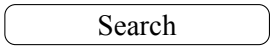
The typology of locality guides restrictive and accurate tier induction

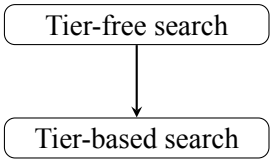
Seoyoung Kim (seoyoungkimk@umass.edu)
University of Massachusetts, Amherst

Contrary to the learner of Hayes and Wilson (2008) where tiers need to be pre-defined and supplied by the analyst (a), Gouskova and Gallagher (2020) introduce the *Inductive Projection Learner* (henceforth, IPL) where tiers are automatically discovered from local trigrams (b). I report two case studies to show that the performance of the IPL can be improved by adding an intermediate step, *Evaluation* (c), in which the necessity and the accuracy of the candidate tiers are evaluated using a novel heuristic derived from the typology of locality.

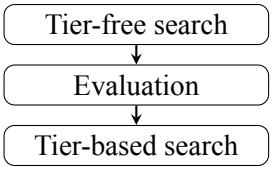
a. Hayes and Wilson (2008)



b. Gouskova and Gallagher (2020)



c. This paper



Typology: There is a robust dichotomy between trigram-bound and unbounded patterns in long-distance interactions. McMullin (2016) shows that all cases of sibilant or nasal harmony are either transvocalic (CVC) or unbounded (C...C). Similarly, McCollum (2019) points out that vowel harmony can be either non-iterative where only a single vowel is harmonized (VCV) or fully iterative throughout the domain (V...V). Finley (2011) and McMullin and Hansson (2019) show that interactions at unbounded distance entail trigram-bound interactions, but trigram-bound interactions do not always generalize to unbounded distances. Thus, a transvocalic or non-iterative restriction is assumed to expand unboundedly once it holds over another consonant or vowel, respectively. Put differently, checking whether a pattern observed as a local trigram holds over an extra segment makes a good heuristic for checking its unboundedness.

The IPL first discovers a tier-free grammar. A trigram with a medial placeholder (henceforth, [ ]), which refers to any segment of the language, is a hint that classes on either side interact non-locally regardless of the identity of the placeholder (e.g., *[+lateral][ ][+lateral]). If the tier-free grammar includes such trigrams, the learner projects the smallest natural class that includes both sides (e.g., [+lateral]) and discovers tier-based constraints on it. Regardless of whether the restriction is transvocalic or unbounded, the IPL will automatically project a tier for every placeholder trigram, which is unnecessary if the pattern is merely trigram-bound. If the restriction is unbounded but has blockers (such as *r*; ***l**a**s**i**l***, ✓***l**a**r**i**l***), the tier will exclude these blockers and inaccurately rule out grammatical sequences. As illustrated in the figure on the right, the blocker (*r*) should also be projected, in order to rule out *lasil and not ✓laril.

l tier	*l l	*l l
l, r tier	*l l	✓l r l
Baseline	* lasil	✓laril

Evaluation I improve the performance of IPL by temporarily projecting either a consonantal or vocalic tier, depending on the placeholder trigram; the example above is a consonantal interaction and thus a C tier is temporarily projected. In order to determine whether the pattern observed as a placeholder trigram expands unboundedly, I re-evaluate the same trigram (*[+lateral][ ][+lateral]) on a different level: on the temporary C tier. If forms like ✓***l**a**s**i**l*** and ✓***l**a**r**i**l*** are not underattested, meaning that the restriction does not hold over another consonant and therefore is only transvocalic, *[+lateral][ ][+lateral] will be weighted lowly because the medial placeholder refers to any consonant on this C tier. In this case, no tier will be projected from this trigram. Conversely, if

