A challenge for tier-based strict locality from Uyghur backness harmony

Connor Mayer and Travis Major

UCLA

August 11, 2018

Overview

Past work has hypothesized that all phonological stringsets can be generated by **tier-based strictly local (TSL) grammars**.

The standard analysis of backness harmony in Uyghur is **not TSL**.

Either TSL is not sufficient for phonological stringsets, or another analysis of Uyghur must be adopted.

Background

Phonology studies the systematic organization of sounds in languages.

Phonotactics studies restrictions on how sounds may be combined in a given language.

i.e. for a given language, what is the set of possible words?

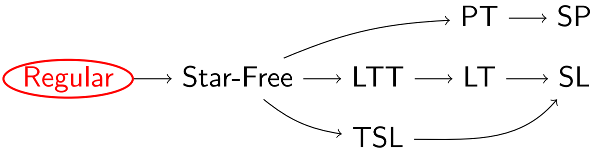
- *blick* is a possible English word
- *bnick* isn't

How complex are phonotactics?

Phonotactics are **regular** [Johnson, 1972, Kaplan and Kay, 1994].

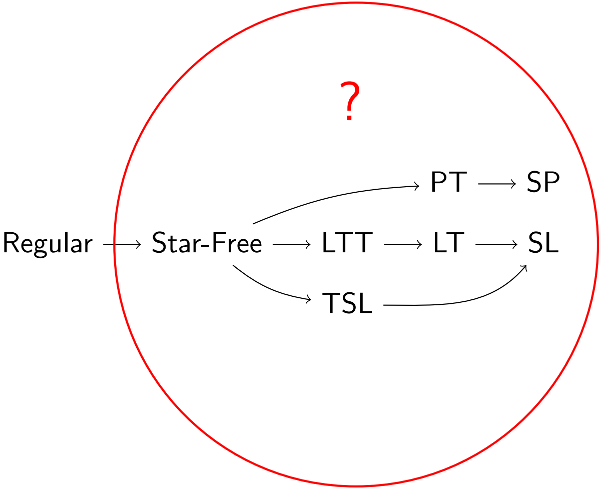
- Can be computed by regular grammars/automata
- But, generates a lot of patterns unattested in natural languages
- Not learnable from positive data [Gold, 1967]

Thus...



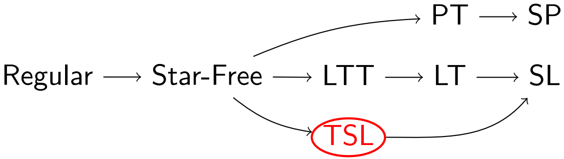
How complex are phonotactics?

The *subregular hypothesis*: phonotactics are **subregular** [Heinz, 2018].



How complex are phonotactics?

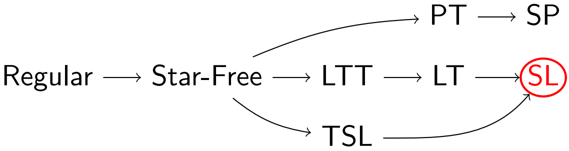
The *weak subregular hypothesis*: phonotactics are **tier-based strictly local (TSL)** [Heinz, 2018].



Strictly local languages

TSL languages are easiest to define starting from strictly local (SL) languages.

Informally: SL languages are generated by grammars that prohibit (or allow) certain *substrings*.



Strictly local languages

- Σ is an alphabet
- $\rtimes$ and $\ltimes$ are beginning and end markers, $\rtimes, \ltimes \notin \Sigma$
- For $s \in \Sigma^*$, $F_k(s)$ is the set of all length-k substrings of $\rtimes^{k-1} s \ltimes^{k-1}$
- A k-SL grammar G is a finite set of strings from $(\{\rtimes, \ltimes\} \cup \Sigma)^k$
- $s \in \Sigma^*$ is well-formed with respect to G iff $F_k(s) \cap G = \emptyset$
- A language L is SL iff there is some k such that L can be generated by a k-SL grammar.

