Learning phonotactics of any span and distance

Polynomial-time learner for MTSL$_k$, $k \geq 2$, with optional overlapping tiers

Ignas Rudaitis, Vilnius University

Vilnius University

Performance on Quechua-derived data

Laryngeal restriction (TSL$_2$)

(Gallagher, 2010)

Tier: aspirates ∪ ejectives

H

Restriction schema: **HH*

Examples: *kintu* 'a bunch', *k'inti* 'a pair', *k*ʰ*astuj* 'to chew'

Counterexamples: **k'int'i*, **k'int*ʰ*i*, **k*ʰ*ast'uj*

∩

Distribution of mid vowels (TSL$_3$)

(Gallagher, 2016)

Tier: [+cons, -uvular] ∪ mid vowels ∪ high vowels

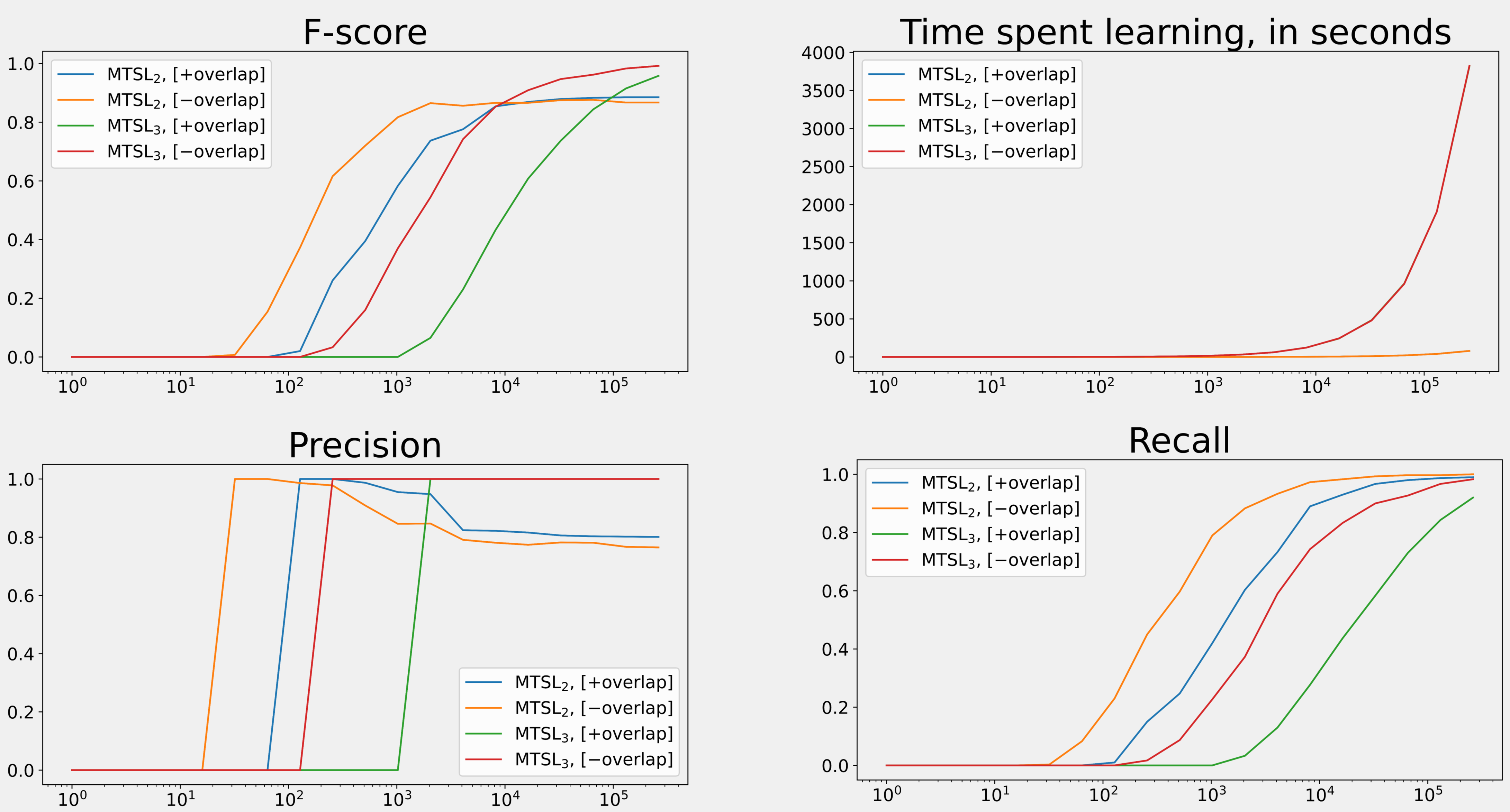
Q̄ *E* *I*

Restriction schema: **Q̄EQ̄*

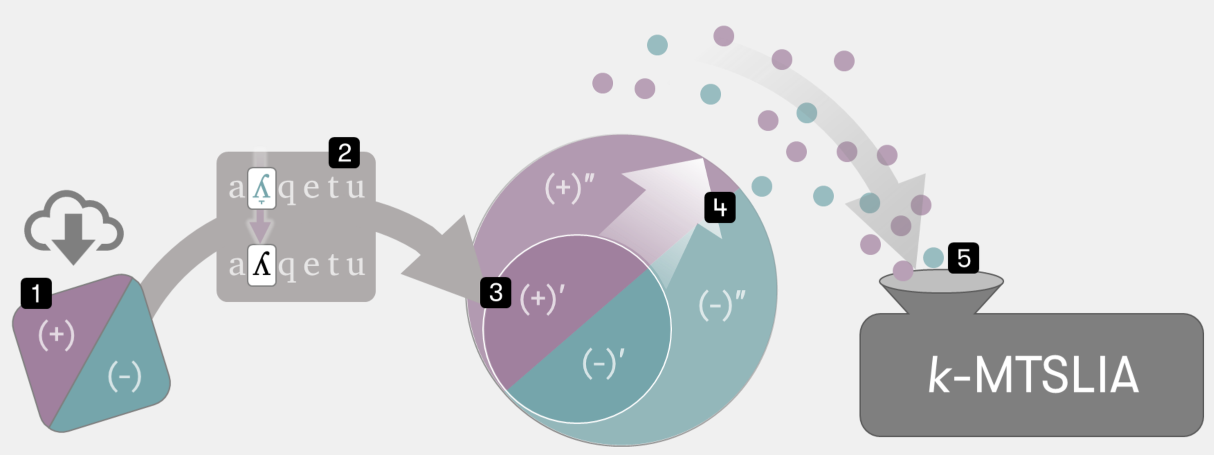
