Formal Language Theory and Phonology: Day 2

Jane Chandlee and Adam Jardine

July 6, 2023

1 Review

The following quote from (Engelfriet and Hoogeboom, 2001) is a nice reminder of why what we're looking at is important:

> It is always a pleasant surprise when two formalisms, introduced with different motivations, turn out to be equally powerful, as this indicates that the underlying concept is a natural one. Additionally, this means that notions and tools from one formalism can be made use of within the other, leading to a better understanding of the formalisms under consideration (p. 216)

Engelfriet, J. and Hoogeboom, H. J. (2001). MSO definable string transductions and two-way finite-state transducers. *ACM Transations on Computational Logic*, 2:216–254

2 The Strictly Local sets

Definition 1 (k-factor) *A string u is a k-factor of another string w iff u is of length k, w is of length $\geq k$, and $w = v_1uv_2$ for some other strings v_1 and v_2; that is, w is the **concatenation** of three strings v_1, u, and v_2 (in that order). If w is of length $< k$, then u is a k-factor of w iff $u = w$.*

concatenation

- Examples:
  - What are the 2-factors of $abab$?
  - What are the 3-factors of $aaba$?
  - What are the 6-factors of $aaba$?
- For a *set* L of strings, the k-factors of L are

$$\bigcup_{w \in L} \{u \mid u \text{ is a } k\text{-factor of } w\}$$

  that is, the union of the set of k-factors for each string w in L.

- We need to be able to distinguish k-factors at the beginning, middle, and ends of words. To do this, we pick two special symbols $\rtimes$ and $\ltimes$ not in Σ that mark the beginning and end of a string, respectively. Let $\rtimes\Sigma^*\ltimes$ denote the set of all strings in Σ^* marked with the boundary symbols.

- Examples:
  - $aaba \rightarrow \rtimes aaba \ltimes$
  - $\lambda \rightarrow \rtimes\ltimes$

Definition 2 (SL$_k$ grammar) *A **SL$_k$ grammar** is a set G of k-factors of $\rtimes\Sigma^*\ltimes$. A string $w \in \Sigma^*$ **satisfies** G (written $w \models G$) if none of the k-factors of $\rtimes w \ltimes$ are in the set G. The set $L(G)$ is the set of strings that satisfy G, i.e.*

$$L(G) = \{w \in \Sigma^* \mid w \models G\}$$

SL$_k$ grammar

satisfies

- What is a SL$_2$ grammar for the set $(ab)^n$?

- What is a SL$_3$ grammar for the set of strings over $\Sigma = \{a, b\}$ that satisfy the generalization "b does not occur three times in a row"?

- Is there a SL$_k$ grammar for $(aa)^n$?

- A set is SL$_k$ if it is described by some SL$_k$ grammar. A set is SL if it is SL$_k$ for some k.
- The abstract characterization for SL is as follows.

Theorem 1 (k-suffix substitution closure (Rogers and Pullum, 2011)) *A set L is SL iff there is some k such that for all strings v_1, v_2, w_1, w_2 whenever there is a string x of length $k-1$, then*

$$w_1xw_2 \in L \text{ and } v_1xv_2 \in L \text{ implies } w_1xv_2 \in L$$

k-suffix substitution closure

Rogers, J. and Pullum, G. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*, 20:329–342

- Let's see how this holds in $(ab)^n$ for $k = 2$.

- Does this hold for $(aa)^n$ for $k = 2$? What about for $k = 3$? For any k?

- Note that k-suffix substitution closure is *a property of the set itself*—it makes no reference to a particular grammar formalism (e.g., SL$_k$ grammars).

3 The Strictly Piecewise sets

Classic example of long-distance sibilant harmony in Navajo (Athabaskan; Southwestern U.S., Navajo Nation; Sapir and Hoijer, 1967):

a.	/sì-ʔá/	[sì-ʔá]	‘a round object lies’
b.	/sì-tí/	[sì-tí]	‘he is lying’
c.	/sì-ɣìʃ/	[ʃì-ɣìʃ]	‘it is bent, curved’
d.	/sì-teːʒ/	[ʃì-teːʒ]	‘they (dual) are lying’

Prove this pattern is **not** SL for any k.

(Heinz, 2010) analyses this type of pattern with a **precedence grammar**, which is another name for **strictly piecewise**.

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

precedence grammar

strictly piecewise

Let’s define a language L_{*bc} with $\Sigma = \{a, b, c\}$ and a constraint *b...c.

What are some strings that **are** in this language? What are some strings that are **not** in this language?

Definition 3 [1] *[subsequence] $w \in \Sigma^*$ is a* ***subsequence*** *of $v \in \Sigma^*$ ($w \sqsubseteq v$) iff $w = \lambda$ or $w = \sigma_1 ... \sigma_n$ and $\exists x_0, ..., x_n \in \Sigma^*$ such that $v = x_0\sigma_1 x_1 ... \sigma_n x_n$.*

subsequence

Definition 4 (k-pieces) *For $w \in \Sigma^*$, the set of* ***k-pieces*** *of w is*

$$P_k(w) = \{v \in \Sigma^{\leq k} : v \sqsubseteq w\}$$

k-pieces

What are the 2-pieces of the string abc?

