Open Library of Humanities

Product- and source-oriented generalizations regulate affix distribution in Polish locative adjectives: A corpus-based analysis

Bartłomiej Czaplicki, University of Warsaw, Poland, bczaplicki@uw.edu.pl

A large-scale corpus-based analysis of affix distribution in Polish locative adjectives provides evidence for (i) a selectional restriction formalized as a product-oriented schema, (ii) selectional restrictions formalized as source-oriented schemas, (iii) distinct native and foreign subgrammars, (iv) phonological arbitrariness of the patterns and (v) non-uniformity of outputs. A selectional restriction, which takes the form of a product-oriented segmental requirement, regulates the distribution of the available affixes. Bases that contain a final sonorant form locative adjectives using a different morphological composition than bases that do not contain a final sonorant. As a result of the application of the product-oriented schema, locative adjectives comply with the requirement irrespective of the segmental structure of their bases. Certain consonants in the input show an idiosyncratic affinity for particular affixes. Crucially, the required generalizations refer to segmental material that is absent from the output but present in the input, which must be captured by source-oriented schemas. Further, it is shown that the preferences in affix selection are different for foreign words than for native words, supporting the view that the grammar is composed of two subgrammars with different properties. Affix distribution shows a degree of phonological arbitrariness in two senses: it does not always result in phonological optimization and the segments or groups of segments that condition it often do not form phonetically natural classes. The non-uniformity of outputs is derived using a model that assumes probabilistic weighting of competing generalizations.

1 Introduction

We present the results of a quantitative corpus-based analysis of the distribution of affixes in Polish locative adjectives. The paper argues for the key relevance of the following aspects: ranked schemas of two types: product- and source-oriented, lexical strata, frequency of use and phonological arbitrariness of morphophonological patterns. A Polish locative adjective (LA) is formed using an optional intermorph and an obligatory suffix. There are four possible intermorphs and two suffixes, as represented in (1). The intermorphs appear between the root and the adjectivizing suffix.

(1)   root + {ij, ɨj, aɲ, ɛɲ} + {sk, tsk}$_{LA}$

In past research, the distribution of affixes across different languages was found to be regulated by various phonological, morphological and semantic factors (see Nevins 2011 for an overview). Among the various factors that have been proposed to derive affix distribution of derivational suffixes are selectional restrictions (Fabb 1988; Plag 1996). For example, the English verbalizing suffix *-ize* occurs only on bases that end in an unstressed syllable, while *-en* attaches only to monosyllables that end in an obstruent (Fabb 1988). In this paper, we look at a selectional restriction that refers to a particular segmental structure of the output as well as those that include reference to the input and output.

Schemas exhibit properties that set them apart from traditional re-write rules as conceptualized in early generative research (e.g. Chomsky & Halle 1968): (i) their rate of productivity (strength) depends on the number of words they derive, (ii) they can be stated at various degrees of generality (e.g. segment, feature, syllable, etc.), (iii) they are primarily morphologically conditioned and (iv) they do not necessarily invoke phonological naturalness. Supporting evidence for these properties of schemas can be found in, for example, Bybee (2001), Ellis (2002) (frequency effects), Booij & Audring (2017), Czaplicki (2019; 2020; 2021) (morphological conditioning), Bybee (2001) (phonological arbitrariness).

Two types of schemas (generalizations) have been identified in the literature: product-oriented and source-oriented (Bybee 2001; Pierrehumbert 2006). Product-oriented schemas define the shape of the output without reference to the input. For example, in English past tense verbs end in an alveolar stop (Bybee 2001). Source-oriented schemas refer to the input and output. For example, in Polish the root-final alternation in [drɔŋ-a] 'rod' *g.sg.* and [drɔw̃ʐ-ɛk] *diminutive* can be expressed with the schema $g \leftrightarrow ʐ$ / _$ek_{diminutive}$, which means that the root-final [g] corresponds to [ʐ] in the diminutive formed with the suffix *-ek* (Burzio 2002; Czaplicki 2021). Pierrehumbert (2006) argues on the basis of experimental data that the /k/ : /s/ alternation in English words like *publi*[k] : *publi*[s]*ity* must be expressed by a source-oriented schema $k \leftrightarrow s$ / _*ity*, rather than a product-oriented schema, as such generalizations are formed on the basis of

pairs of related words. Source-oriented schemas are not equivalent to rules, in the sense that they are not generalizations about possible transformations (they do not involve a change-context split). Rather, they are paradigmatic mappings between related words (Kapatsinski 2013).

