Learning tiers for long-distance phonotactics

Adam Jardine

University of Delaware

February 20, 2015

Introduction

- Languages have long-distance phonotactic patterns
- Children learn these patterns
- How could *anything* learn these patterns?

Introduction

- This talk presents a new learning algorithm for long-distance phonotactic dependencies
- Algorithm is provably correct with specific criteria for learning
- Theoretical result: phonological concepts of *tier* & *locality* are sufficient to induce a tier and grammar from positive data

Long-distance dependencies

- Latin: [l] and [r] alternate, ignoring intervening Cs and Vs (Jensen, 1974)

navalis	'naval'	militaris	'military'
episcopalis	'infinitalis'	floralis	'floral'
infinitalis	'negative'	sepulkralis	'funeareal'
solaris	'solar'	litoralis	'of the shore'
lunaris	'lunar'		

- Finnish: vowels in a word are either [+back] or [−back], ignoring Cs and /i,e/ (Ringen, 1975)

pöütä-nä	'table-ESS'
väkkärä-nä	'pinwheel-ESS'
ulko-ta	'outside-ABL'
pappi-na	'priest-ESS'

Trans.	[−B]	[+B]
i	ü	u
e	ö	o
	ä	a

Learning long-distance dependencies

- One solution: stipulate tier (Hayes and Wilson, 2008)
- Latin

navalis	'naval'	militaris	'military'
episcopalis	'infinitalis'	floralis	'floral'
infinitalis	'negative'	sepulkralis	'funeareal'
solaris	'solar'	litoralis	'of the shore'
lunaris	'lunar'		

- Strictly Local (SL) learning (García et al., 1990; Heinz, 2010): Contiguous chunks of a certain size
- SL grammar for liquids: $\{ll, rr\}$ are banned *on the stipulated tier*

Learning long-distance dependencies

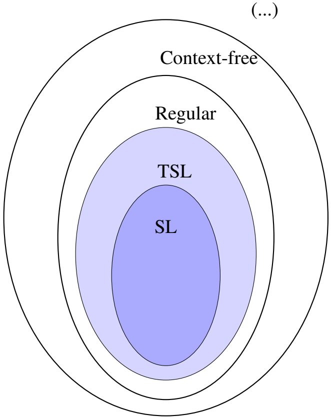
- Such a learner would have a lot of tiers to consider:
  Vowels (Turkish; Clements and Sezer, 1982)
  Vowels except /i,e/ (Finnish; Ringen, 1975)
  Liquids (Sundanese; Cohn, 1992)
  Liquids and non-coronals (Latin; Cser, 2010)
  Sibilants (Samala; Rose and Walker, 2004)
  Sibilants and round vowels, /r/, and some voiced
  obstruents (Koorete; McMullin and Hansson, 2014)
  ...
- An inventory of n phonemes has 2^n possible tiers
- Is it possible to *discover* a tier?

Learning a tier

- Goldsmith and Riggle (2012) offer solution based on mutual information
- They demonstrate an algorithm that works on Finnish data
- Some unanswered questions:
  - Does it work for any tier?
  - What kind of data does it need to see?

Tier-based SL

- *Tier-based Strictly Local (TSL)* formal languages (Heinz et al., 2011): generalization of SL with variable tier
- Hypothesized upper complexity bound for phonotactics (Heinz et al., 2011; Rogers et al., 2013; McMullin and Hansson, 2014)
- Evidence that such boundaries are psychologically real (Lai, 2013, 2014; McMullin and Hansson, 2014)



- I present a learner that *provably* learns TSL, given certain data

TSL grammars

- Given an inventory Σ, a TSL grammar is $G = (T, \overline{S})$ where:
  - $T \subseteq \Sigma$
  - $\overline{S}$ are *banned tier substrings* of elements of T
  - G checks strings if they include a member of $\overline{S}$, *ignoring any elements not on T*
- Example: $\Sigma = \{a, b, c\}$, $T = \{a, b\}$, $\overline{S} = \{aa, bb\}$
  - ✓ $aba, bab, bcacb, bcaccccbca, \ldots$
  - ✗ $aca, acccccca, cabccccbcc, \ldots$
- Latin: $\Sigma = \{l, r, C, V\}$, $T = \{l, r\}$, $\overline{S} = \{ll, rr\}$
- TSL learned like SL *if* we know T (Heinz et al., 2011)
- What if we don't?

