Morphotactics as Tier-Based Strictly Local Dependencies

Alëna Aksënova Thomas Graf Sedigheh Moradi

Department of Linguistics
Stony Brook University
Stony Brook, NY 11794, USA

*mail@thomasgraf.net
alena.aksenova@stonybrook.edu, sedigheh.moradi@stonybrook.edu

Abstract

It is commonly accepted that morphological dependencies are finite-state in nature. We argue that the upper bound on morphological expressivity is much lower. Drawing on technical results from computational phonology, we show that a variety of morphotactic phenomena are tier-based strictly local and do not fall into weaker subclasses such as the strictly local or strictly piecewise languages. Since the tier-based strictly local languages are learnable in the limit from positive texts, this marks a first important step towards general machine learning algorithms for morphology. Furthermore, the limitation to tier-based strictly local languages explains typological gaps that are puzzling from a purely linguistic perspective.

1 Introduction

Different aspects of language have different levels of complexity. A lot of recent work in phonology (see Graf (2010), Heinz (2011a; 2011b; 2015), Chandlee (2014), Jardine (2015) and references therein) argues that phonological well-formedness conditions are subregular and hence do not require the full power of finite-state automata. This is particularly noteworthy because computational phonology still relies heavily on finite-state methods (Kaplan and Kay, 1994; Frank and Satta, 1998; Riggle, 2004). A similar trend can be observed in computational syntax, where the original characterization as mildly context-sensitive string languages (Huybregts, 1984; Shieber, 1985) is now being reinterpreted in terms of subregular tree languages (Graf, 2012; Graf and Heinz, 2015). Curiously missing from these investigations is morphology.

While linguistic theories sometimes consider morphology a part of syntax, computational morphology recognizes that the weak generative capacity of morphology is much closer to phonology than syntax. Consequently, computational morphology involves largely the same finite-state methods as computational phonology (Koskenniemi, 1983; Karttunen et al., 1992). This raises the question whether morphology, just like phonology, uses only a fraction of the power furnished by these tools. A positive answer would have important repercussions for linguistics as well as natural language processing. The subregular classes identified in computational phonology are learnable in the limit from positive text (Heinz et al., 2012), so a subregular theory of morphology would greatly simplify machine learning while also explaining how morphological dependencies can be acquired by the child from very little input. A subregular model of morphology would also be much more restricted with respect to what processes are predicted to arise in natural languages. It thus provides a much tighter typological fit than the regular languages. In this paper, we argue that the subregular view of morphology is indeed correct, at least for morphotactics.

Morphotactics describes the syntax of morphemes, that is to say, their linear order in the word and the conditions that license their presence or enforce their absence. One can distinguish *surface morphotactics* from *underlying morphotactics*. The former regulates the shape of the pronounced surface strings, whereas the latter is only concerned with the arrangements of the morphemes in the initial representation rather than how said morphemes are realized in specific environments. We only consider underlying morphotactics in this paper.

The following example may clarify the distinction. In German, the past participle of a verb is

Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 121–130,
Berlin, Germany, August 11, 2016.

formed via a circumfix. The first part of the circumfix is always the prefix *ge-*, whereas the second part may be the suffix *-en* or *-t* depending on the verb stem. In addition, the suffixes can also occur on their own, e.g. on infinitives or the third person singular form of the verb. Surface morphotactics thus has to ensure that *ge-* always appears with one of these two suffixes, and that the form of the suffix matches the stem. At the same time, it does not need to worry about matching *-en* or *-t* with *ge-* since these forms can occur independently as realizations of different morphemes. Underlying morphotactics, on the other hand, is unaware of the surface realizations and only knows that some abstract prefix *GE-* must always occur with the abstract suffix *-EN*, and the other way round. The fact that *-EN* has a surface realization that is indistinguishable from the infinitival marker, which does not require the prefix *GE-*, is irrelevant for underlying morphotactics. More succinctly: underlying morphotactics regulates the distribution of morphemes, surface morphotactics the distribution of allomorphs.

This paper considers a variety of phenomena — circumfixation, variable affix ordering, unbounded prefixation — and concludes that they all belong to the class of tier-based strictly local languages. We first show that even though many morphotactic dependencies are strictly local, that is not the case for all of them (Sec. 2.1). While some of these outliers are strictly piecewise (Sec. 2.2), tier-based strictly local grammars are needed to handle the full range of data points (Sec. 2.3). This prompts our conjecture that all dependencies that are part of underlying morphotactics stay within the class of tier-based strictly local languages. We then use this hypothesis in Sec. 3 to explain two typological gaps with respect to compounding markers and circumfixation.