A pivotal element of the proposed analysis of LAs is a certain selectional restriction that defines the preferred segmental composition: a sonorant preceding the adjectivizing suffix *-sk* in LAs. It is formalized as the product-oriented schema in (2). A quantitative analysis of the corpus data serves to test the hypothesis that the use of the intermorphs in LAs is partially regulated by the need to satisfy this restriction. More specifically, it is shown that the selection of a particular affix depends on the presence in the base of a base-final sonorant. When the base ends in a sonorant, no intermorph is selected. An intermorph is selected when the base does not end in a sonorant. As a result, the generated LAs meet the description of the schema. In addition, the data point to a need for source-oriented schemas, as generalizations pertaining to some groups of segments or individual segments must include reference to the input. This finding undermines Bybee's (2001: 129) claim that "any morphological pattern that can be described by a source-oriented rule can also be described by a product-oriented one" (see Becker & Gouskova 2016 for a similar conclusion). Schemas are used as OT constraints, that is, they are ranked with respect to other constraints, which reflects their relative importance (Burzio 2002). The approach invoking schemas is preferable over an approach based on phonological optimization (e.g. Carstairs 1988), as the required generalizations are phonologically arbitrary. In addition, the schemas are ranked probabilistically in response to the non-uniformity of the outputs observed in the data.

(2)   R-sk-$_{LA}$ (R = sonorant)

There is ample evidence that the phonology of foreign words is different from the phonology of native words (Selkirk 1982; Itô & Mester 1995; Itô & Mester 1999; Inkelas 1999; Pater 2010). It is demonstrated that the subgrammar of foreign words evidences a stronger tendency to use the intermorphs than the subgrammar of native words, which implicates that the two subgrammars have distinct properties.

Frequency is a well-established determinant of the productivity of linguistic patterns and stability of words. The number of words that represent a pattern in a corpus (the pattern's type frequency) is a reliable predictor of the pattern's productivity (MacWhinney 1978; Bybee 1985; 1995). The frequency of a word in a corpus (the word's token frequency) determines its morphological stability (Bybee 2001). It is shown that the productivity of various ways of forming LAs cannot be predicted from their type frequency alone. The LAs with the intermorphs are used more often (i.e. they are more productive) than their type frequency would predict. This is especially true in the foreign stratum and indicates that the intermorphs have become (probabilistic or categorical) cues to foreign status. It is argued that the schema in (2) is responsible

for the observed preference for the intermorphs, even though the patterns with the intermorphs do not show the highest type frequency in the corpus data.

The quantitative analysis also investigates the role of other well-established factors: base-derivative identity, various phonotactic restrictions, similarity avoidance and syllable structure. There is compelling evidence from previous research that each of these factors can have an influence on morphological and phonological composition. The requirement to preserve properties of the base in the derivative leads to violations of certain rules or constraints that are otherwise regular in a given language (Kenstowicz 1996; Steriade 2000). Phonotactic restrictions define impossible combinations of segments in specific domains. For example, the sequence /ps/ is not possible in onsets in English (Harris 1994). Similarity avoidance can lead to the selection of an otherwise unexpected affix (Czaplicki 2022). Finally, there is evidence that syllable structure can determine allomorph distribution. In phonologically conditioned suppletion two phonologically dissimilar alternants need to be stored, however, their distribution is phonologically conditioned (Carstairs 1988; 1990). In Korean, there are two allomorphs of the nominative singular ending: *-i* and *-ka*. The vowel-initial ending is selected after stem-final consonants (e.g. /mom-i/ 'body' nom.sg.), while the consonant-initial ending is chosen after stem-final vowels (e.g. /kʰo-ka/ 'nose' nom.sg.) (Nevins 2011). The result in both cases is the unmarked CVCV syllable structure. In this paper, I identify a propensity for avoiding extrasyllabic consonants with dedicated phonological and morphological mechanisms: vowel insertion and the use of an intermorph. In addition, certain distributional patterns are not optimizing and must be categorized as phonologically arbitrary, a finding consistent with the model proposed in Paster (2006).

The conceptual framework for the analysis is a mixture of ideas from usage-based linguistics (product- and source-oriented schemas) and Harmonic Grammar (weighted constraints, Pater 2010). The set of product-oriented schemas linked to a particular meaning can be conceptualized as a description of the set of word forms associated with that meaning (Bybee 1985). Schemas are thus related to the learned phonotactic constraints of Harmonic Grammar, because the phonotactic grammar can also be viewed as a description of the lexicon (Kapatsinski 2013). Weighted or ranked schemas are useful for modeling patterns that are nonuniversal, learned by generalizing over the lexicon and involve competition between several possible outputs (e.g. Kapatsinski 2013 employs weighted schemas in a Maximum Entropy grammar).