Definition 5 (SP$_k$ grammar) *A* ***SP$_k$ grammar*** *is a set G of k-pieces of $\rtimes\Sigma^*\ltimes$. A string $w \in \Sigma^*$* ***satisfies*** *G (written $w \models G$) if none of the k-pieces of $\rtimes w \ltimes$ are in the set G. The set $L(G)$ is the set of strings that satisfy G, i.e.*

$$L(G) = \{w \in \Sigma^* \mid w \models G\}$$

SP$_k$ grammar

satisfies

[1] The following definitions are adapted from Jim Rogers’s LING890 course notes (UDel, Spring 13).

What is the grammar for L_{*bc}?

Back to Navajo: this pattern is **symmetric**, meaning the hypothetical form *ʃitis is also ungrammatical. What is the SP$_2$ grammar for this pattern?

symmetric

As noted in Heinz (2010), Tsuut'ina (formerly known as Sarcee; Athabaskan; Calgary, Canada) also has sibilant harmony, but it is **asymmetric**: a [+anterior] segment like [s] can *follow* [−anterior] [ʃ], but it still can't *precede* it. As an SP$_2$ grammar, how does this pattern **differ** from Navajo?

asymmetric

A language is SP$_k$ if it is $L(G)$ for some SP$_k$ grammar G. It is SP if it is SP$_k$ for some k.

One abstract characterization of SP is the property of being **subsequence closed**:

subsequence closed

Theorem 2 *L is SP iff the following holds:* $w \in L$ *and* $v \sqsubseteq w \implies v \in L$.

Informally: in an SP language, if a string is in that language, then so must be all of its subsequences. We can see this more easily with an example where it fails (i.e., a non-SP language).

As discussed by Heinz (2010), consonant harmony with blocking is not SP. Consider a version of sibilant harmony in which the disagreeing sibilants are permitted provided a coronal obstruent intervenes between them: so *sipiʃ is ungrammatical but sitiʃ is grammatical.

Assuming $k = 2$, show that this language is not subsequence-closed. Does it help to instead assume $k = 3$? (Spoiler: no. But why not?)

Citing a typological observation by Hansson (2001) and Rose and Walker (2004), Heinz (2010) claims blocking patterns such as this are unattested. Subsequent work, however, challenged that claim - in fact the above example is based on a reported pattern in Slovenian (Jurgec, 2011).

McMullin (2016) uses the existence of consonant harmony with blocking to argue that **tier-based strictly local (TSL)** languages are a better characterization of long-distance phonotactics than SP. We won't have time to cover TSL in this course, but relevant sources will be included in the further readings list. In short, TSL is just SL over a subset of the alphabet called the **tier**.

tier-based strictly local (TSL)

tier

For example, instead of using subsequences and SP to characterize sibilant harmony, we can first project a tier of sibilants and then define SL constraints like *sʃ, *zʃ, etc. When there is blocking, the blocking segments are simply included on the tier and the same constraints can be used.

The tier segments of sitiʃ are stʃ, and this string does not contain the prohibited 2-factor *sʃ. But the tier segments of sipiʃ are sʃ, which does contain the prohibited 2-factor.

4 Typological predictions

Rogers and Pullum (2011) define additional language classes above SL, including locally testable and locally threshold testable. SP is also contained by a piecewise testable class. But as discussed in Heinz (2018), these classes appear to be overly complex with respect to phonotactic patterns.

5 Next time

Reading: Heinz and Lai (2013)

Task: Find a function that is not left-subsequential and *prove* that it isn't.

References

Engelfriet, J. and Hoogeboom, H. J. (2001). MSO definable string transductions and two-way finite-state transducers. *ACM Transations on Computational Logic*, 2:216–254.

Hansson, G. (2001). *Theoretical and typological issues in consonant harmony*. PhD thesis, University of California Berkeley.

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

Heinz, J. (2018). The computational nature of phonological generalizations. In Hyman, L. and Plank, F., editors, *Phonological Typology*, Phonetics and Phonology, chapter 5, pages 126–195. De Gruyter Mouton.

Heinz, J. and Lai, R. (2013). Vowel harmony and subsequentiality. In Kornai, A. and Kuhlmann, M., editors, *Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13)*, pages 52–63, Sofia, Bulgaria.

Jurgec, P. (2011). *Feature spreading 2.0: a unified theory of assimilation*. PhD thesis, University of Tromsø.

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

Rogers, J. and Pullum, G. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*, 20:329–342.

Rose, S. and Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, 80:475–531.

Sapir, E. and Hoijer, H. (1967). *The phonology and morphology of the Navaho language*. University of California Press, Berkeley.