Tier-based modeling of gradience and distance-based decay in phonological processes

Kevin McMullin and Phillip Burness
University of Ottawa, Canada

Mathematics of Language 17
December 13, 2021

Overview

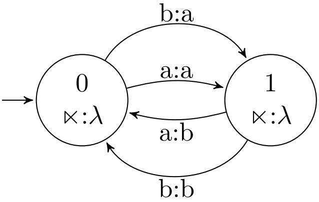
- Background
  - Notation
  - Transducers
  - Tier-based Strictly Local functions
- Allowing for optionality
- Distance-based decay
  - Case study 1: Malagasy
  - Case study 2: Hungarian
- Conclusion

Notation

- The empty string is λ
- String concatenation is indicated with $\cdot$
  - $ab \cdot b = abb$
  - Omitted when context permits
- The exponent -1 represents string subtraction
  - $(ab)^{-1} \cdot abb = b$
- Σ and Δ are the input and output alphabets
- Σ^* is all strings of any length made using elements of Σ

Transducers

- Incrementally write an output as they read an input
- Always in one of a finite set Q of states
- Non-final transitions of the shape (q, i, o, r)
  - When i is read while in state q, append o and move to r
- Final transitions with the shape (q, f)
  - When the derivation ends in state q, append f



Strictly k-Local (SL_k) functions

- Developped by Chandlee (2014) and Chandlee et al. (2014, 2015)
- Functional analogues to the SL_k languages (McNaughton and Papert, 1971; Rogers and Pullum, 2011; Rogers et al., 2013)
- Transducer state labels record the $k-1$ most recent:
    - Seen input letters (Input Strictly k-Local = ISL_k)
    - Produced output letters (Output Strictly k-Local = OSL_k)
- Model phonological transformations with triggering contests of bounded size (Chandlee and Heinz, 2018)

Tier-based Strictly k-Local (TSL_k) functions

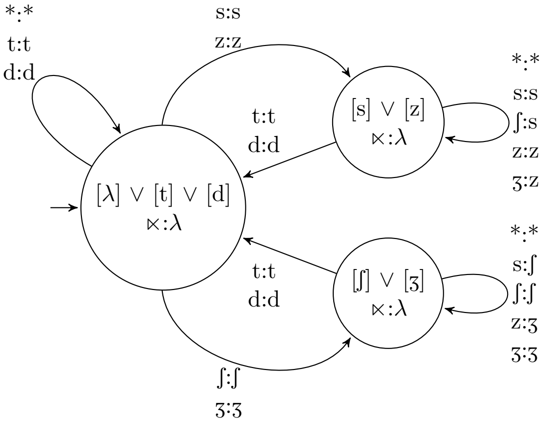
- Exactly like SL_k functions except we limit which letters may occupy space in state labels (Hao and Andersson, 2019; Hao and Bowers, 2019; Burness and McMullin, 2019)
- The *tier* is the set of such "relevant" letters
- Functional analogues to the TSL_k languages (Heinz et al., 2011)
- By relativizing the transducer's attention, we can capture processes whose triggers may be arbitrarily far from their targets (Andersson et al., 2020; Burness et al., 2021)

Example: Slovenian sibilant harmony

(1) Slovenian sibilant harmony
Blocked by coronal stops (Jurgec, 2011, pp. 330-331)

a.	/spi-ʃ/	[ʃpi-ʃ]	'sleep-2SG'
b.	/pozabi-ʃ/	[poʒabi-ʃ]	'forget-2SG'
c.	/stoji-ʃ/	[stoji-ʃ]	'stand-2SG'
d.	/zida-ʃ/	[zida-ʃ]	'build-2SG'