The paper is structured as follows. The next section presents the main generalizations. Section 3 defines frequency and, with the use of data extracted from a corpus, investigates the type frequency of the competing ways of forming LAs. Section 4 presents the results of a corpus-based analysis. In section 5, I put forward a constraint-based analysis of the data and model the resulting grammar using a probabilistic framework, Noisy Harmonic Grammar. Section 6 comments on the possible role of lexical frequency. Section 7 discusses the main implications of the analysis. Section 8 provides the main conclusions.

2 Locative adjectives – basic generalizations

The suffix /-sk-/ is one of the two most productive and semantically versatile suffixes used to form adjectives, including locative adjectives, in Polish (Satkiewicz 1969: 128; Kallas 1999: 494–499; Czaplicki 2014a: 136–142), as exemplified in (3a). In the nominative singular, it is followed by an ending designating masculine, feminine and neuter genders, *-i, -a* and *-e*, respectively. The suffix /-sk-/ has the allomorph /-tsk-/, where /ts/ is an affricate, as illustrated in (3b). The two allomorphs show restrictions to base-final consonants. When the base ends in /s x ʂ g/ the allomorph /-sk-/ is commonly used and the base-final consonant is dropped, e.g. /nɨs-a/ > /nɨ-sk-i/. The allomorph /-tsk-/ often occurs when the base ends in /t k ts tʂ tɕ/. The base-final consonant is lost (Kreja 1989: 50), e.g. /rɛjkjavik/ > /rɛjkjavi-tsk-i/ (Though these generalizations are not exceptionless, see section 4). As the suffix /-tsk-/ appears mostly in complementary distribution with /-sk-/ (/-tsk-/ appears after voiceless non-labial stops and affricates, while /-sk-/ appears elsewhere), it will be classified as a contextually conditioned allomorph of /-sk-/. The following patterns will be analyzed as instances of coalescence. A base-final fricative coalesces with the suffix-initial fricative /s/, giving rise to /s/ (e.g. ʂ + sk → sk), while a base-final stop or affricate coalesces with the suffix-initial /s/ giving rise to the affricate /ts/ (e.g. t + sk → tsk). However, the behavior of velar plosives is somewhat unexpected. While the base-final /k/ coalesces with /-sk-/, giving /-tsk-/, the base-final /g/ is lost and appears with /-sk-/ instead, e.g. /xag-a/ 'the Hague' > /xa-sk-i/. This pattern shows a certain degree of arbitrariness, discussed in more detail in section 5.1.

The suffix /-sk-/ can be used alone or with one of the four intermorphs, i.e. /-ij-/, /-aɲ-/, /-ɛɲ-/ and /-ɔf-/ (Kreja 1989: 49; Gussmann 2007: 152–157), as illustrated in (3c–f).[1] The sources concur that the intermorphs do not contribute any meaning to adjectives in /-sk-/ (e.g. Kreja 1989: 48–49; Kowalik 1997: 48–55; Gussmann 2007: 157). In fact, Kreja (1989) treats the intermorphs as suffix extensions. The intermorph /-ij-/ has the allomorph /-ɨj-/, used after non-palatal consonants /t d s z r dʐ/. The intermorph /-ɔf-/ is predominately used to form adjectives from proper nouns denoting people (/lɛɲin/ 'Lenin' – /lɛɲin-ɔf-sk-i/), from nouns denoting proponents of scientific disciplines and political movements (/naʑist-a/ 'Nazi' – /naʑist-ɔf-sk-i/) and from other personal nouns (/gɛj/ 'gay' – /gɛj-ɔf-sk-i/) (Satkiewicz 1969: 145; Szymanek 2010). Crucially, the intermorph /-ɔf-/ is used in adjectives referring to people.[2] The intermorphs

---

[1] In fact, according to multiple descriptions of Polish, nasals are pronounced as nasal glides before fricatives (e.g. Wierzchowska 1960). This means that in the narrow transcription the intermorphs are pronounced with a nasal glide, i.e. as [-ãȷ̃-sk-] and [-ɛ̃ȷ̃-sk-]. However, I use the broad transcription /-aɲ-sk-/ and /-ɛɲ-sk-/ mainly for better readability but also for consistency with other sources. It is important that segments [j] and [ȷ̃] are contrastive before fricatives, which means that nasal gliding is a purely phonetic effect with no impact on contrast.

Gussmann (2007) discusses /-ij-sk-/, /-aɲ-sk-/, and /-ɔf-sk-/ but does not discuss /-ɛɲ-sk-/.