2 Subregular Patterns in Morphology

The regular languages are one of the best understood language classes, with many attractive properties. Yet it is often forgotten that this class properly includes many weaker ones (McNaughton and Pappert, 1971), some of which have recently attracted much interest in computational phonology. At the very bottom of the hierarchy one finds strictly local and strictly piecewise languages (Rogers et al., 2010), and a little bit higher up the tier-based strictly local languages (Heinz et al., 2011). The subregular hierarchy includes many other classes (see Fig. 1), but the previous three are noteworthy because they are conceptually simple and efficiently learnable in the limit from positive data (Heinz et al., 2012; Jardine and Heinz, 2016) while also furnishing sufficient power for a wide range of phonological phenomena (Heinz, 2015; Jardine, 2015).

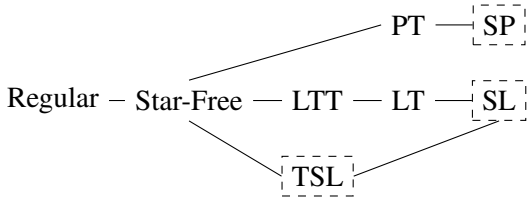

Figure 1: The subregular hierarchy as given in Heinz et al. (2011); language classes in dashed boxes are studied in this paper

In this section, we investigate the role of strictly local, strictly piecewise and tier-based strictly local patterns in morphotactics. We show that some but not all patterns are strictly local or strictly piecewise, whereas all typologically instantiated patterns seem to fit in the class of tier-based strictly local languages.

2.1 Strictly Local

A string language L over alphabet Σ is *strictly local* (SL) iff there is some $k \in \mathbb{N}$ such that L is generated by a strictly k-local grammar G. Such a grammar consists of a finite set of k-grams, each one of which describes an illicit substring. More precisely, given a string $w \in \Sigma^*$, let $\hat{w}^k := \rtimes^k w \ltimes^k$ (where $\rtimes, \ltimes \notin \Sigma$) and $k\text{-grams}(w) := \{s \mid s \text{ is a substring of } \hat{w}^{k-1} \text{ of length } k\}$. Then G generates string w iff $k\text{-grams}(w) \cap G = \emptyset$. That is to say, G generates every string over Σ that does not contain an illicit substring.

Most phonological dependencies can be described in strictly local terms — see Heinz (2015) for numerous examples. Consider for instance the well-known process of *word-final obstruent devoicing* that forces voiced obstruents at the end of the word to be realized as voiceless: moro**z** [maros] 'frost' in Russian, Ba**d** [ba**t**] 'bath' in German). If one considers phonotactics rather than mappings from underlying representations to surface forms, this is tantamount to a ban against word-final voiced obstruents. Said ban, in turn, is captured by a strictly 2-local grammar G that contains all bigrams of the form $v\ltimes$, with v a voiced

obstruent.

The specification of SL grammars can be simplified by applying mappings. In the case at hand, one could define a function f that replaces every voiced obstruent by the designated symbol $\Diamond$ so that the grammar G can be reduced to the single bigram $\Diamond\ltimes$. One has to be careful, though. The SL languages are not closed under relabelings, in fact, every regular language is the image of a strictly 2-local language under some relabeling. However, the directionality of the $\Diamond$-relabeling above is the opposite: first the relabeling is applied, and then the grammar filters out strings in the image of that relabeling. As long as the relabeling is a many-to-one map between alphabets (and thus does not introduce distinctions that weren't already part of the original alphabet), this provably does not increase weak generative capacity for any of the formalisms discussed in this paper.

We make use of such relabelings in the following sections in order to convert natural language patterns into more easily described formal languages. For morphotactics, though, this raises the issue how the size of the atomic units should be chosen. One could posit that morphology, just like phonology, treats every phonological segment as a symbol. In that case, stems and morphemes are strings of symbols. Alternatively, one may treat each morpheme, including stems, as an atomic symbol. This is an important decision when it comes to modeling the interactions of morphology and phonology such as phonologically conditioned allomorphy. Fortunately our results are independent of this choice, due to the productive nature of compounding.