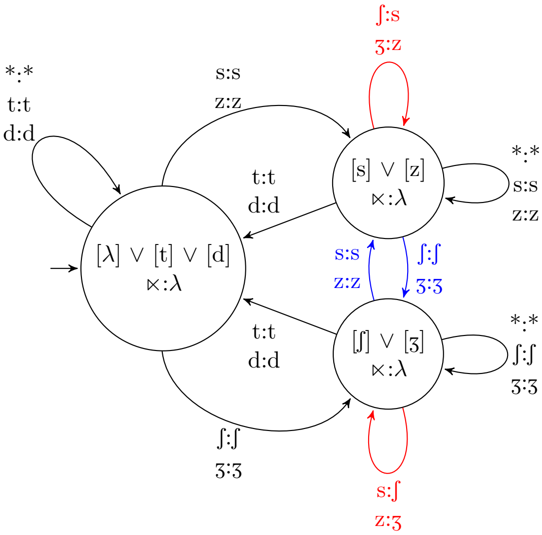
Example: Slovenian sibilant harmony



Relaxing determinism

- The Slovenian process is actually *optional* (Jurgec, 2011)
- By definition, however, TSL functions are deterministic
- Of course, we can always add extra transitions to achieve non-deterministic outcomes
- To preserve (tier-based) strict locality, extra transitions still need to land in the state corresponding to the:
    - Recently read tier elements (for quasi-ITSL models)
    - Recently produced tier elements (for quasi-OTSL models)

Relaxing determinism



Weighting transitions

- Choosing randomly between the two options in the previous transducer means there is a 50% chance of applying harmony.
- Giving transitions weights proportional to the desired traversal probability can achieve other ratios
- Generally, quasi-TSL models create a conditional distribution for each input string such that:
  - There is a finite number of output possibilities
  - The probabilities of the outputs sum to 1
- Crucially, the distributions are shaped according to tier-based strict locality

Utility of multiple tiers

- Work by Burness and McMullin (2020, 2021) investigates functions that track multiple tiers simultaneously
- We adhere to the requirements for their *target-specified* subclass of MTSL_k functions
    - (i) each input element pays attention only to its own tierset
    - (ii) each such *target-specified* tierset must form a strict superset-subset hierarchy
- This structure makes it possible to model some peculiarities of Bukusu liquid harmony:
    - Obligatory in adjacent syllables
    - Optional at larger distances (de Blois, 1975; Odden, 1994; Hansson, 2010)

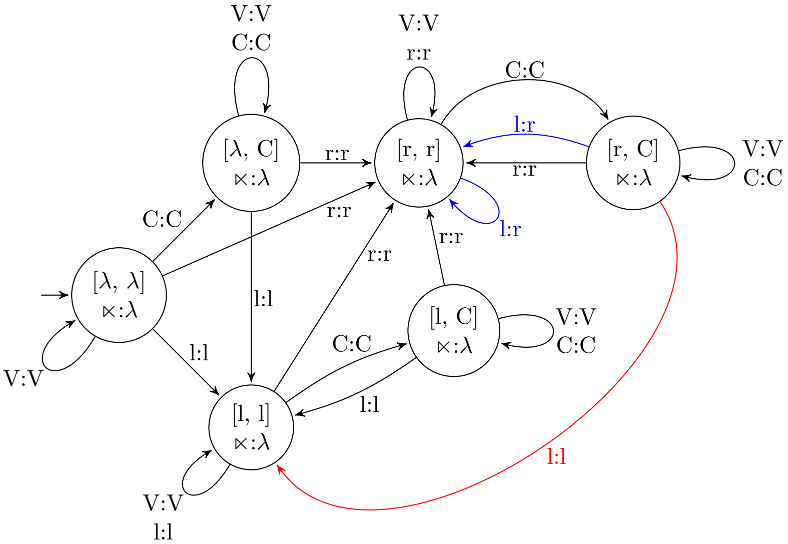
Utility of multiple tiers

(2) Bukusu liquid harmony (Odden, 1994; Hansson, 2010)

a.		xam-ila	'milk-APPL'
b.		lim-ila	'cultivate-APPL'
c.	kar-ira	*kar-ila	'twist-APPL'
d.	ruk-ira	ruk-ila	'plait-APPL'

- The language's syllable structure is almost exclusively CV
- We can take advantage of this with a tier of all consonants and a tier of just liquids
- Using these, we can distinguish between transvocalic and beyond-transvocalic scenarios

