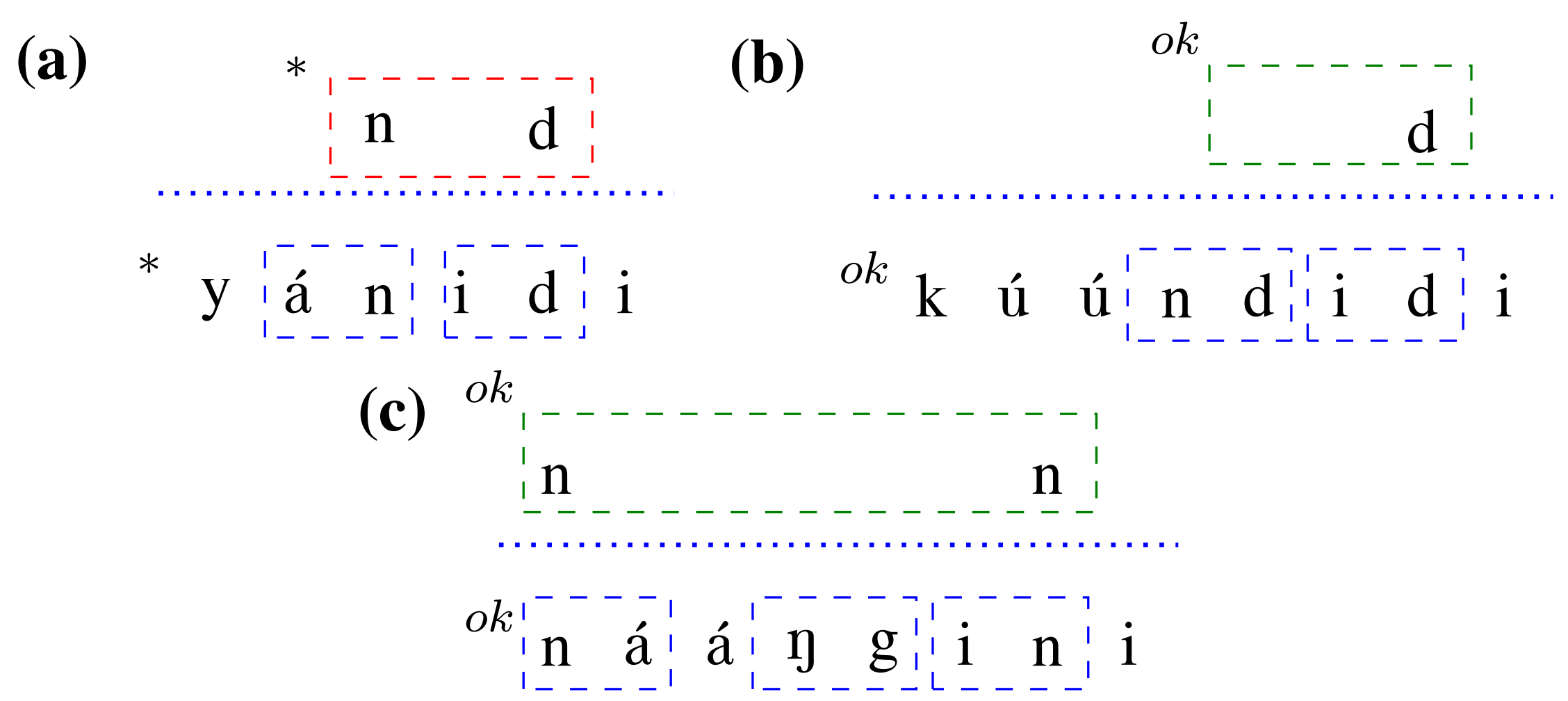


Figure: ITSL$^2_2$ analysis of Yaka nasal harmony from De Santo & Aksënova (2021), illustrating a 2-local projection and 2-local tier constraints. (a) is ill-formed because of tier-adjacent *[nd], but [n,d,g,N] are projected on the tier only when not in a nasal-stop cluster in the input (cf. (b), (c)).; data from Walker (2000)