To better understand why different representations could in principle affect subregular complexity, note first that whether a stem is represented as a single, atomic symbol or as a sequence of phonological segments seems to determine if prefixes and suffixes might be separated by an unbounded amount of symbols. Consider a circumfix *u- -v*, where neither part of the affix may occur without the other. A concrete example is the nominalization circumfix *ke- -an* in Indonesian (Mahdi, 2012; Sneddon, 1996):

(1) a. tingii
high

b. ke- tinggi -an
NMN- high -NMN
'altitude'

If a stem is a single symbol x, then x and uxv are well-formed whereas ux and xv are not due to *u- -v* being a circumfix whose subparts cannot occur in isolation. This generalization is easily captured by the strictly 3-local grammar $\{\rtimes xv, ux\ltimes\}$. However, if stems are sequences of symbols, then the well-formed patterns are of the form x^+ or ux^+v (since the length of stems is in principle unbounded). The illict strings, on the other hand, are of the form $\rtimes x^+v$ and $ux^+\ltimes$. But no strictly local grammar can generate the former patterns without also generating the latter. That is due to the strictly local languages being closed under suffix substitution closure.

Suffix Substitution Closure Language L is SL iff there exists a $k \in \mathbb{N}$ such that for all strings u_1, v_1, u_2, v_2 and any string x of length $k-1$, if u_1xv_1, $u_2xv_2 \in L$, then $u_1xv_2 \in L$.

If there is no upper bound on the length of stems, then we can infer from $x^k \in L$ and $ux^kv \in L$ that both $x^kv \in L$ and $ux^k \in L$. It seems, then, that circumfixes are strictly local only if each stem is an atomic symbol.

But this line of reasoning erroneously assumes that the circumfix can only apply to individual stems, which ignores the availability of compounding. Returning to Indonesian, we see that its nominalization marker is not restricted to single stems and can also apply to compounds.

(2) a. maha siswa
big pupil
'student'

b. ke- maha siswa -an
NMN- big pupil -NMN
'student affairs'

Compounding is an unbounded process, so even if each stem is mapped to a single symbol x, one ends up with the same patterns as with the segmental mapping approach: x^+ and ux^+v are well-formed, while ux^+ and x^+v are ill-formed. Since the choice of representation does not affect the subregular complexity results, we opt for the segmental mapping, which does not require us to use compounding in all our natural language data points.

The details of the segmental mapping are as follows: within a stem, all segments are replaced by some distinguished symbol. We choose x for

this purpose. All morphemes, on the other hand, are replaced by single symbols. Symbols are chosen to maximize clarity of exposition, so that we sometimes assign each morpheme a unique symbol and sometimes map irrelevant morphemes to a randomly chosen filler symbol. For some linguistic phenomena we follow linguistic convention in assuming that the underlying representations contain additional distinguished symbols to mark the edges of the stem — this will be mentioned explicitly for all relevant cases.

The preceding discussion yielded as a nice side-result that circumfixation is not SL unless each part of the circumfix can also occur on its own. Few circumfixes display that kind of freedom, wherefore not all aspects of morphotactics are SL. However, large swaths of morphology still are, with a couple of examples from English given below:

(3) a. un- do
a- xx

b. break -able
xxxxx -b

(4) a. punch -ed
xxxxx -c

b. put -ε
xxx -c

Any kind of affix that only consists of one part and whose distribution is determined within a locally bounded domain is part of strictly local morphotactics. Although we did not carry out any rigorous quantitative comparisons, we believe the majority of morphological dependencies to belong to this class.

2.2 Strictly Piecewise

While SL covers a wide range of phenomena, it isn't just circumfixes that require more power. Problems arise whenever a dependency involves both the domain of prefixes and the domain of suffixes — because they can be separated by arbitrarily many symbols — and such configurations are not limited to circumfixes. In most languages the ordering of affixes tends to be fixed, but there are languages in which affixes are ordered relatively freely and do not follow a strict template, thereby creating non-local dependencies.

Let us consider the following data from Swahili:

(5) a. a- vi- soma
SBJ:CL.1- OBJ:CL.8- read
-vyo
-REL:CL.8
u-v-xxxx-c
'reads'

b. a- si- vyo- vi-
SBJ:CL.1- NEG- REL:CL.8- read
soma
-OBJ:CL.8
u-w-c-v-xxxx
'doesn't read'

This data is taken from Stump (2016). Based on his discussion of *vyo*, the following forms are ungrammatical.