Utility of multiple tiers



Obtaining appropriate weights

- The Slovenian and Bukusu transducers can be recast as acceptors over an alphabet of transducer actions
  - $\Sigma = \{$<s, s>, <s, ʃ>, ...$\}$
- Doing so makes them deterministic
  - Inputs may follow more than one path, but each output possibility is obtained from exactly one path
- Training deterministic probabilistic acceptors is easy (Vidal et al., 2005a,b; de la Higuera, 2010)
  - Read through the sample
  - Count the number of times each transition was followed
  - Divide by the number of times its origin state was visited
- To then ensure that we generate conditional distributions for each input string
  - Normalize the weights by state AND by input letter

Distance-based decay

- In some patterns, the probability of application diminishes exponentially with increasing distance from the trigger
- Such *distance-based decay* (Zymet, 2015) is seen in:
  - Malagasy vowel dissimilation (Zymet, 2015)
  - Hungarian vowel harmony (Hayes and Londe, 2006; Hayes et al., 2009)
  - Latin liquid dissimilation (Zymet, 2015)
  - Navajo sibilant harmony (Martin, 2005)
- In constraint-based frameworks, decay is achieved by scaling constraint weights with distance (Kimper, 2011; Zymet, 2015)
- The same result can be achieved by modifying TSL transducers

Malagasy dissimilation

- Malagasy passive imperative suffix /-u/ dissimilates to [-i] if it would follow [u]
  - /babu-u/ → [babu-i] 'plunder-PASS.IMP'
- Dissimilation is blocked by intervening [i] or [e]
  - /turi-u/ → [turi-u] 'preach-PASS.IMP'
  - /ure-u/ → [ure-u] 'massage-PASS.IMP'
- Dissimilation can occur across the other vowel [a]
  - /gurabah-u/ → [gurabah-i] 'splutter-PASS.IMP'
- Dissimilation applies only to the /u/ in this suffix (Zymet, 2020)

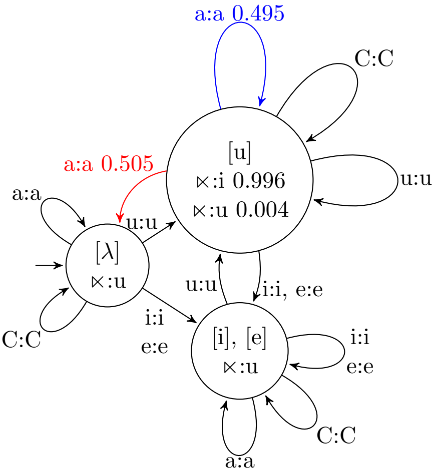
Malagasy dissimilation

- The probability that dissimilation occurs diminishes across multiple syllables
- Token frequency in a dictionary (de la Beaujardière, 2004)

Adjacent syllable	989/993	99%
Across one syllable	201/397	51%
Across two syllables	4/32	13%
Across three syllables	0/4	0%

- Syllable structure is largely CV, so the probability of dissimillation is roughly $1/2^x$ where x is the number of [a]s
- We can model this using a transducer by including both a “transparent” and “opaque” transition for [a]

Malagasy transducer



Deriving the Malagasy weights

- Even when recast as an acceptor, the machine is non-deterministic
- We cannot directly count how many times each transition is traversed by a sample, but we can estimate them
- For transition (q, σ, q') and string $x \cdot \sigma \cdot y$
  - Forward probability = that of being in q after reading x
  - Backward probability = that of producing y starting in q'

Deriving the Malagasy weights

- The count contributed by the reading step is the product of the transition, forward, and backward probabilities, all divided by the string probability (de la Higuera, 2010)
- We can use the estimated counts to calibrate the machine exactly as with a deterministic acceptor
- Repeating the estimation-calibration steps will eventually converge onto a local optimum
  - Baum-Welch algorithm (Baum et al., 1970; Baum, 1972)

