Learning phonotactic restrictions on multiple tiers

Kevin McMullin (uOttawa), Alëna Aksënova (Stony Brook), Aniello De Santo (Stony Brook)

Overview We present an algorithm for efficiently learning formal languages containing multiple Tier-based Strictly 2-Local (TSL_2; Heinz et al., 2011) dependencies that operate simultaneously on different tiers. The algorithm does not require a priori knowledge of what the restrictions are, which elements are on the tiers, or how many tiers are required, and we demonstrate its success with learning simulations for a number of complex phonotactic patterns. This constitutes an important advance with respect to the viability of TSL characterizations of linguistic patterns, which have been argued to be relevant for long-distance phonotactics (McMullin, 2016) as well as morphological and syntactic dependencies (Aksënova et al., 2016; Graf, 2017).

Background The grammar of a TSL_2 language is a 2-tuple, $G = \langle T, R \rangle$, where the tier T is a subset of the segment inventory and R is a set of $*xy$ restrictions enforced on the tier (non-tier segments are ignored). TSL_2 grammars can be used to model phonotactic patterns such as sibilant harmony in Tamashek Tuareg (Berber; Heath, 2005), where [s] and [ʃ] may not co-occur at any distance. This pattern holds morpheme-internally, and is further exemplified by alternations of the causative prefix /s(ː)-/ in words such as [**ʃ**-ùkməʃ] 'make scratch!' (cf. *[**s**-ùkməʃ]). The formal grammar for this pattern would ban any sequences of *[sʃ, ʃs] on the tier of sibilants: $G_1 = \langle \{\text{s}, \text{ʃ}\}_T, \{\text{sʃ}, \text{ʃs}\}_R \rangle$. This characterization of long-distance dependencies not only offers a close approximation of the range of attested phonotactic patterns (bounded and unbounded locality, patterns with blocking, exclusion of complex pathologies; McMullin and Hansson, 2016), but the class of TSL_k languages has been proven to be efficiently learnable from positive data for any $k \geq 1$ (Jardine and McMullin, 2017).

However, there are many languages whose phonotactics cannot be captured with a single TSL grammar (McMullin, 2016). Indeed, sibilant harmony in Tamashek Tuareg co-exists with long-distance labial dissimilation, seen in alternations of the agentive prefix /m-/ in, e.g., [ɑ-**n**-ánɑ**m**] 'one who is fond' (cf. *[ɑ-**m**-ánɑ**m**]; Heath, 2005). Though the pattern is TSL_2 on its own, with $G_2 = \langle \{\text{m}\}_T, \{\text{mm}\}_R \rangle$, note that simply combining G_2 with G_1 (from above) into a single grammar $G_3 = \langle \{\text{s}, \text{ʃ}, \text{m}\}_T, \{\text{sʃ}, \text{ʃs}, \text{mm}\}_R \rangle$ would erroneously permit words such as *[**s**-ùk**m**əʃ]. Since there is no restriction against tier-adjacent [sm] or [mʃ], [m] inadvertently acts as a blocker for the unrelated pattern of sibilant harmony.

The overall phonotactics of such languages, which require distinct tiers for each set of restrictions, can instead be modeled with *conjunctions* of multiple TSL_2 grammars ($MTSL_2$), but doing so introduces additional complexity. Moreover, Aksënova and Deshmukh (2018) argue that identifying multiple cooperating grammars poses a significant learnability problem if any conceivable combination of tiers (of which there are $2^{2^{|\Sigma|}}$) is allowed. They therefore hypothesize that the space of (human-)learnable grammars is a reduced set of MTSL grammars, motivating a number of possible restrictions with typological evidence.

Algorithm We will show that the Multiple Tier-based Strictly 2-Local Inference Algorithm (MTSL2IA), presented with pseudocode below, learns an $MTSL_2$ grammar from positive data in polynomial time and data (de la Higuera, 1997). The algorithm exploits the fact that if a bigram $\rho_1\rho_2$ is banned on some tier, then it will never appear in string-adjacent contexts. For each $\rho_1\rho_2$ absent from the training data, the goal is therefore to determine which segments can be safely removed from the associated tier. To do so, the algorithm incorporates the notion of a 2-*path* (Jardine and Heinz, 2016). Intuitively, a 2-path can be thought of as a precedence relation $(\rho_1 \ldots \rho_2)$ accompanied by the set X of symbols that intervene between ρ_1