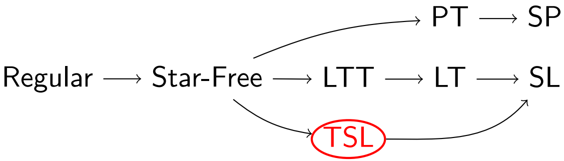
Strictly local languages

Let $\Sigma = \{a, b, c\}$. Suppose we want to generate a language L where b and c cannot be adjacent.

- Define a 2-SL grammar $G = \{bc, cb\}$
- $ababca \notin L$ because $F_2(ababca) = \{\rtimes a, ab, ba, \underline{\mathbf{bc}}, ca, a\ltimes\}$ ✗
- $ababaca \in L$, because $F_2(ababaca) = \{\rtimes a, ab, ba, ac, ca, a\ltimes\}$ ✓

Tier-based strictly local languages

TSL grammars [Heinz et al., 2011] are like SL grammars where we first remove irrelevant symbols before checking for illicit substrings.



Regular → Star-Free → LTT → LT → SL

PT → SP

TSL

Tier-based strictly local languages

A k-TSL grammar is a tuple (T, G) where

- $T \subseteq \Sigma$
- G is a finite set of strings from $(\{\rtimes, \ltimes\} \cup T)^k$

The tier representation of a string is generated by a projection function that 'erases' irrelevant symbols:

$$E_T(\sigma_1 \cdots \sigma_n) = u_1 \cdots u_n$$

where $u_i = \sigma_i$ iff $\sigma_i \in T$ and $u_i = \lambda$ (the empty string) otherwise.

- $s \in \Sigma^*$ is well formed with regard to a k-TSL grammar (T, G) iff $F_k(E_T(s)) \cap G = \emptyset$
- A language L is TSL iff there is some k such that L can be generated by a k-TSL grammar

Tier-based strictly local languages

Let $\Sigma = \{a, b, c\}$. Suppose we want to define a language L that does not allow words that contain both b and c.

- SL won't work because any number of a's can go between b and c
- Define a 2-TSL grammar where $T = \{b, c\}$ and $G = \{bc, cb\}$
- e.g. $E_T(abaaaca) = bc$ and $F_2(E_T(abaaaca)) = \{\rtimes b, \underline{\textbf{bc}}, c\ltimes\}$ ✗

Why TSL as an upper bound?

TSL grammars provide a desirable upper bound for phonological complexity.

Powerful enough...

- Captures long distance *harmony* patterns, where non-adjacent segments in a word must agree for some property.
- e.g. sibilant anteriority harmony in Aari [Hayward, 1990]:

UR	**SR**	**Gloss**
/baʔ-s-e/	[baʔ**s**e]	‘he brought’
/ʃed-er-s-it/	[**ʃ**eder**ʃ**it]	‘I was seen’
	*[**ʃ**eder**s**it]	
/ʒa̤ːg-er-s-e/	[**ʒ**a̤ːger**ʃ**e]	‘it was sewn’
	*[**ʒ**a̤ːger**s**e]	

Why TSL as an upper bound?

.. and restrictive enough!

- e.g. a language where words must have an even number of vowels is regular but not TSL

Learnable in polynomial time from positive data
[Jardine and Heinz, 2016, Jardine and McMullin, 2017]

Learnable in artificial grammar learning experiments
[McMullin and Ólafur Hansson, 2016, McMullin, 2016]

Is TSL enough?

TSL is restricted to a single tier

- Multiple long-distance patterns sometimes cannot be handled by a single TSL grammar
- Even worse if these patterns *conflict*

There are a handful of known examples of segmental phonology that are not TSL for these reasons.

- Tamashek Tuareg and Imdlawn Tashlhiyt sibilant harmony [McMullin, 2016]
- Sanskrit n-retroflexion harmony [Graf and Mayer, in prep.]
- **Uyghur backness harmony**

Uyghur backness harmony

Uyghur is a southeastern Turkic language.

- About 10 million speakers in China and neighboring countries.

- Backness harmony requires suffix forms to agree in backness with vowels and certain consonants within a stem [Lindblad, 1990, Vaux, 2000]

  - We use the locative suffix /-DA/ as a prototypical example
  - Backness agreement is reflected in the vowel: /a/ or /æ/
  - Voicing changes in the initial segment are not relevant: /t/ or /d/

Uyghur backness harmony

Table: The Uyghur vowel system. Harmonizing vowels are colored.