(6) a. * a- vyo- vi-
SBJ:CL.1- REL:CL.8- OBJ:CL.8-
soma
read
u-c-v-xxxx

b. * a- vyo- vi-
SBJ:CL.1- REL:CL.8- OBJ:CL.8-
soma -vyo
read -REL:CL.8
u-c-v-xxxx-c

c. * a- si- vyo-
SBJ:CL.1- NEG- REL:CL.8-
vi- soma -vyo
OBJ:CL.8- read REL:CL.8-
u-w-c-v-xxxx-c

d. * a- si- vi- soma
SBJ:CL.1- NEG- OBJ:CL.8- read
-vyo
REL:CL.8-
u-w-v-xxxx-c

Different generalizations can be drawn from these data sets, some of which are more complex than others.

The first generalization states that *vyo* is only licensed if it follows either a segment that is part of a stem or the prefix *si*. This is a strictly 2-local pattern, and it explains both (6a) and (6b). Alternatively, one may conclude that (6b) is ill-formed because there is more than one occurrence of *vyo*. Such a ban against two instances of *vyo* is also supported by the ill-formedness of (6c), which is unexpected under the first generalization. Preventing the presence of two instances of *vyo* is beyond the power of any SL grammar G: if $uvx^+c \subset L(G)$

and $uwcvx^+ \subset L(G)$, then $L(G)$ must also contain strings of the form $uwcvx^+c$ (due to suffix substitution closure).

The second generalization is similar to the phonological requirement that no word may contain more than one primary stress, which is *strictly piecewise* (SP; Heinz (2010), Rogers et al. (2010)). SP grammars work exactly the same as SL grammar except that instead of illicit substrings they list illicit subsequences. Given a string w, its set of k-sequences is $k\text{-seqs}(w) := \{s \mid s \text{ is a subsequence of } \hat{w}^{k-1} \text{ of length } k\}$. A strictly k-piecewise grammar G is a finite set of k-grams over $\Sigma \cup \{\rtimes, \ltimes\}$, and the language generated by G is $L := \{w \mid k\text{-seqs}(w) \cap G = \emptyset\}$.

The ban against two occurrences of *vyo* is strictly 2-piecewise — the grammar only need to contain the bigram *vyo vyo*. The intersection of the strictly 2-local and strictly 2-piecewise languages does not contain (6a)–(6c), as desired. But it does contain (6d). Both generalizations miss that even though *vyo* can occur as a prefix and as a suffix, it is a prefix if and only if *si* is present. This kind of conditional positioning cannot be captured by SL grammars, and the culprit is once again suffix substitution closure. But SP grammars by themselves are not sufficient, either.

Suppose we increase the locality rank from 2 to 3 and include *si x vyo* as an illicit subsequence in our SP grammar. This forces *vyo* to be a prefix in the presence of *si*. However, it still incorrectly allows for *vyo* to be a prefix in the absence of *si*. No SP grammar can prevent this outcome. The problem is that any word of the form *u vyo v x* contains only subsequences that also occur in the well-formed *u si vyo v x*. Consequently, any SP grammar that generates the latter also generates the former. It is only in combination with the SL grammar that we can correctly rule out prefix *vyo* without a preceding *si*. Swahili's inflectional morphology thus provides evidence that SL is not enough to handle all aspects of morphotactics and must be supplemented by some mechanism to handle long-distance dependencies, with SP being one option.

But even the combination of SL and SP cannot capture all non-local dependencies. In Swahili, the inability of SP mechanisms to enforce the presence of *si* with prefix *vyo* could be remedied by the strictly local requirement that *vyo* may only occur after *si* or a stem. This elegant interaction of SL and SP is not always possible, however. The most noteworthy case are circumfixes. Consider some arbitrary circumfix *u- -v*. Clearly all subsequences of ux^+ are subsequences of ux^+v, so if the latter is generated by some SP grammar then by definition the former must be, too. The underlying problem is that SP grammars can only enforce the absence of an affix, not its presence. Circumfixes where the presence of one affix entails the presence of the other affix thus are not SP. It seems that we must move higher up the subregular hierarchy in order to accommodate circumfixes, which will also have the welcome side-effect of providing a simpler account for the distribution of Swahili *vyo*.

2.3 Tier-Based Strictly Local

As pointed out in the previous section, the Swahili pattern isn't too dissimilar from the phonological requirement that no word has more than one primary stress. However, the distribution of primary stress is more specific than that: every phonological word has exactly one primary stress. Ensuring the presence of at least one primary stress is beyond the capabilities of SP grammars — once again this holds because every subsequence of an ill-formed word without primary stress is also a subsequence of the well-formed counterpart with exactly one primary stress. A better model for primary stress assignment is furnished by *tier-based strictly local* (TSL; Heinz et al. (2011)) grammars, which also happen to be powerful enough for circumfixation.