[2] As a result of its specialization to adjectives denoting people, the intermorph /-ɔf-/ has acquired more specific meaning.

/-ij-/, /-aɲ-/ and /-ɛɲ-/ are used to form adjectives referring to place names (toponyms). As the present discussion focuses on various ways of forming LAs, the intermorph /-ɔf-/ will be omitted from further consideration, as it does not compete with the other intermorphs, /-ij-/, /-aɲ-/ and /-ɛɲ-/.³ Each of the ways of forming LAs in (3a–e) is productive; see the evidence in section 3. For clarity, adjective formation using /-sk-/ and /-tsk-/ alone (i.e. bare suffixes without an intermorph) will be referred to as /-∅-sk-/ and /-∅-tsk-/ in what follows.

(3)

	Noun	Adjective		Noun	Adjective
a.	-∅-sk-		b.	-∅-tsk-	
	nɨs-a 'Nysa'	nɨ-sk-i		lɛɡɲits-a 'Legnica'	lɛɡɲi-tsk-i
	xin-ɨ 'China'	xiɲ-sk-i		kuvɛjt 'Kuwait'	kuvɛj-tsk-i
	puwav-ɨ 'Puławy'	puwaf-sk-i		rɛjkjavik 'Rejkjavik'	rɛjkjavi-tsk-i
	ɛɡipt 'Egypt'	ɛɡip-sk-i		madrɨt 'Madrid'	madrɨ-tsk-i
	nɔrvɛgj-a 'Norway'	nɔrvɛ-sk-i		atlantɨk 'Atlantic'	atlantɨ-tsk-i
c.	-ij-sk-/-ɨj-sk-		d.	-aɲ-sk-	
	kanad-a 'Canada'	kanad-ɨj-sk-i		marɔk-ɔ 'Morocco'	marɔk-aɲ-sk-i
	mɔnak-ɔ 'Monaco'	mɔnak-ij-sk-i		amɛrɨk-a 'America'	amɛrɨk-aɲ-sk-i
	tɔg-ɔ 'Togo'	tɔg-ij-sk-i		tɨbɛt 'Tibet'	tɨbɛt-aɲ-sk-i
	tʂilɛ 'Chile'	tʂil-ij-sk-i		kub-a 'Cuba'	kub-aɲ-sk-i
	dɛli 'Delhi'	dɛl-ij-sk-i		alask-a 'Alaska'	alask-aɲ-sk-i
e.	-ɛɲ-sk-		f.	-ɔf-sk-	
	budapɛʂt 'Budapest'	budapɛʂt-ɛɲ-sk-i		ʐɨd-a 'Jew' gen.sg.	ʐɨd-ɔf-sk-i
	lim-a 'Lima'	lim-ɛɲ-sk-i		mistʂ 'master'	mistʂ-ɔf-sk-i
	bɛrn-ɔ 'Bern'	bɛrn-ɛɲ-sk-i			
	krɔsn-ɔ 'Krosno'	krɔɕɲ-ɛɲ-sk-i			

³ Though there are a handful of LAs formed using /-ɔf-/, e.g. /brɔdwɛj/ 'Broadway' > /brɔdwɛj-ɔf-sk-i/.

LAs in /-∅-sk-/ and its allomorph /-∅-tsk-/, illustrated in (3a) and (3b), commonly trigger modifications of the root of the base in the form of consonant mutations or deletions (e.g. n ~ ɲ, v ~ f, ʐ ~ ∅, g ~ ∅, t ~ ∅, d ~ ∅ and k ~ ∅). Mutations or deletions are not found before the intermorphs /-ij-/ and /-aɲ-/, as shown in (3c) and (3d).[4] The intermorph /-ɛɲ-/ may occasionally trigger mutations of the base-final consonant(s) (usually /n ~ ɲ/).