	Front		Back	
	Unrounded	Round	Unrounded	Round
High	i	**y**		**u**
Mid	e	**ø**		**o**
Low	**æ**		**a**	

Table: The harmonizing Uyghur dorsal consonants

	Front	Back
Voiceless	k	q
Voiced	g	ʁ

Uyghur backness harmony

The suffix must match the backness of the final harmonizing vowel in the stem.

Form	Gloss	Harmony type
aʁinæ-**dæ** friend-LOC	"on the friend"	Closest front vowel
qoichi-**da** shepherd-LOC	"on the shepherd"	Closest back vowel

Uyghur backness harmony

Even if there are conflicting harmonizing consonants.

Form	Gloss	Harmony type
rak-**ta** shrimp-LOC	“on the shrimp”	Closest back vowel across front dorsal
mæʃq-**tæ** exercise-LOC	“on the exercise”	Closest front vowel across back dorsal

Uyghur backness harmony

If there is no harmonizing vowel, the stem must match the backness of the final harmonizing dorsal consonant (/k/, /g/, /q/, /ʁ/).

Form	Gloss	Harmony type
gezit-**tæ** newspaper-LOC	"on the newspaper"	Closest front dorsal
qirʁiz-**da** Kyrgyz-LOC	"on the Kyrgyz"	Closest back dorsal

Uyghur backness harmony

If there are neither harmonizing vowels nor harmonizing dorsal consonants, the stem is arbitrarily specified for backness.

Form	Gloss	Harmony type
it-**ta** dog-LOC	"on the dog"	No harmonizers, arbitrarily back
biz-**dæ** we-LOC	"on us"	No harmonizers, arbitrarily front

Alternative analyses

There may be alternative analyses of Uyghur backness harmony that mitigate the issues to be described (see paper)

- No transparent vowels [McCollum, 2018]
- Backness harmony is a lexicalized pattern

We're in the process of collecting data on this!

The formal complexity of Uyghur backness harmony

We show that Uyghur backness harmony is not TSL under the assumed analysis.

Because segmental content is not crucially important, we use a more abstract notation:

- $V_f = y|ø|æ$
- $V_b = u|o|a$
- $C_f = k|g$
- $C_b = q|ʁ$
- S_f and S_b are front and back suffix forms
- $\Sigma_h = \{V_f, V_b, C_f, C_b, S_f, S_b\}$

These abbreviations group together segments that are *functionally equivalent*, and omit segments that are *transparent*.

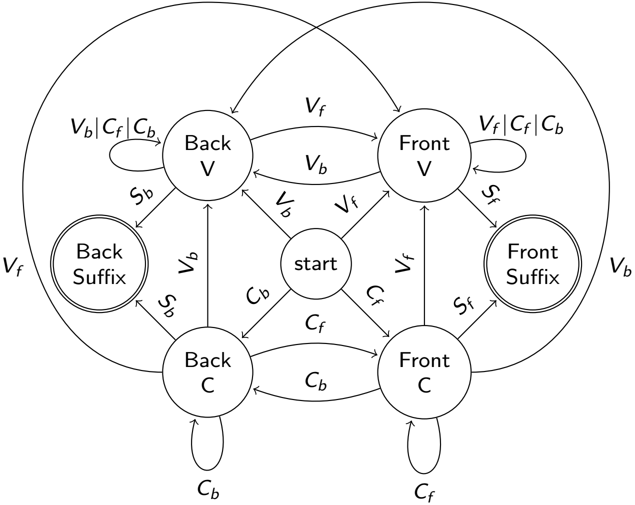
Uyghur backness harmony is regular

The following regular expression captures licit forms under backness harmony.

$$(\overline{(S_f|S_b)}^* V_f \overline{(V_b|S_f|S_b)}^* S_f)|(\overline{(S_f|S_b)}^* V_b \overline{(V_f|S_f|S_b)}^* S_b)$$
$$|(\overline{(V_f|V_b|S_f|S_b)}^* C_f C_f{}^* S_f)|(\overline{(V_f|V_b|S_f|S_b)}^* C_b C_b{}^* S_b)$$

Thus Uyghur backness harmony is at most regular.