The TSL Learning Algorithm

- Problem: Given an inventory Σ and a sufficient data set D of words, what is the TSL grammar for the language that generated D?
- This problem is *solvable* (Jardine and Heinz, in prep.)
- Partial solution given here to illustrate the main idea
- Goal: find nontier elements (freely distributed)
- Key: learn about tier piece-by-piece, building on knowledge

The TSL Learning Algorithm

- *Path*: $\langle \sigma_1, X, \sigma_2 \rangle$
  "σ_1 precedes σ_2 (at any distance), and
  X is the set of symbols which appear between them"

(1) $$\mathit{paths}(CVClVr) = \{ \begin{array}{l} \langle C, \{\}, V \rangle \\ \langle C, \{V\}, C \rangle \\ \quad \ldots \\ \langle C, \{V, C, l, \}, r \rangle \\ \quad \ldots \\ \langle V, \{C, l\}, V \rangle \\ \quad \ldots \\ \langle V, \{\}, r \rangle \end{array} \}$$

The TSL Learning Algorithm (simple version)

- Problem: Given an inventory Σ and a sufficient data set D of words, what is the TSL grammar $(T, \overline{S})$ for the language that generated D?
- Solution:
  a. Calculate all paths for all words in D
  b. Start with $T = \Sigma$ as guess for tier
  c. Look at set of paths $\langle \sigma_1, X, \sigma_2 \rangle$ where σ_1, σ_2 on tier and X are non-tier elements (σ_1 and σ_2 are *tier adjacent*)
  d. Is there any member of T which is tier adjacent to every other member?
  e. If so, remove that member from T, and repeat from step (c)
  f. If not, return T, and $\overline{S}$ is any tier elements not tier-adjacent

The TSL Learning Algorithm (simple version)

- $\Sigma = \{C, V, l, r\}$
- T initialized to $\{C, V, l, r\}$
- Considering paths $\langle \sigma_1, \{\}, \sigma_2 \rangle$, are any symbols free?

$$T = \{C, V, r, l\}$$

navalis	militaris
episcopalis	floralis
infinitalis	sepulkralis
solaris	litoralis
lunaris	migrus
certe	

Data
(Latin; Jensen, 1974; Odden, 1994; Cawley, 2014)

The TSL Learning Algorithm (simple version)

- Yes! C is a free element:

$\langle \#, \{\}, C\rangle$	**n**avalis
$\langle C, \{\}, \#\rangle$	navali**s**
$\langle C, \{\}, C\rangle$	epi**sc**opalis
$\langle l, \{\}, C\rangle$	sepu**lk**ralis
$\langle C, \{\}, l\rangle$	**fl**oralis
$\langle r, \{\}, C\rangle$	ce**rt**e
$\langle C, \{\}, r\rangle$	mi**gr**us
$\langle V, \{\}, C\rangle$	n**av**alis
$\langle C, \{\}, V\rangle$	na**va**lis

- C is thus removed, and $T = \{V, r, l\}$

$$T = \{C, V, r, l\}$$

navalis	militaris
episcopalis	floralis
infinitalis	sepulkralis
solaris	litoralis
lunaris	migrus
certe	

Data

(Latin; Jensen, 1974; Odden, 1994; Cawley, 2014)

- Considering paths $\langle \sigma_1, \{\}, \sigma_2\rangle$, and $\langle \sigma_1, \{C\}, \sigma_2\rangle$ are any symbols free?

The TSL Learning Algorithm (simple version)

- Yes! V is a free element:

$\langle \#, \{\}, V \rangle$	**e**piscopalis
$\langle V, \{\}, \# \rangle$	cert**e**
$\langle l, \{\}, V \rangle$	nava**li**s
$\langle V, \{\}, l \rangle$	nav**al**is
$\langle V, \{\}, r \rangle$	c**er**te
$\langle r, \{\}, V \rangle$	mig**ru**s
$\langle V, \{\}, V \rangle$	none!
$\langle V, \{C\}, V \rangle$	n**ava**lis

- V is thus removed, and $T = \{r, l\}$

$T = \{V, r, l\}$

navalis	militaris
episcopalis	floralis
infinitalis	sepulkralis
solaris	litoralis
lunaris	migrus
certe	

Data
(Latin; Jensen, 1974; Odden, 1994; Cawley, 2014)

- Considering paths $\langle \sigma_1, X, \sigma_2 \rangle$, where $X \subseteq \{C, V\}$, are any symbols free?