What is the status of the intermorphs? It will be shown that the intermorphs are separate morphological chunks and that they are productively used to form new LAs. Intermorphs are defined as "morpheme-like sequences which are inserted between the base and the suffix" (Gussmann 2007: 157). LAs, exemplified in the previous section, can be analyzed in two different ways: either (1) the suffixes /-sk-/, /-aɲsk-/, /-ijsk-/, /-ɛɲsk-/ and /-ɔfsk-/ have no internal structure or (2) there is the shared adjectivizing suffix /-sk-/ which can combine with several possible intermorphs: /-aɲ-/, /-ij-/, /-ɛɲ-/, and /-ɔf-/. Following Kowalik (1997) and Gussmann (2007), the latter approach is advocated for a number of reasons. First, on the assumption that the intermorphs are independent of the adjectivizing /-sk-/, LAs are formed using an optional intermorph and the suffix /-sk-/, the latter being the most versatile and productive suffix used to form adjectives. In this way, the suffix can appear on its own or in combination with the intermorphs. Second, many intermorphs can be used in combination with several suffixes, in addition to /-sk-/. For example /-ij-/ can be used in combination with the adjectivizing suffix /-n-/, as in /dɛprɛs-ij-n-ɨ/ 'depressive', and /-ɔf-/ ~ /-ɔv/ can be used with /-n-/ and /-it-/, as in /list-ɔv-n-ɨ/ 'related to letters' and /prats-ɔv-it-ɨ/ 'hardworking', in addition to the suffix /-sk-/ (Kowalik 1997: 48). Third, several affixes function both as intermorphs and as final suffixes. For example, /-ɔf-/ ~ /-ɔv-/ can function as an intermorph, as in /sɨn-ɔf-sk-i/ 'related to a son', or as the final suffix used to form relational adjectives, as in /narɔd-ɔv- ɨ/ 'national'.[5] Fourth, there is evidence that affixes can acquire new functions. Kapatsinski (2021) argues that the Russian adjectivizing suffix /-ɔv-/ is related to the formally identical genitive plural suffix through paradigmatic associations. In sum, there is convincing evidence that both the adjectivizing suffix /-sk-/ and the intermorphs /-aɲ-/, /-ij-/, /-ɛɲ-/ and /-ɔf-/ function to some extent independently of one another. It follows that the intermorphs should be granted the status of separate affixes.

The claim that intermorphs are separate chunks of morphological structure can be easily reconciled with the finding that affix boundaries show a range of strengths, which means that there is ranking in the decomposability of morphologically complex words (Hay 2003). While

[4] The intermorph /-ij-/ begins with a high front vowel, yet it does not trigger phonological palatalization of the preceding consonant. This observation is fully consistent with the results of Czaplicki's (2019) quantitative study indicating that consonant mutations in Polish are morphologically conditioned. Phonological conditioning (e.g. palatalization before front vowels), though historically relevant, no longer plays an important role.

[5] Kreja (1989: 49) suggests that the intermorph /-ɔf-/ in /-ɔf-sk-/ originated from the final suffix /-ɔv-/ and specialized in forming adjectives from noun bases denoting people.

intermorphs are decomposable, their decomposability is lower than that of other affixes, as their distribution is restricted. Following this line of reasoning, the intermorph + affix boundary is weaker than many other affix boundaries: /root + + IM + sk/, where IM indicates an intermorph, + + represents a stronger and + a weaker affix boundary.

The sequences /ij-/ and /ɛɲ-/ are derived in two different ways in LAs. In the LAs in (4a) the sequences arise through vowel insertion and are not intermorphs. They are derived by means of the bare suffix, as the consonants /j/ and /ɲ/ are part of the roots, while the vowels /i/ (/ɨ/) and /ɛ/ result from insertion. On the other hand, in (4b), the formally identical sequences are intermorphs, as their material is completely absent from the root. Thus, the status of the sequences /-ij-/ and /-ɛɲ-/ in LAs depends on whether they have corresponding segments in the input (i.e. their base nouns), as in (4a), or not, as in (4b).

(4)　　　　　/ij/　　　　　　　　　　　/ɛɲ/
　　　a.　bɛlgj-a – bɛlgij-sk-i　　　viln-ɔ – vilɛɲ-sk-i
　　　　　'Belgium'　　　　　　　　　'Vilnius'
　　　　　indɔnɛzj-a – indɔnɛzij-sk-i　grɔdn-ɔ – grɔdʑɛɲ-sk-i
　　　　　'Indonesia'　　　　　　　　'Grodno'
　　　b.　tʂil-ɛ – tʂil-ij-sk-i　　　　krɛt-a – krɛt-ɛɲ-sk-i
　　　　　'Chile'　　　　　　　　　　'Crete'
　　　　　tsɨpr – tsɨpr-ij-sk-i　　　　budapɛʂt – budapɛʂt-ɛɲ-sk-i
　　　　　'Cyprus'　　　　　　　　　'Budapest'
　　　　　kɔng-ɔ – kɔng-ij-sk-i　　　lim-a – lim-ɛɲ-sk-i
　　　　　'Congo'　　　　　　　　　'Lima'