Examples: *erqe* 'son', *q*ʰ*eʎu* 'lazy', *p'esqo* 'bird'

Counterexamples: **erpe*, **t*ʰ*eʎu*, **p'esko*

Outcomes



Deriving the learning samples

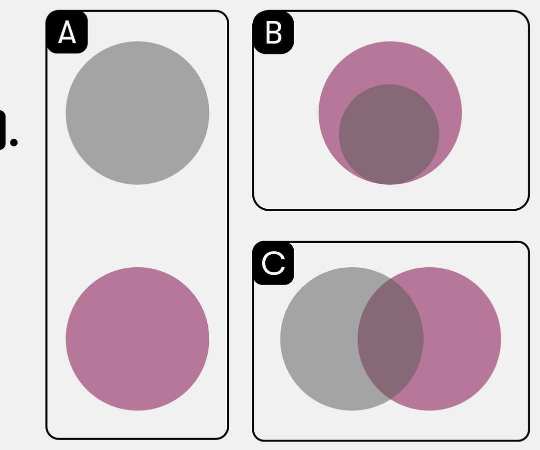


Legend

1. We obtain the Quechua data included with Gouskova & Gallagher (2020)'s code.
2. We eliminate pre- and post-uvular allophones of all consonants, so that the mid vowels pattern with uvulars in a long-distance manner, like they do in Gallagher (2016)'s description.
3. The original dataset is too small to achieve convergence.
4. We expand it by sampling from its trigram model and reinforcing the restrictions in question.
5. We feed the resulting samples into *k*-MTSLIA.