The TSL Learning Algorithm (simple version)

- No! We don't see paths:
  $\langle l, X, l\rangle$
  $\langle r, X, r\rangle$
  (where $X \subseteq \{C, V\}$)
- l and r are thus not removed
- ll and rr are added to $\bar{S}$
- Algorithm halts

$$T = \{r, l\}$$

navalis	militaris
episcopalis	floralis
infinitalis	sepulkralis
solaris	litoralis
lunaris	migrus
certe	

Data
(Latin; Jensen, 1974; Odden, 1994; Cawley, 2014)

The TSL Learning Algorithm (simple version)

- TLA correctly returns $T = \{r, l\}, \overline{S} = \{ll, rr\}$ on this small data set
- Idea: find free elements, using a decreasing T to aid search
- Result on natural language data: on harmonic forms from Goldsmith and Riggle (2012)'s Finnish corpus
  - over inventory {u, o, a, ü, ö, ä, i, e, J, T, S, N}
  - TLA learns
    $T = \{\text{u}, \text{o}, \text{a}, \text{ü}, \text{ö}, \text{ä}\}$
    $\overline{S} = \{\text{uü}, \text{uö}, \text{uä}, \text{oü}, \text{oä}, ..., \text{äa}\}$
- It can be *proven* that this algorithm will always be correct, given the right data

Discussion

- Full version of algorithm accounts for ‘free blockers’ — members of tier which do not play role in generalizaton
  - E.g., *g*, *m* in Latin (*legalis* ‘legal’ **legaris*)
- Full version provably learns entire TSL class (Jardine and Heinz, in prep)
- However, it requires more data

Discussion

- Neither version of algorithm does well with raw natural language data
- Issue: local dependencies prevent removal from tier
- Ex. from Finnish: [d] not followed by other consonants
- Integrating features (as in test Finnish inventory) will likely help with raw data
- Another issue: what if we need more than one tier?
- Future work: do humans learn like this? (Finley, 2012; McMullin and Hansson, 2015)

Conclusion

- Introduced algorithm that finds a tier and long distance dependencies
- Key is incrementally learning the tier and generalizing on that knowledge
- Algorithm is guaranteed to learn a *class* of phonotactic constraint, given the right data
- Theoretical result: phonological concepts of tier and locality are sufficient for finding particular tier and particular long-distance dependencies
- While some limitations, brings us closer to theory of how grammars for long distance phonological phenomena can be reliably acquired from corpora (by humans?)

Acknowledgements

This research is indebted to the ideas and advice of Jeffrey Heinz and Rémi Eyraud, as well as discussion with the members of Irene Vogel's acquisition seminar at UD (Taylor Lampton Miller, Hyun Jin Hwangbo, Kalyn Matocha, Nicole Demers) and Kevin McMullin. Thanks also to Jason Riggle for providing the Finnish corpus.

References I

Cawley, K. (2014). Latin dictionary and grammar aid (website). http://archives.nd.edu/latgramm.htm.

Clements, G. N. and Sezer, E. (1982). Vowel and consonant disharmony in Turkish. In van der Hulst, H. and Smith, N., editors, *The Structure of Phonological Representations (Part II)*. Foris, Dordrecht.

Cohn, A. (1992). The consequences of dissimilation in Sundanese. *Phonology*, 9:199–220.

Cser, A. (2010). The -alis/aris- allomorphy revisited. In Kastovsky, D., Rainer, F., Dressler, W. U., and Luschützky, H. C., editors, *Variation and change in morphology: selected papers from the 13th international morphology meeting*, pages 33–51. Philadelphia: John Benjamins.

Finley, S. (2012). Testing the limits of long-distance learning: Learning beyond the three-segment window. *Cognitive Science*, 36:740–756.

García, P., Vidal, E., and Oncina, J. (1990). Learning locally testable languages in the strict sense. In *Proceedings of the Workshop on Algorithmic Learning Theory*, pages 325–338.

References II

Goldsmith, J. and Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

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

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

Jardine, A. and Heinz, J. (2015). Learning tier and grammar parameters for tier-based strictly local languages. Paper to be submitted to the Annals of Mathematics and Artificial Intelligence special issue on Mathematical Theories of Natural Language Processing.

Jensen, J. (1974). Variables in phonology. *Language*, 50:675–686.

Lai, R. (2013). *Domain Specificity in Learning Phonology*. PhD thesis, University of Delaware.

References III

Lai, R. (2014). Learnable versus unlearnable harmony patterns. *Linguistic Inquiry*. To appear.

McMullin, K. and Hansson, G. O. (2014). Long-distance phonotactics as Tier-Based Strictly 2-Local languages. In *Proceedings of the Annual Meeting on Phonology 2015*. Talk; manuscript to be submitted.

McMullin, K. and Hansson, G. O. (2015). Locality in long-distance phonotactics: evidence for modular learning. In *NELS 44, Amherst, MA*. in press.

Odden, D. (1994). Adjacency parameters in phonology. *Language*, 70(2):289–330.

Ringen, C. (1975). *Vowel Harmony: Theoretical Implications*. PhD thesis, Indiana University.

Rogers, J., Heinz, J., Fero, M., Hurst, J., Lambert, D., and Wibel, S. (2013). Cognitive and sub-regular complexity. In *Formal Grammar*, volume 8036 of *Lecture Notes in Computer Science*, pages 90–108. Springer.

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