A TSL grammar is an SL grammar that operates over a *tier*, a specific substructure of the string. Given a tier-alphabet $T \subseteq \Sigma$, let E_T be a mapping that erases all symbols in a string that do not belong to T. First, $E_T(\varepsilon) = \varepsilon$. Then for $a \in \Sigma$ and $w \in \Sigma^*$,

$$E_T(aw) := \begin{cases} a \cdot E_T(w) & \text{if } a \in T \\ E_T(w) & \text{otherwise} \end{cases}$$

The T-tier of a string w is its image under E_T. A tier-based strictly k-local grammar G is a pair $\langle K, T\rangle$ where K is a strictly k-local grammar over tier-alphabet T. The grammar generates the language $L(G) := \{w \mid E_T(w) \in L(K)\}$. Note that every SL language is a TSL language with $T = \Sigma$.

The distribution of primary stress is tier-based strictly 2-local. Assuming that primary stress is indicated as some diacritic on symbols, the tier-alphabet T contains all symbols with this diacritic.

This is tantamount to projecting a tier that only contains segments with primary stress. The grammar then contains the bigram $\rtimes\ltimes$ to block words with an empty primary stress tier, i.e. words that contain no primary stress. In addition, every bigram uv for $u, v \in T$ is added to rule out words with more than one primary stress. The requirement of exactly one primary stress per word thus boils down to having exactly one segment on the primary stress tier, which is a strictly local dependency over that tier.

The Swahili pattern from the previous section can also be analyzed as tier-based strictly local, and the same is true for circumfixation. For Swahili we project a tier that contains only the affix *vyo*, and we do not allow more than one segment on this tier. As a result, *vyo* occurs at most once per word. To ensure that *vyo* is a prefix whenever *si* is present, we furthermore postulate a marker # that indicates the edges of the stem. The projected tier then includes all instances of *vyo*, *si* and the marker # (of which there are exactly two). On this tier, the 4-gram *si*##*vyo* correctly excludes all ill-formed cases of *vyo* as a suffix, whereas $\rtimes$*vyo*## prevents *vyo* from occurring as a prefix in the absence of *si*. Adapting the ban against two instances of *vyo* to this slightly expanded tier is left as an exercise to the reader.

In order to regulate the distribution of circumfixes such as *u- -v*, we have to project a tier that contains only those subparts *u* and *v*. If the affixes can never occur by themselves, then we block $\rtimes v$ and $u\ltimes$. Removing one of those two bigrams creates asymmetric cases where one of the affixes — but not the other — is sometimes allowed to be present by itself. We also add uu and vv to block strings where the prefix parts outnumber or are outnumbered by the suffix parts of the circumfix. Note that this has the side effect of also prohibiting unbounded circumfixation, a point we return to in Sec. 3.2.

At this point, we can safely say that natural language morphotactics is at least TSL (barring the discovery of any intermediate classes between SL and TSL, or SP and TSL). SL is sufficiently powerful for large parts of morphology, but any kind of dependency that involves both prefixes and suffixes is likely not SL. Some patterns that are not SL are SP, but these also turn out to be TSL. To the best of our knowledge, there are no morphological dependencies that are SP but not TSL (even though the two language classes are incomparable). We thus put forward the following conjecture:

Subregular Morphotactics All morphotactic dependencies are tier-based strictly local.

As any universal claim about the real world, our conjecture cannot be proved conclusively — the fact that no counterexamples have been encountered does not guarantee that counterexamples will never be encountered. But there are additional reasons to consider TSL a better fit for morphotactics than one of the more powerful classes.

Moving beyond TSL in the subregular hierarchy would take us into the class of star-free languages, which are equivalent to the string sets definable in first-order logic with the transitive closure of the successor relation. As mentioned before, every language that is generated by a tier-based strictly k-local grammar can be identified in the limited from positive text, provided the learner knows the value of k. The class of star-free languages, on the other hand, is not learnable in the limit from positive text. It also makes largely incorrect typological predictions: Unlike TSL, the class of star-free languages is closed under union and relative complement. But the union or relative complement of two morphotactic systems attested in natural languages rarely yields linguistically plausible morphotactics. Similarly, it is trivial to write first-order formulas for highly unnatural patterns, e.g. that every word containing two instances of a but less than three bs must contain no substring of the form cd^+c. These points show that moving from TSL to star-free means losing essential properties of natural language morphotactics.