MTSL and overlapping tiers

- The MTSL class of stringsets is the intersection closure of TSL.
- Aksënova & Deshmukh (2018) consider a subclass of MTSL, where the tier alphabets of its consituent TSL stringsets do not overlap, as in **A**, **B**, but not ***C**.
- We will use the name {A, B}-MTSL to refer to this restricted subclass of MTSL.
- Aksënova & Deshmukh (2018) predicted that only {A, B}-MTSL would be relevant for phonotactics.
- So far, the distinction between {A, B}-MTSL and {A, B, C}-MTSL has received little attention in the literature.
- The only learning algorithm for MTSL to date, namely McMullin et al. (2019), can only learn {A, B}-MTSL.
- Gleim & Schneider (2023) provide counterexamples to the {A, B} hypothesis, necessitating new algorithms.
- Our learner, *k*-MTSLIA, is capable of identifying {A, B, C}-MTSL stringsets.



A novel treatment of 2-paths

- Some previous work on learning TSL$_2$, MTSL$_2$, and MITSL$^2_2$ relies on the notion of 2-paths (Jardine & Heinz, 2016; McMullin et al. 2021; De Santo & Aksënova, 2021).
- 2-paths (e.g. x c d y) are essentially bigrams (x y) that are separated by intervening elements ({c d}).
- The set of the interveners must be disjoint with the bigram itself: x y y is **not** a 2-path.
- When learning MTSL$_2$, what does it mean to witness a 2-path in the input data?

2-path in data	Reading	Explanation
u v	Impossible to restrict **uv* without contradicting the data.	*uv* will project regardless of the tier.
x b y	To restrict **xy* on a tier, we must also have *b* on that tier.	*xby* will project as *xy* otherwise.
x c d y	To restrict **xy* on a tier, we must also have *c* or *d* on that tier.	*xcdy* will project as *xy* otherwise.

- 2-paths are **propositions** about the possible ways to restrict the bigrams (x y).
- We conjoin these propositions into one large CNF formula, which becomes the MTSL grammar.
- The size of the formula grows linearly with the size of the data, even when the number of relevant tiers explodes.
- Also, under this treatment, we trivially generalize to ***k*-paths**, $k \geq 2$.

Summary

- Our learner successfully learned the intersection of a TSL$_2$ and a TSL$_3$ pattern.
- However, this required an unnaturally large amount of learning data.
- The learner also lifts the assumption of non-overlapping tiers, made relevant by Gleim & Schneider (2023)'s data.

Code

Abstract

References

Aksënova, A., & **Deshmukh**, S. (2018). Formal restrictions on multiple tiers. In *Proceedings of SCiL 2018* (pp. 64-73). • **De Santo**, A., & **Aksënova**, A. (2021). Learning Interactions of Local and Non-Local Phonotactic Constraints from Positive Input. In *Proceedings of SCiL 2021* (pp. 167-176). • **Gallagher**, G. E. S. (2010). *The perceptual basis of long-distance laryngeal restrictions* (PhD thesis, MIT). • **Gallagher**, G. (2016). Vowel height allophony and dorsal place contrasts in Cochabamba Quechua. *Phonetica*, *73*(2), 101-119. • **Gouskova**, M., & **Gallagher**, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, *38*, 77-116. • **Gleim**, D., & **Schneider**, J. (2023). Phonological processes with intersecting tier alphabets. *Proceedings of SCiL*, *6*(1), 243-249. • **Jardine**, A., & **Heinz**, J. (2016). Learning tier-based strictly 2-local languages. *TACL* 4, 87-98. • **McMullin**, K., **Aksënova**, A., & **De Santo**, A. (2019). Learning phonotactic restrictions on multiple tiers. *Proceedings of SCiL*, *2*(1), 377-378.