and ρ_2. Formally, each 2-path is therefore a 3-tuple of the form $\langle \rho_1, X, \rho_2 \rangle$. For example, the string *abcc* includes the following 2-paths: $\langle a, \emptyset, b \rangle, \langle a, \{b\}, c \rangle, \langle a, \{b, c\}, c \rangle, \langle b, \emptyset, c \rangle, \langle b, \{c\}, c \rangle$. In short, by examining the set of 2-paths present in the training data, we can determine which segments are freely distributed with respect to a bigram $\rho_1\rho_2$ that is known to be banned on some tier. Specifically, if all of the attested $\langle \rho_1, X, \rho_2 \rangle$ 2-paths that include an intervening σ are likewise attested *without* an intervening σ, the algorithm removes σ from the tier, since the presence of $\rho_1 \ldots \rho_2$ is not dependent on an intervening σ.

Data: A finite input sample $I \subset \Sigma^*$
Result: MTSL$_2$ grammar of the form $G = \bigwedge \langle T_i, R_i \rangle$
Initialize $B =$`bigrams`$(\Sigma^*) -$`bigrams`(I);
foreach $\rho_1\rho_2 \in B$ **do**
  Initialize $R_i = \rho_1\rho_2$, $T_i = \Sigma$; (note: i ranges from 1 to $|B|$)
  foreach $\sigma \in \Sigma - \{\rho_1, \rho_2\}$ **do**
    if $\forall \langle \rho_1, X, \rho_2 \rangle \in$ `2-paths`(I) *s.t.* $\sigma \in X, \langle \rho_1, X - \{\sigma\}, \rho_2 \rangle \in$ `2-paths`(I)
    then $T_i = T_i - \{\sigma\}$ (i.e., remove σ from T_i);
  end
  $G_i = \langle T_i, R_i \rangle$
end
Return $G = G_1 \wedge G_2 \wedge ... \wedge G_n$

Algorithm 1: The Multiple Tier-based Strictly 2-Local Inference Algorithm

Discussion As illustrated by the Tamashek Tuareg example, conjunctions of TSL grammars are necessary if we want to account for the phonotactics of natural languages. We present computational simulations demonstrating that the MTSL2IA succeeds on a number of complex phonotactic patterns. However, we note that the algorithm relies on the assumption that each bigram restriction is enforced on at most one tier. A small portion of logically-possible MTSL patterns therefore remain out of reach at present, but the problematic cases are among those which Aksënova and Deshmukh (2018) claim to be unattested (those with overlapping tiers, such that $T_1 \not\subseteq T_2$ and $T_1 \cap T_2 \neq \emptyset$). Specifically, the MTSL2IA fails if these overlapping tiers are associated with a single $^*\rho_1\rho_2$ restriction (i.e., when it is blocked by a different symbol on each tier), but it will succeed when they are associated with different restrictions. While a more careful analysis of the learning space is required, our results further point in the direction of a sub-class of long-distance dependencies more restricted than the entire range of TSL conjunctions. Most importantly, the learner presented above succeeds on the types of patterns that have eluded other algorithms (Jardine and Heinz, 2016; Jardine and McMullin, 2017), and is thus a concrete step toward a comprehensive, efficient subregular learner for long-distance dependencies attested in natural language.

Aksënova, A. and Deshmukh, S. (2018). Formal restrictions on multiple tiers. In *Proceedings of SCiL 2018*. • **Aksënova, A., Graf T., and Moradi S.**. (2016). Morphotactics as tier-based strictly local dependencies. In *Proceedings of SIGMorPhon 2016*. • **Graf, Thomas**. (2017). Why movement comes for free once you have adjunction. In *Proceedings of CLS 53*. • **Heath, J.** (2005). *A grammar of Tamashek (Tuareg of Mali)*. Mouton de Gruyter, Berlin. • **Heinz, J., Rawal, C., and Tanner, H. G.** (2011). Tier-based strictly local constraints for phonology. In *Proceedings of ACL 49th* , Portland, OR. • **de la Higuera, C.** (1997). Characteristic sets for polynomial grammatical inference. *Machine Learning*, 27:125-138. • **Jardine, A. and Heinz, J.** (2016). Learning tier-based strictly 2-local languages. *TACL*, 4:87–98. • **Jardine, A. and McMullin, K.** (2017). Efficient learning of Tier-based Strictly k-Local languages. In *Proceedings of LATA 11th*. • **McMullin, K.** (2016). *Tier-based locality in long-distance phonotactics: learnability and typology*. PhD thesis, U. of British Columbia. • **McMullin, K. and Hansson, G. Ó.** (2016). Long-distance phonotactics as Tier-based Strictly 2-Local languages. In *Proceedings of AMP 2014*.