The instances of vowel insertion in (4a) are driven by syllable structure and more specifically by avoidance of extrasyllabic consonants. An extrasyllabic consonant refers to a consonant (a sonorant) that is "trapped" between two obstruents and in this way violates the Sonority Sequencing Principle (Selkirk 1982). In this analysis, the extrasyllabic consonant is found in the *potential* LA formed using -∅-*sk*. For example, the LA of /grɔdn-ɔ/ is /grɔdʑɛɲ-sk-i/ with vowel insertion. The *potential* but non occurring LA formed with -∅-*sk* would contain the extrasyllabic consonant /n/, */grɔdn-sk-i/, which is avoided. In the case of /pilzn-ɔ/ 'Pilzno' the potential LA with -∅-*sk* would contain an extrasyllabic /n/, i.e. */pilzn-sk-i/. The attested LA avoids the extrasyllabic consonant by using an intermorph, /pilzn-ɛɲ-sk-i/. Thus, the notion of extrasyllabic consonant as used in this analysis refers to a *potential* output with the -∅-*sk* suffix. In this light, the use of an epenthetic vowel or an intermorph can be viewed as a strategy to avoid extrasyllabic consonants.

Similarly, the sequence /-aɲ/ can have a different status in LAs. In the LA in (5a), the sequence is present in the base, which means that it is not an intermorph. In contrast, the sequence in (5b) is classified as an intermorph, as it is absent from the base.[6]

---

[6] The decision to treat LAs such as /grɔdʑɛɲ-sk-i/ < /grɔdn-ɔ/ 'Grodno' and /pɔznaɲ-sk-i/ < /pɔznaɲ/ 'Poznań' as not containing an intermorph has consequences for the analysis. It correctly derives the fact that the intermorph-

(5)　　　/aɲ/
　　　a.　pɔznaɲ – pɔznaɲ-sk-i　　　　b.　marɔk-ɔ – marɔk-aɲ-sk-i
　　　　　'Poznań'　　　　　　　　　　　　'Morocco'

The hypotheses for the corpus-based analysis of the distribution of affixes in LAs are formulated in (6).

(6)　　　H1: The distribution of affixes is governed by the product-oriented schema *R-sk*.
　　　　　H2: The distribution of affixes is governed by source-oriented generalizations.
　　　　　H3: The distribution of affixes is partially phonologically arbitrary.
　　　　　H4: There is a difference between the distribution of affixes in the native and foreign stratum.
　　　　　H5: The distribution of affixes across the attested outputs is probabilistic.

3 Frequency and productivity

There is growing evidence that frequency plays an important role in morphophonology (Mańczak 1980; Bybee 1995; Bybee 2001; Ellis 2002; Albright 2002; Albright & Hayes 2003; Baayen et al. 2003; Dąbrowska 2008; Czaplicki 2013a; 2013b; 2014a; 2014b; 2021). The generalizability of a pattern crucially depends on the number of stored words that exhibit the pattern (lexical gang effects; MacWhinney 1978; Stemberger & MacWhinney 1988; Alegre & Gordon 1999). The higher the number of words that adhere to a given pattern (i.e. the larger the gang and the higher the type frequency), the more likely the pattern is to be extended to novel words. More specifically, when two (or more) patterns are available in a particular context, the pattern with a higher type frequency is the one most likely to become generalized to novel words (all else being equal). In other words, pattern extension is the most likely to recruit the most robust of the several patterns used in a particular morphological context.[7] There is accumulating evidence that analyses probing the role of type frequency in pattern extension should be more nuanced. Albright (2002) presents evidence from Italian for a grammar in which general rules exist alongside more specific but more reliable generalizations describing subregularities for the same process. Albright (2002) refers to such subregularities as *islands of reliability* and defines them as environments in which the reliability value of a rule is higher than the general reliability of the rule (Albright 2002: 686). A grammar showing islands of reliability requires modeling

---

like sequence in the derivative has a correspondent (or correspondents) in the base (an instance of base-derivative identity). However, it does not take into consideration the fact that the output contains a string which from the perspective of product-oriented generalizations is identical to an intermorph. The inclusion of the R-sk schema in the analysis, defined in section 5.1, addresses this issue, as the schema is product-oriented. All the LAs that show a sonorant before the *-sk* suffix irrespective of is provenance comply with the schema.

[7] This accords with Bybee (1995), who argues that while type frequency is an important predictor of pattern productivity, token frequency plays no role.

competition between general and local pattern extensions. Albright & Hayes (2003) argue that the productivity of rules is a function of their relative reliability. Reliability of a rule is affected by the number of words that adhere to the rule (its type frequency), but also by the number of exceptions to the rule. Thus, a rule with high type frequency, but with many exceptions is not as productive as a comparable rule with few exceptions. Albright & Hayes (2003) argue that several competing patterns (general and specific) might be extended in a particular context, with their productivity dependent on their relative reliability.