Uyghur backness harmony is regular



Challenges for TSL

The vowel component in isolation can be captured by defining a 2-TSL grammar over the tier

$$T_v = \{V_f, V_b, S_f, S_b\}$$

where

$$G_v = \{V_f S_b, V_b S_f\}$$

- *mæʃq-**ta** → $V_f C_b S_b$
- $E_{T_v}(V_f C_b S_b) = \underline{V_f S_b}$ ✗

- mæʃq-**tæ** → $V_f C_b S_f$
- $E_{T_v}(V_f C_b S_f) = V_f S_f$ ✓

Challenges for TSL

The consonant component in isolation can be captured by defining a 2-TSL grammar over the tier

$$T_c = \{C_f, C_b, S_f, S_b\}$$

where

$$G_c = \{C_f S_b, C_b S_f\}$$

- *qirʁiz-**dæ** → $C_b C_b S_f$
- $E_{T_v}(C_b C_b S_f) = C_b \underline{C_b S_f}$ ✗
- qirʁiz-**da** → $C_b C_b S_b$
- $E_{T_v}(C_b C_b S_b) = C_b C_b S_b$ ✓

Challenges for TSL

If a TSL formulation were able to capture the interaction between the vowel and consonant patterns, it would need to be over the tier

$$T = T_v \cup T_c \cup \{\rtimes\}$$

$\rtimes$ is necessary because we need to be able to look back to the beginning of the tier to determine if there is a vowel to harmonize with.

But any number of harmonizing dorsals can intervene between the final harmonizing vowel and suffix!

Challenges for TSL

Let $C = C_f | C_b$ and define a k-TSL grammar for some fixed k where G contains the following k-factors:

$$V_b C^{k-2} S_f$$

$$V_f C^{k-2} S_b$$

$$\rtimes C^{k-3} C_b S_f$$

$$\rtimes C^{k-3} C_f S_b$$

This accepts strings like

$$V_b {C_f}^{k-1} S_f$$

but such forms violate backness harmony!

Challenges for TSL

k-factors cannot see the vowel and suffix at the same time!

Uyghur backness harmony cannot be TSL!

MTSL

An intuitive extension to TSL is the *intersection* of multiple TSL grammars.

TSL is not closed under intersection in general.

The class of intersections of TSL languages is the *multi-tier strictly local* (MTSL) languages [de Santo and Graf, 2017].

MTSL $\subsetneq$ Star-Free

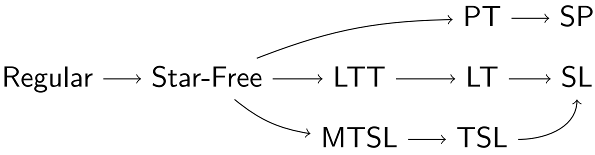
Because violations of each grammar are given equal weight, even this more powerful class cannot capture Uyghur backness harmony.

- e.g. grammatical forms like [mæʃq-**tæ**] violate the consonant harmony grammar

Interim summary

Uyghur backness harmony is not TSL nor MTSL.

What about the other languages in the subregular hierarchy?



Regular ⟶ Star-Free ⟶ LTT ⟶ LT ⟶ SL

Star-Free ⟶ PT ⟶ SP

Star-Free ⟶ MTSL ⟶ TSL ⟶ SL

Other subregular languages

Uyghur backness harmony can be generated by star-free grammars because they can encode precedence relations:

$$\forall x[S_b(x) \Rightarrow \forall y[V_f(y) \Rightarrow \exists z[V_b(z) \wedge y < z < x]]]$$

$$\forall x[S_f(x) \Rightarrow \forall y[V_b(y) \Rightarrow \exists z[V_f(z) \wedge y < z < x]]]$$

$$\forall x[S_b(x) \wedge \neg\exists y[V_f(y) \vee V_b(y)] \Rightarrow \forall z[C_f(z) \Rightarrow \exists w[C_b(w) \wedge z < w < x]]]$$

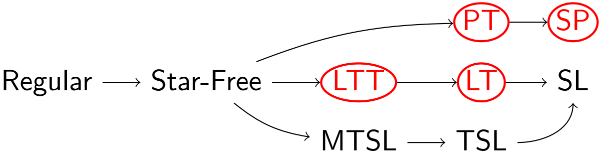
$$\forall x[S_f(x) \wedge \neg\exists y[V_f(y) \vee V_b(y)] \Rightarrow \forall z[C_b(z) \Rightarrow \exists w[C_f(w) \wedge z < w < x]]]$$

Star-free languages are not learnable in the limit [Gold, 1967], and may be too expressive to be a good model of natural language.