Future work may of course identify more adequate classes in the vicinity of TSL. Given our current, more limited knowledge of the subregular hierarchy, however, the strongest empirically defensible stance is that tier-based strict locality is both sufficient and necessary for natural language morphotactics.

3 Beyond Tier-Based Strictly Local?

If the subregular hypothesis is correct, then no morphological pattern may exceed the computational power furnished by tier-based strictly local grammars. In particular, whenever the combination of two attested TSL patterns is not TSL, that combination should not be attested. The subreg-

ular hypothesis thus provides a principled explanation for typological gaps. In this section we consider two such cases related to compounding markers and the limits of circumfixation.

3.1 Case Study 1: Compounding Markers

Compounding describes the combination of two or more stems to form a compound lexeme, where the stems may belong to different categories. Languages differ with respect to whether compounding is (at least sometimes) explicitly marked. In the following we exhibit two TSL compounding patterns from Turkish and Russian whose combination is not typologically attested. We then explain why this combined pattern is not TSL, deriving the otherwise puzzling typological gap.

Turkish possessive compounds (see Aslan and Altan (1998) for a detailed description) obey the general pattern *stem-stem*$^+$*-o*, where *o* stands for the compounding marker *-sı*.

(7) a. bahçe kapı -sı
garden gate -COMP
xxxxx-xxxx-o
'garden gate'

b. türk kahve -sı
turkish coffee -COMP
xxxx-xxxxx-o
'Turkish coffee'

c. türk bahçe kapı -sı (*-sı)
turkish garden gate -COMP (*-COMP)
xxxx-xxxxx-xxxx-o(-*o)
'Turkish garden gate'

The compounding marker is added when two stems are combined. Addition of further stems does not increase the number of compounding markers, it is always a single marker for the whole word. The resulting pattern *stem-(stem*$^+$*-o)* is tier-based strictly 2-local under the assumption that a designated symbol occurs between stems, say $\square$. For then we can project a tier that contains only $\square$ and *o*, with the only illicit bigrams on this tier being $\rtimes o$, $o\square$, and $\square\ltimes$.

Russian compounding, on the other hand, follows the pattern *(stem-o)***-stem*, which means that the addition of a new stem to the compound requires the appearance of the compounding marker *-o-* between the stems:

(8) a. vod -o- voz
water -COMP- carry
xxx-o-xxx
'water-carrier'

b. vod -o- voz -o- voz
water -COMP- carry -COMP- carry
xxx-o-xxx-o-xxx
'carrier of water-carriers'

Assuming once again the presence of the special symbol # — which marked the edges of stems in the previous section — we can show this pattern to also be tier-based strictly 2-local. In this case, the illicit bigrams are ##, *oo*, $\rtimes o$, and $o\ltimes$. Observe that we can remove the first one of these bigrams to allow for cases where the compounding marker is optional.

One may wonder now whether it is possible for natural languages to display a combination of the compounding patterns seen with Russian and Turkish. From a linguistic perspective, the expected answer is yes. If compounding can be marked by a suffix as in Turkish, and compounding can introduce a marker with each step as in Russian, then it should be possible to introduce a suffix with each compounding step. But to the best of our knowledge, no language instantiates this system. From a computational perspective, on the other hand, this typological gap is expected because the described system is not TSL — as a matter of fact, it isn't even regular.

A language L that suffixes a marker to a compound with each compounding step would produce compounds where the number of compound markers is proportional to the number of stems. Let h be a map that replaces all stems by s, all compound markers by o, and all other material by some other symbol. Intersecting $h(L)$ with the regular language s^+o^+ yields the language $s^m o^n$, $m > n$. This string set is easily shown to be context-free (e.g. via the Myhill-Nerode theorem), and since regular languages are closed under homomorphisms and intersection, it follows that L cannot be regular. But every TSL language is regular, so the combination of Russian and Turkish outlined above is not TSL. The absence of this compounding pattern in the typology of natural languages thus lends further support to our conjecture that natural language morphotactics is limited to TSL dependencies.

3.2 Case Study 2: Unbounded Affixation

Circumfixation already played an important role in motivating TSL as a reasonable lower bound on how much power is required for natural lan-

guage morphotactics. We now show that just like compounding markers, circumfixation also suggests that TSL marks the upper bound on required expressivity. In particular, unbounded affixation is widely attested across languages, whereas unbounded circumfixation is not.