Throughout the analysis, I use the data from the *plTenTen19: Corpus of the Polish Web*, which is made up of texts collected from the internet in 2019 and comprises more than 4.2 billion words. *Sketchengine* was used to search for the relevant lemmas. This search engine allows for searching for lemmas tagged as adjectives of a specific shape using a wild card, e.g. *ijsk*. Separate searches were performed for each segment preceding the adjectivizing affixes (e.g. *bijsk*, *ryjsk*, *bsk*, *esk*, *ick*) and the extracted words were integrated into a single list.

LAs were then manually filtered out from the whole set and manually coded for the relevant variables: *suffix, foreign, extrasyllabic* and *consonant*. The different ways of forming LAs were coded using the variable *suffix* with six categories: sk, tsk, ijsk, aɲsk, ɛɲsk, V ("V" codes vowel insertion). In accordance with the assumptions provided in the previous section, the intermorphs /-ij-/, /-aɲ-/ and /-ɛɲ-/ were coded only if the relevant sequence was completely absent from the base. That is, /marɔk-aɲ-sk-i/ and /tsɨpr-ij-sk-i/ were coded as having the intermorphs, /-aɲ-/ and /-ij-/, respectively, while /bɛlgij-sk-i/ and /pɔznaɲ-sk-i/ were coded without the intermorphs (the former was coded as showing vowel insertion, V, the latter was coded with the suffix /∅-sk/). The variable *foreign* was coded with two categories: native ("no") and foreign ("yes"). Foreign LAs are those that refer to places that are located outside of Poland. The variable *extrasyllabic* was coded as "no" and "yes" (for sonorants), in accordance with the definition provided in the previous section. The variable *consonant* included all the attested base-final consonants. More detailed information about the coding of the variable *consonant* is given in section 4. Raw token frequency counts were extracted using frequency lists in *Sketchengine* to provide an estimate of the lexical frequency of the LAs.

In the remainder of this section, I focus on Polish LAs with the aim of showing that (i) the type frequency of adjectives without an intermorph, i.e. in /-∅-sk-/ (/-∅-tsk-/), is significantly higher than the type frequency of adjectives in /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/, and (ii) while all four patterns are productive, the patterns with the intermorphs are more productive than the pattern without them. In order to assess the frequency of the four ways of forming adjectives in /-sk-/, I extracted LAs in /-∅-sk-/ (including LAs in /-∅-tsk-/ and those exhibiting vowel insertion), /-ij-sk-/ (including /-ɨj-sk-/), /-aɲ-sk-/ and /-ɛɲ-sk-/ from the corpus. The results are presented in **Table 1**.

	-∅-sk-	*-ij-sk-*	*-aɲ-sk-*	*-ɛɲ-sk-*	*Sum*
Type freq.	2,194	125	141	43	2,503
%	87.7	5	5.6	1.7	100

Table 1: Type frequency of LAs in /-∅-sk-/, /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/ in the corpus.

Based on the data in **Table 1**, LAs without the intermorphs, /-∅-sk-/ (and its allomorph /-∅-tsk-/), greatly outnumber LAs with the three intermorphs, /-ij-/, /-aɲ-/ and /-ɛɲ-/. The number of adjectives in /-∅-sk-/ is 17.6 times higher than those in /-ij-sk-/, 15.6 times higher than those in /-aɲ-sk-/ and 51 times higher than those in /-ɛɲ-sk-/. It also appears that the intermorphs /-ij-/ and /-aɲ-/ occur equally frequently in LAs, while the intermorph /-ɛɲ-/ is three times less frequent than the other two. The data in **Table 1** provide enough evidence to conclude that LAs in /-∅-sk/ greatly outnumber LAs in /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/.

Baayen & Lieber (1991) and Baayen (1993) argue that the number of occasionalisms, including hapax legomena (words that occur in a corpus only once), is an adequate estimate of the productivity of a pattern. Following Baayen & Lieber (1991), the productivity of a pattern is calculated as the ratio of occasionalisms to the token frequency of all the words exhibiting this pattern. This measure of productivity is called *productivity in the narrow sense* and is used to assess the probability of encountering new formations among all derivatives of a certain morphological category (Plag 2012). **Table 2** provides the results of the estimations of the productivity measure for LAs in /-∅-sk-/, /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/. Following the discussion in Haspelmath & Sims (2010: 135–136) suggesting that for very large corpora the count of "true" hapax legomena may not always be representative, it is assumed that occasionalisms are words with the frequency of 50 or below. In **Table 2**, the productivity measure is the ratio of the number of words with frequency 50 or below to the token frequency of all the words showing the pattern (aggregated token frequency). It appears that the productivity measures for adjectives in /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/ are higher than the productivity measure for adjectives in /-∅-sk-/. These results suggest that while all four patterns are productive (occasionalisms are attested for all of them), the patterns with the intermorphs are more productive than the pattern without them.[8]