Other subregular languages

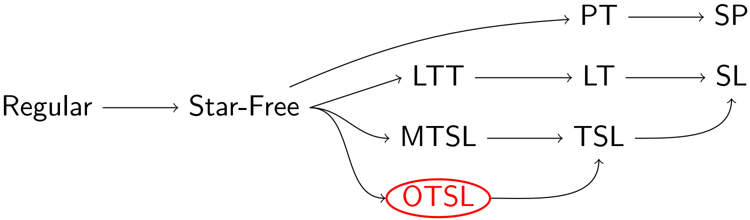
But it does not fall into any other commonly discussed classes (see paper).

- Not strictly piecewise or piecewise testable.
- Not locally testable or locally threshold testable
- Not interval-bounded strictly piecewise [Graf, 2017].



Output tier-based strictly local

Can be captured by a natural extension of TSL: output tier-based strictly local (OTSL) [Graf and Mayer, in prep.].



Output tier-based strictly local

TSL projection function E_T is a 1-ISL or 1-OSL map [Chandlee, 2014].

- Generalize to a k-OSL map
  - i.e. consider the preceding $k-1$ symbols on the tier when deciding whether to project

Uyghur backness harmony can be captured with a 2-OTSL grammar.

- V_f, V_b, S_f, and S_b are always projected
- C_f and C_b are projected if the previous symbol is not V_f or V_b
- $G = \{C_f S_b, C_b S_f, V_f S_b, V_b S_f\}$

Unclear how useful this formalism is for modeling natural language.

Conclusion

Segmental patterns that are not TSL are uncommon.

Uyghur backness harmony is more complex than most of these patterns.

- Suggests that hypotheses about phonotactic complexity should be revised, OR
- Uyghur backness harmony needs to be better understood

This pattern shows an interesting divergence in complexity between formal language models and Optimality Theory!

An OT analysis

Uyghur backness harmony is simple to model in OT.

mæʃq-DA	Harmonize V	Harmonize C
☞ a. mæʃq-**tæ**		*
b. mæʃq-**ta**	*!	

Two things to consider:

- OT lends itself very well to an analysis of such a pattern
- These patterns appear to be quite uncommon

Such patterns may be useful in considering how formal language models can integrate with existing linguistic analyses.

Acknowledgements

Thanks to Tim Hunter, Kie Zuraw, Thomas Graf, two anonymous reviewers, and the attendees of the UCLA phonology seminar for their invaluable feedback. We would also like to thank our Uyghur consultants for sharing their language and culture.

References I

Chandlee, J. (2014).
Strictly local phonological processes.
PhD thesis, The University of Delaware.

de Santo, A. and Graf, T. (2017).
Structure sensitive tier projection: Applications and formal properties.
Ms., Stony Brook University.

Gold, E. M. (1967).
Language identification in the limit.
Information and Control, 10:447–474.

Graf, T. (2017).
The power of locality domains in phonology.
Phonology, 34:385–405.

References II

Hayward, R. J. (1990).
Notes on the aari language.
In Hayward, R. J., editor, *Omotic language studies*, pages 425–493. School of Oriental and African Studies, University of London, London.

Heinz, J. (2018).
The computational nature of phonological generalizations.
In *Phonological Typology*.

Heinz, J., Rawal, C., and Tanner, H. G. (2011).
Tier-based strictly local constraints for phonology.
Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers, pages 58–64.

Jardine, A. and Heinz, J. (2016).
Learning tier-based strictly 2-local languages.
Transactions of the ACL, 4:87–98.