A number of languages allow some of their affixes to occur multiple times. For instance, the Russian temporal prefix *posle-* can appear iteratively in the beginning of a word like *zavtra* 'tomorrow'.

(9) a. posle- zavtra
AFTER- tomorrow
a-xxxxxx
'the day after tomorrow'

b. posle- posle- zavtra
AFTER- AFTER- tomorrow
a-a-xxxxxx
'the day after the day after tomorrow'

The very same pattern is also found in German, with *morgen* 'tomorrow', *über-morgen* 'the day after tomorrow', *über-über-morgen* 'the day after the day after tomorrow', and so on. German also has the pattern *ur-groß-vater*, *ur-ur-groß-vater*, which is the analog of English *great grandfather*, *great great grandfather* and its iterations (various linguistic diagnostics show that these are morphological words rather than phrases). Note that in all those cases it is impossible to insert other prefixes between the iterated prefix: **ur-groß-ur-ur-groß-vater*. In sum, some affixes can be freely iterated as long as no other affixes intervene.

These patterns are all TSL by virtue of being strictly 2-local. We illustrate this claim with German. We ignore the problem of how *groß* can be restricted to occur only with specific stems (if stems are atomic symbols, this is trivial, otherwise it requires a strictly k-local grammar over the string of phonological segments where k is large enough to contain both *groß* and the relevant stems). We also assume, as before, that there is some marker # that marks the edges of stems. Then to ensure that the relevant strings of prefixes follow the pattern *ur*groß*#, the sequences *großur*, *großgroß*, and *ur*# are marked as illicit. Unbounded prefixation thus stays within the class of TSL dependencies.

An interesting counterpart to Russian and German is Ilocano (Galvez Rubino, 1998), which uses the circumfix *ka- -an* with a function similar to *posle* and *über*.

(10) a. bigat
morning
xxxxx
'morning'

b. ka- bigat -an
NEXT- morning -NEXT
a-xxxxx-a′
'the next morning'

Crucially, Ilocano differs from Russian and German in that the circumfix cannot be iterated.

(11) * ka- ka- bigat -an -an
NEXT- NEXT- morning -NEXT -NEXT
a-a-xxxxx-a′-a′
'the next morning after the next one'

Given our previous discussion of circumfixation in Sec. 2.3, Ilocano clearly instantiates a tier-based strictly 2-local pattern, too.

As before, there is little linguistic reason why unbounded circumfixation should be blocked. If affixation can be unbounded to construct more and more complex versions of *day after tomorrow*, and *day after tomorrow* can be derived via circumfixation, then one would expect unbounded circumfixation to be a viable option. But once again there is a clear computational reason as to why this does not happen: the corresponding morphotactic system would no longer be TSL.

Let L be some minor variant of Russian where *posle-* is instead a circumfix *pos- -le*. As before we let h be a homomorphism that replaces all stems by s, the two parts of the circumfix by o, and all other material by some distinct symbol. The intersection of $h(L)$ with the regular language o^+so^+ is the context-free string set o^nso^n. Therefore L is supra-regular and cannot be tier-based strictly local. Unbounded circumfixation simply cannot be reconciled with the subregular hypothesis.

4 Conclusion

The received view is that all of morphology is easily modeled with finite-state machines (Koskenniemi, 1983; Karttunen et al., 1992). We contend that this view is overly generous and that tighter bounds can be established, at least for specific subparts of morphology. Morphotactics defines the restrictions on the possible orderings of morphological units, and we argued based on data from typologically diverse languages that the power of natural language morphotactics is severely restricted:

Subregular Morphotactics All morphotactic dependencies are tier-based strictly local.

In contrast to regular languages, tier-based strictly local languages are efficiently learnable in the limit from positive text (Heinz et al., 2012; Jardine and Heinz, 2016). Our result thus marks a first step towards provably correct machine learning algorithms for natural language morphology.

Admittedly, morphotactics is just one of several parts of morphology. We put aside allomorphy and only considered the distribution of morphemes in the underlying forms. Even within that narrow area we did not thoroughly explore all facets, for instance infixation and incorporation. Nonetheless our results show that the success of the subregular perspective need not be limited to phonology. At least morphotactics can be insightfully studied through this lens, too. In addition, there has been a lot of progress in extending the subregular hierarchy from languages to transductions (see Chandlee (2014) and references therein), and we are confident that these results will allow us to expand the focus of investigation from morphotactics to morphology at large.