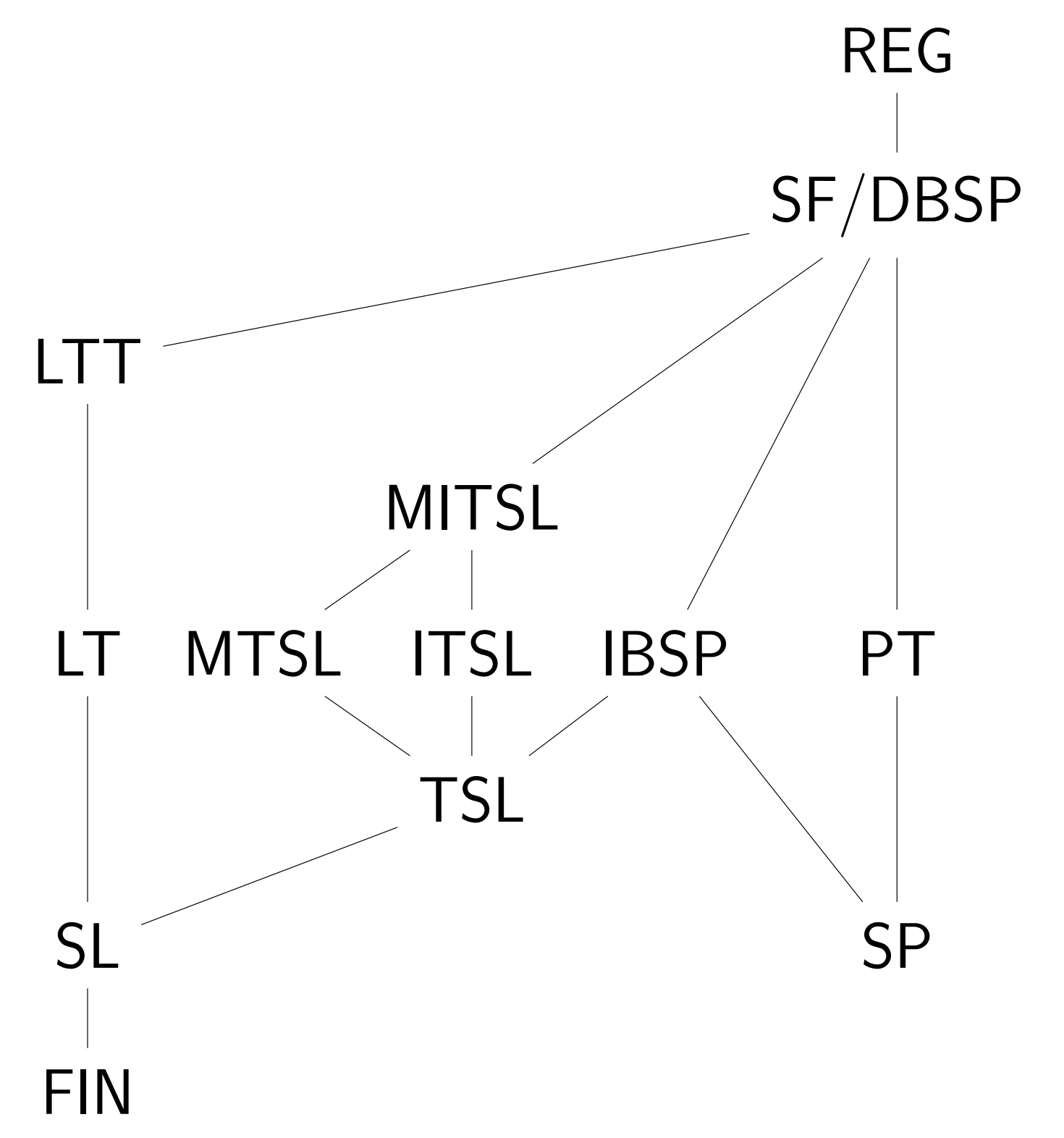


Figure: Subsumption of subregular classes, with the TSL extensions as of De Santo & Graf (2019).

Get the code!

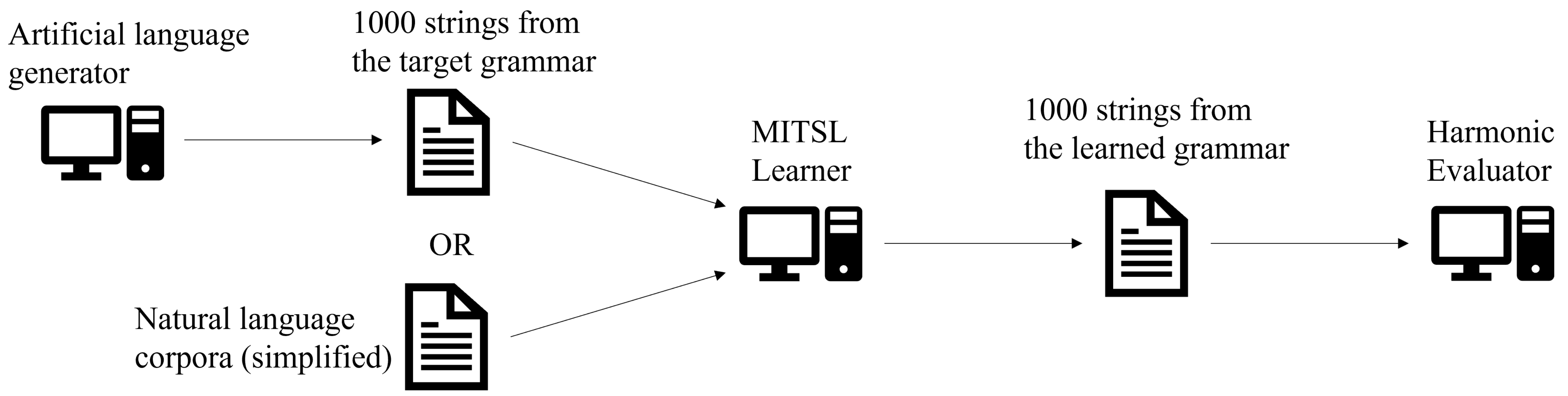
Introduction

- Formal Language Theory provides insights into properties underlying typologically attested patterns (Heinz, 2018)
- MITSL handles **multiple** interactions of **local and non-local phonotactic** constraints (De Santo & Graf, 2019)
- MITSL$^2_2$ can be learned efficiently from positive data (De Santo & Aksënova, 2021)
- **Here:** An implementation and extensive evaluation of (De Santo & Aksënova, 2021)