Deciding on the transitions

- We opted for a forwards-selection approach
- Assessed the performance of extra transitions using log-likelihood ratio tests
- Round 1:
  - All consonants except [v], [t], [f], and [h] significantly improved performance individually
  - The vowel [a] increased performance the most by far
- Round 2:
  - Only [dʒ] and [z] improved performance
  - Just barely significant, so we opted to only include an extra transition for [a]

Deciding on the transitions

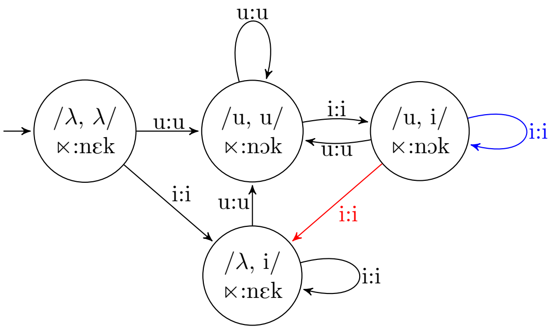
- The significance of most extra consonant transitions indirectly reflected the significance of the extra [a] transition
- Another possibility is that the vowel [a] is much more similar to the tier elements [u], [i], and [e]
- The high similarity makes [a] more likely to interfere with keeping an [u] in memory

	Observed	Predicted
Adjacent syllable	99%	99%
Across one syllable	51%	49%
Across two syllables	13%	24%
Across three syllables	0%	12%

Hungarian vowel harmony

- General description (Hayes and Londe, 2006; Hayes et al., 2009; Kimper, 2011; Ozburn, 2019)
  - Suffix vowels agree in backness with nearest root vowel
  - Non-low, unrounded front vowels are transparent
  - Nearly obligatory across a single transparent vowel, but exponentially less likely across 2+ transparent vowels
- This last fact makes it difficult to accurately model with a quasi-TSL transducer
  - In such a transducer, a transparent segment either always or never has the opportunity to cause "forgetting"
- A 2+ threshold is possible with a quasi-MTSL transducer

The 2+ threshold



- Fragment of a quasi-MTSL transducer that decides between the front and back allomorphs of the dative suffix
- The first *superset* tier includes all vowels
- The second *subset* tier excludes transparent vowels

Training quasi-TSL and quasi-MTSL models

- Corpus of nouns and their proportion of front versus back allomorphs of the dative
  - Collected from Google search results by Hayes and Londe (2006) and Hayes et al. (2009)
- A quasi-TSL and a quasi-MTSL transducer were trained on these data using the Baum-Welch procedure

Comparing quasi-TSL and quasi-MTSL models

- The Malagasy models being compared had nested parameters
- The quasi-TSL and quasi-MTSL models being compared here, however, are not nested
  - This makes it impossible to compare them with log-likelihood ratio tests
- We opted to assess their relative performance using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Results

Transparent syllables	Observed average	TSL average	MTSL average
0	0.99	0.99	0.99
1	0.67	0.64	0.67
2	0.18	0.33	0.19
3	0.00	0.03	0.01

Model	Free params	Log likelihood	AIC	BIC
TSL	7	−266.90	547.8	597.86
MTSL	23	−249.73	545.46	709.94

Conclusion

- Existing transducer models of long-distance phonological processes are strictly categorical
- They are readily modified for probabilistic application
- The resulting quasi-TSL and quasi-MTSL transducers can be trained on real data using the Baum-Welch algorithm
- Quasi-(M)TSL models capture distance-based decay in a cognitively plausible manner
- Quasi-MTSL models can distinguish between spans of 1 and 2+ transparent elements

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada

Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

References I

Andersson, S., Dolatian, H., and Hao, Y. (2020). Computing vowel harmony: The generative capacity of search and copy. In Baek, H., Takahashi, C., and Hong-Lun Yeung, A., editors, *Proceedings of the 2019 Annual Meeting on Phonology*.

Baum, L. E. (1972). An inequality and associated maximization technique occurring in the statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171.