forms like ***l**a**s**i**l*, ***l**a**r**i**l* are underattested, indicating that the pattern holds over an extra consonant and is therefore unbounded, the trigram will be weighted highly on the C tier, in which case a [+lateral] tier will projected from *[+lateral][ ][+lateral]. Detecting blockers requires more specific representation of the placeholder because it is crucial to know specifically which segments, rather than any consonant (or any vowel), lift the restriction. Thus, instead of a single trigram, I re-evaluate *a set of* trigrams where the middle placeholder is replaced by every consonant (or vowel) of the language. If forms like ***l**a**C**i**l*** (such as ***l**a**s**i**l***, ***l**a**b**i**l***, ***l**a**k**i**l***, etc) are underattested while ✓***l**a**r**i**l*** are not, indicating that the otherwise illegal sequences *l...l are allowed with an intervening *r*, *[+lateral]C[+lateral] with any consonant should be weighted highly unless the C is a *r*. In this case, the blocker *r* will be projected along with the interacting classes ([+lateral]) of the trigram.

Evaluating necessity Lamba, a Bantu language, exhibits transvocalic nasal harmony in a form of alternation (Odden 1994); the suffixes /-ile/ and /-ele/ surface as [-ine] and [-ene] after a nasal within a trigram window, as in [**n**-i**n**e] 'drink-perf' *vs.* [**m**as-i**l**e] 'plastered-perf'. To capture this pattern, a local trigram *[+nasal][ ][+lateral] is sufficient and a tier is unnecessary. I ran the IPL on a Lamba sublexicon (1,293 words ending in *-ile*, *-ele*, *-ine*, *-ene*, gain 10, gamma 3). A placeholder trigram that can penalize part of nasal and lateral co-occurrences, *[+cor,−lateral][ ][+cor,+voice,−nas] ($w = 10.8$, *{n}[ ]{d, ʒ, l}), was learned in the tier-free search at best, because trigram-bound patterns are inherently hard to find if there are lots of long words. Notably, this constraint automatically led to projecting an unnecessary [+coronal+voice] tier. Instead, I reweighted this trigram on the C tier, in order to determine whether a tier should be projected from it. The constraint was weighted 0, which indicates that a subsequence of *n* and {*d, ʒ, l*} is not prohibited over a consonant in the learning data. Based on the typological observation, this result can be interpreted as that the restriction does not generalize unboundedly and there does not have to be a tier projected from this trigram; the evaluation step successfully prevented the unnecessary tier from being projected.

Evaluating accuracy In Shona, a high vowel cannot follow mid vowels, unless *a* intervenes (**ei*, **oi vs.* ✓*eai*, ✓*oai*; Beckman 1997). I trained the IPL on Shona (4,688 verbal stems, gain 190), and the trigram *[−high,−low][ ][+high,−back] led to a projection of [−low]. In the tier-based search, *[−high][+syll,+high,−back] ($w = 3.7$, *{e, o}[ ]∗{i}) was learned on [−low], which incorrectly rules out legal forms on [−low], as shown in the figure below. There were four placeholder trigrams in the tier-free search in total. I replaced the placeholder middle grams with every segment that will be visible on the intermediate vowel tier: all 5 vowels of Shona. Among the 20 constraints (4×5) that were reweighted, *[−high,−low][+low][+high,−back] (*{e, o}{a}{i}) was singled out, weighted significantly lower than the other 19 constraints. The result accurately captures the fact that the restriction *[−high,−low][ ][+high,−back] does not hold over a blocker [a], indicating that not only the non-low vowels but also the low vowel should be projected; evaluation successfully projects the blocking segment on the tier.

[−low] tier	*e i	o*o i
V tier	e a i a	o o a i a
Baseline	✓ʧejamisa	✓pofomadzira

Conclusions I argued that the candidate tiers suggested by local trigrams in the IPL should be further validated before being projected. First, the pattern that is observable as a trigram might not generalize outside the trigram window, in which case a tier-based constraint is unnecessary. Second, the candidate tier might exclude blockers. I exploit a typological observation as a heuristic to aid in determining the necessity and accuracy of the candidate tiers: whether the pattern holds over another segment or a subset of segments. Through learning simulations on Lamba and Shona, I showed that tiers can be induced more restrictively and accurately by adding an Evaluation step.