Learning MITSL$^2_2$ Patterns

- De Santo & Aksënova extend McMullin et al. (2019)'s MTSL$_2$ algorithm to MITSL$^2_2$
- MITSL2IA builds on the intuition that if a bigram $\rho_1\rho_2$ is banned on some tier, then it will never appear in string-adjacent contexts
- We can determine which segments are freely distributed with respect to a bigram $\rho_1\rho_2$ which is not attested and thus assumed to be banned on some tier
- MITSL2IA is guaranteed to learn target grammars efficiently if the input sample is characteristic, but does this align with how phenomena of interest are represented in naturalistic data sets?
- Our implementation of MITSL2IA is available at `https://github.com/jacobkj314/MITSL2IA`
- By inspecting the output, it is possible to infer whether/why the input data was insufficient for the learner to converge on the target grammar in some cases

The Evaluation Pipeline (Aksënova, 2020)


Artificial language generator

1000 strings from the target grammar

OR

Natural language corpora (simplified)

MITSL Learner

1000 strings from the learned grammar

Harmonic Evaluator

- We implemented MITSL2IA in Python 3 following requirements of SigmaPie
- We evaluated it on artificial and simplified natural language phenomena in different subregular classes
- Artificial datasets contained 1000 randomly sampled strings, and up to 130K words for the simplified natural language corpora
- The learned grammars were then given to string generators, and we computed the proportion of strings in the newly generated sample that were well-formed according to the target grammar
- We defined an injection procedure process to explain strings generated by the learned grammar that were not accepted by the target grammar and form new strings to augment the input sample
- Re-running the learner on the data augmented with the "missing" samples resulted in a 100% performance in all cases

References

Aksënova, A. (2020). Tool-assisted induction of subregular languages and mappings (Doctoral dissertation, State University of New York at Stony Brook). · De Santo, A., Aksënova, A. (2021, February). Learning Interactions of Local and Non-Local Phonotactic Constraints from Positive Input. In Proceedings of the Society for Computation in Linguistics 2021 (pp. 167-176). · De Santo, A., Graf, T. (2019). Structure sensitive tier projection: Applications and formal properties. In Formal Grammar: 24th International Conference, FG 2019, Riga, Latvia, August 11, 2019, Proceedings 24 (pp. 35-50). Springer Berlin Heidelberg. · Heinz, J.: The computational nature of phonological generalizations. In: Hyman, L., Plank, F. (eds.) Phonological Typology, chap. 5, pp. 126–195. Phonetics and Phonology, Mouton De Gruyter (2018) · McMullin, K., Aksënova, A., De Santo, A. (2019). Learning phonotactic restrictions on multiple tiers. Proceedings of the Society for Computation in Linguistics, 2(1), 377-378. · Walker, R. (2000, September). Yaka nasal harmony: Spreading or segmental correspondence?. In Annual meeting of the berkeley linguistics society (Vol. 26, No. 1, pp. 321-332).

Acknowledgement. This project was supported by a Research Incentive Seed Grant from the University of Utah awarded to Aniello De Santo (Project #: 51900472).

-

	Aksënova (2020)				This Paper
	SP	SL	TSL	MTSL	MITSL
Word-final devoicing					
T	✗	✓	✓	✓	✓
A	68%	100%	100%	100%	100%
N_G	58%	100%	100%	100%	100%
Single vowel harmony without blocking					
T	✓	✗	✓	✓	✓
A	100%	83%	100%	100%	100%
N_F	100%	72%	100%	100%	100%
Single vowel harmony with blocking					
T	✗	✗	✓	✓	✓
A	84%	89%	100%	100%	99%
Several vowel harmonies without blocking					
T	✓	✗	✓	✓	✓
A	100%	69%	100%	100%	100%
Several vowel harmonies with blocking					
T	✗	✗	✓	✓	✓
A	76%	59%	100%	100%	99%
N_T	76%	70%	67%	95%	99%
Vowel harmony and consonant harmony without blocking					
T	✓	✗	✗	✓	✓
A	100%	64%	74%	100%	100%
Vowel harmony and consonant harmony with blocking					
T	✗	✗	✗	✓	✓
A	83%	64%	69%	100%	100%
Unbounded tone plateauing					
T	✓	✗	✗	✗	✓
A	100%	85%	90%		100%
Two locally-driven long-distance assimilations (ITSL restrictions)					
T	✗	✗	✗	✗	✓
A					100%

Table: (T)heoretical expectations and performance of 5 subregular learners on (A)rtificial and simplified (N)atural language input data-sets. MITSL corresponds to this work. N_G: German; N_F: Finnish; N_T: Turkish.