References II

Burness, P. and McMullin, K. (2019). Efficient learning of Output Tier-Based Strictly 2-Local functions. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 78–90. Association for Computational Linguistics.

Burness, P. and McMullin, K. (2020). Multi-tiered strictly local functions. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 245–255. Association for Computational Linguistics.

References III

Burness, P. and McMullin, K. (2021). Learning multiple independent tier-based processes. In Chandlee, J., Eyraud, R., Heinz, J., Jardine, A., and van Zaanen, M., editors, *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of *Proceedings of Machine Learning Research*, pages 66–80. PMLR.

Burness, P., McMullin, K., and Chandlee, J. (2021). Long-distance phonological processes as tier-based strictly local functions. *Glossa*, 6.

Chandlee, J. (2014). *Strictly Local Phonological Processes*. Doctoral dissertation, University of Delaware.

Chandlee, J., Eyraud, R., and Heinz, J. (2014). Learning Strictly Local subsequential functions. *Transactions of the Association for Computational Linguistics*, 2:491–503.

References IV

Chandlee, J., Eyraud, R., and Heinz, J. (2015). Output Strictly Local functions. In *Proceedings of the 14th Meeting on the Mathematics of Language (MOL 2015)*, pages 112–125.

Chandlee, J. and Heinz, J. (2018). Strict Locality and phonological maps. *Linguistic Inquiry*, 49:23–60.

de Blois, K. F. (1975). *Bukusu Generative Phonology an Aspects of Bantu Structure*. Number 85 in Annales. Musée Royal de l'Afrique Centrale, Tervuren.

de la Beaujardière, J.-M. (2004). Malagasy dictionary and encyclopedia of Madagascar.

de la Higuera, C. (2010). *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press, New York.

References V

Hansson, G. Ó. (2010). *Consonant Harmony: Long-Distance Interaction in Phonology*. Number 145 in University of California Publications in Linguistics. University of California Press, Berkeley, CA.

Hao, Y. and Andersson, S. (2019). Unbounded stress in subregular phonology. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology and Morphology*, pages 135–143, Florence, Italy. Association for Computational Linguistics.

Hao, Y. and Bowers, D. (2019). Action-sensitive phonological dependencies. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology and Morphology*, pages 218–228, Florence, Italy. Association for Computational Linguistics.

References VI

Hayes, B. and Londe, Z. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23:59–104.

Hayes, B., Zuraw, K., Siptar, P., and Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85:822–863.

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, OR. Association for Computational Linguistics.

Jurgec, P. (2011). *Feature Spreading 2.0: A Unified Theory of Assimilation*. Doctoral dissertation, University of Tromso.

References VII

Kimper, W. A. (2011). *Competing Triggers: Transparency and Opacity in Vowel Harmony*. Doctoral dissertation, University of Massachusetts Amherst.

Martin, A. (2005). *The Effects of Distance on Lexical Bias: Sibilant Harmony in Navajo Compounds*. Master's thesis, University of California, Los Angeles.

McNaughton, R. and Papert, S. A. (1971). *Counter-Free Automata*. MIT Press, Cambridge, MA.

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

Ozburn, A. (2019). A target-oriented approach to neutrality in vowel harmony: Evidence from Hungarian. *Glossa*, 4:1–36.

References VIII

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

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

Vidal, E., Tollard, F., de la Higuera, C., Casacuberta, F., and Carrasco, R. (2005a). Probabilistic finite-state machines - Part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025.

References IX

Vidal, E., Tollard, F., de la Higuera, C., Casacuberta, F., and Carrasco, R. (2005b). Probabilistic finite-state machines - Part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1026–1039.

Zymet, J. (2015). Distance-based decay in long-distance phonological processes. In Steindl, U., Borer, T., Fang, H., Pard, A. G., Guekguezian, P., Hsu, B., O'Hara, C., and Ouyang, I. C., editors, *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, pages 72–81, Sommerville, MA. Cascadilla Press.

Zymet, J. (2020). Malagasy ocp targets a single affix: Implications for morphosyntactic generalization in learning. *Linguistic Inquiry*, 51:624–634.