References III

Jardine, A. and McMullin, K. (2017).
Efficient learning of tier-based strictly *k*-local languages.
In Drewes, F., Martín-Vide, C., and Truthe, B., editors, *Language and Automata Theory and Applications, 11th International Conference*, pages 64–76. Springer.

Johnson, C. D. (1972).
Formal Aspects of Phonological Decription.
Mouton, The Hague.

Kaplan, R. and Kay, M. (1994).
Regular models of phonological rule systems.
Computational Linguistics, 20:331–378.

Lindblad, V. M. (1990).
Neutralization in Uyghur.
University of Washington.

References IV

McCollum, A. (2018).
Locality, transparency, and uyghur backness harmony.
In *The Twenty-Sixth Manchester Phonology Meeting Abstracts Booklet*.

McMullin, K. (2016).
Tier-based locality in long-distance phonotactics: Learnability and typology.
PhD thesis, University of British Columbia.

McMullin, K. and Ólafur Hansson, G. (2016).
Long-distance phonotactics as tier-based strictly 2-local languages.
Procedings of the Annual Meetings on Phonology 2014.

Vaux, B. (2000).
Disharmony and derived transparency in Uyghur vowel harmony.
Proceedings of NELS 30, pages 671–698.

Appendix: Acoustic study

General question: is there phonetic evidence for a phonemic backness constrast between /i/ and /ɨ/?

Specific question: Do vowels in forms with no harmonizing segments show F2 differences predictable from the suffixes they take?

Appendix: Acoustic study

Tables: Word lists for speakers 1 and 2. Bolded forms indicate disagreements in stem backness between the speakers.

Front		**Back**	
/bil/	‘know’	/ʧiʃ/	‘tooth’
/bir/	‘one’	/dil/	‘heart’
/biz/	‘we’	**/mis/**	**‘copper’**
/din/	**‘religion’**	/pil/	‘elephant’
/iʃ/	‘work’	/sirt/	‘outside’
/ʤin/	**‘Djinn’**	/siz/	‘draw’
/min/	‘ride’	/til/	‘tongue’
/sir/	**‘brush’**	/tiz/	‘knee’
/siz/	‘you’		

Front		**Back**	
/bil/	‘know’	/ʧiʃ/	‘tooth’
/bir/	‘one’	/dil/	‘heart’
/biz/	‘we’	**/din/**	**‘religion’**
/min/	‘ride’	/it/	‘dog’
/mis/	**‘copper’**	**/ʤin/**	**‘Djinn’**
/siz/	‘you’	/lim/	‘beam’
		/pil/	‘elephant’
		/pir/	‘master’
		/sir/	**‘brush’**
		/sirt/	‘outside’
		/siz/	‘draw’
		/til/	‘tongue’
		/tiz/	‘knee’

Appendix: Acoustic study

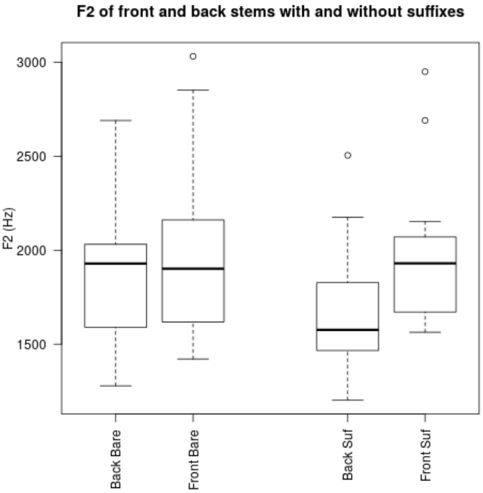
Speakers produced the words in the carrier sentence

tursun hazir	________	*dɛdi*
Tursun again	________	say.PAST
Tursun said	________	again.

Elicited words in two forms:

- No harmonizing suffix
  - Bare for nouns
  - Suffix *-di* for verbs
- With harmonizing suffix
  - *-DA* for nouns
  - *-mAQ* for verbs

Appendix: Acoustic study



- No difference in F2 in bare forms between front and back stems
- Back suffixes pull vowels in stem back (coarticulation)
- No clear evidence of a phonemic distinction