It will also be interesting to see how uniform the complexity bounds are across different modules of morphology. In phonology, suprasegmental dependencies tend to be more complex than segmental ones (Jardine, 2015). Most constructions in this paper involve derivational morphology, but the affix *vyo* in Swahili is related to inflectional morphology and turned out to have a distribution that is neither SL nor SP (although it can be captured with a combination of the two). So both derivational and inflectional morphotactics occupy points in $\mathrm{TSL} \setminus (\mathrm{SL} \cup \mathrm{SP})$. In this regard it is also worth noting that some phonological processes such as tone plateauing belong to $\mathrm{SP} \setminus \mathrm{TSL}$, whereas no morphological dependencies seem to be part of this subclass. We hope to address these and related issues in future work.

References

Erhan Aslan and Asli Altan. 1998. The role of *-(s)ı* in turkish indefinite nominal compounds. *Dil*, 131:57–75.

Jane Chandlee. 2014. *Strictly Local Phonological Processes*. Ph.D. thesis, University of Delaware.

Robert Frank and Giorgio Satta. 1998. Optimality theory and the generative complexity of constraint violability. *Computational Linguistics*, 24:307–315.

Carl R. Galvez Rubino. 1998. *Ilocano: Ilocano-English, English-Ilocano: Dictionary and Phrasebook*. Hippocrene Books Inc., U.S.

Thomas Graf and Jeffrey Heinz. 2015. Commonality in disparity: The computational view of syntax and phonology. Slides of a talk given at GLOW 2015, April 18, Paris, France.

Thomas Graf. 2010. Logics of phonological reasoning. Master's thesis, University of California, Los Angeles.

Thomas Graf. 2012. Locality and the complexity of minimalist derivation tree languages. In Philippe de Groot and Mark-Jan Nederhof, editors, *Formal Grammar 2010/2011*, volume 7395 of *Lecture Notes in Computer Science*, pages 208–227, Heidelberg. Springer.

Jeffrey Heinz, Chetan Rawal, and Herbert G. Tanner. 2011. Tier-based strictly local constraints in phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 58–64.

Jeffrey Heinz, Anna Kasprzik, and Timo Kötzing. 2012. Learning with lattice-structure hypothesis spaces. *Theoretical Computer Science*, 457:111–127.

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

Jeffrey Heinz. 2011a. Computational phonology — part I: Foundations. *Language and Linguistics Compass*, 5:140–152.

Jeffrey Heinz. 2011b. Computational phonology — part II: Grammars, learning, and the future. *Language and Linguistics Compass*, 5:153–168.

Jeffrey Heinz. 2015. The computational nature of phonological generalizations. Ms., University of Delaware.

M. A. C. Huybregts. 1984. The weak adequacy of context-free phrase structure grammar. In Ger J. de Haan, Mieke Trommelen, and Wim Zonneveld, editors, *Van Periferie naar Kern*, pages 81–99. Foris, Dordrecht.

Adam Jardine and Jeffrey Heinz. 2016. Learning tier-based strictly 2-local languages. *Transactions of the ACL*, 4:87–98.

Adam Jardine. 2015. Computationally, tone is different. *Phonology*. to appear.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-level morphology with composition. In *COLING'92*, pages 141–148.

Kimmo Koskenniemi. 1983. Two-level morphology: A general computational model for word-form recognition and production. Publication 11, University of Helsinki, Department of General Linguistics, Helsinki.

Waruno Mahdi. 2012. Distinguishing cognate homonyms in Indonesian. *Oceanic Linguistics*, 51(2):402–449.

Robert McNaughton and Seymour Pappert. 1971. *Counter-Free Automata*. MIT Press, Cambridge, MA.

Jason Riggle. 2004. *Generation, Recognition, and Learning in Finite-State Optimality Theory*. Ph.D. thesis, University of California, Los Angeles.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Vischer, David Wellcome, and Sean Wibel. 2010. On languages piecewise testable in the strict sense. In Christan Ebert, Gerhard Jäger, and Jens Michaelis, editors, *The Mathematics of Language*, volume 6149 of *Lecture Notes in Artificial Intelligence*, pages 255–265. Springer, Heidelberg.

Stuart M. Shieber. 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8(3):333–345.

James Neil Sneddon. 1996. *Indonesian Comprehensive Grammar*. Routledge, London and New York.

Greg Stump. 2016. Rule composition in an adequate theory of morphotactics. Manuscript, University of Kentucky.