---

[8] According to an alternative explanation for the productivity of the patterns with the intermorphs, the percentages of occasionalisms in **Table 2** should mirror the overall type frequency data given in **Table 1**. Following the reasoning presented in Hayes et al. (2009) and Kapatsinski (2010), users extend the four patterns based on their actual type frequency in the corpus. However, the 25.7% difference (87.7%–62%) between **Table 1** and **Table 2** for the pattern without the intermorphs cannot be easily brushed aside. A binomial test in R (binom.test(206, 206 + 60 + 42 + 24, p = 0.877)) yields p < .001 and a 95% confidence interval from 0.566 to 0.673. This provides statistical support that the pattern with the bare suffix is extended less often than its type frequency predicts. A reviewer suggests that type frequency within the foreign stratum can be used to predict the usage of the intermorphs in the foreign stratum (in the foreign stratum 77.9% of LAs are used with the bare suffix). Though certainly interesting, this line of reasoning cannot be used to explain how the difference in affix preferences between the two strata arose.

	Number of occasionalisms	% of occasionalisms	Aggregated token frequency	Productivity measure
-∅-sk-	206	62	32,220,630	0.00000639
-ij-sk-	60	18	1,557,075	0.00003853
-aɲ-sk-	42	13	2,819,366	0.0000149
-ɛɲ-sk-	24	7	511,606	0.00004691

Table 2: Productivity measures of LAs in /-∅-sk-/, /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/ in the corpus.

Based on the type frequency data, I conclude that the pattern of forming LAs using /-∅-sk-/ is significantly more robust in the lexicon than the three patterns using the intermorphs, that is, /-ij-sk-/, /-aɲ-sk-/ and /-ɛɲ-sk-/. However, the three patterns with the intermorphs are more productive than the pattern without them. Therefore, we need to explain why the three patterns with the intermorphs are currently more productive than the pattern without them, even though the latter exhibits a significantly higher type frequency.

4 Corpus-based analysis

This section aims to determine which factors play a role in the distribution of affixes in LAs. I focus on base-final consonants and extrasyllabicity. In addition, I attempt to find out whether the native stratum shows different affix preferences than the foreign stratum. The first multinomial analysis serves to answer this question by looking at the whole set of LAs. Once this prediction is confirmed, two analyses are run on the data: one for the native LAs and one for the foreign LAs.

4.1 The whole set of data

The entire set of LAs extracted from the corpus comprises 2,503 words. The corpus data clearly show that the distribution of the intermorphs and suffixes is not governed by categorical rules. It is common to find variation in the choice of affixes. For some bases, three competing LAs with different affixes appear in usage (see section 6 for details). For example, the LAs of /alask-a/ 'Alaska' are /ala(-)sk-i/, /alask-aɲ-sk-i/ and /alask-ij-sk-i/. The token frequencies of the three adjectives are different: 656, 554 and 17. Similarly, /angɔl-a/ 'Angola' produces /angɔl-sk-i/ (805), /angɔl-aɲ-sk-i/ (202) and /angɔl-ij-sk-i/ (18). An investigation of the context of usage of the forms in the corpus reveals no evidence that any of the forms that appear in competition have acquired a different meaning from the others (but see footnote 11). Due to the observed variation

in the data, we should expect to identify a gradient, rather than categorical, impact of a host of factors on the selection of an affix. The factors, which include base-final consonant(s), lexical strata and syllable structure, are subjected to statistical analyses in this section. The results are then translated into a constraint-based analysis that employs schemas as well as markedness and faithfulness constraints in section 5.

A multinomial logistic regression analysis was run on the data using *multinom()* function in the *nnet* package (Venables & Ripley 2002) in R (R Core Team 2021). The analysis uses *suffix* as the dependent variable and *consonant, extrasyllabic*, and *foreign* as predictors. Three interactions are also included: *consonant* and *extrasyllabic*, *extrasyllabic* and *foreign*, and *consonant* and *foreign*. The coding of the variables is given in (7). The possible affixes are coded in *suffix*. The category "V" stands for vowel insertion. The bare suffix /-∅-sk-/ (sk) is the reference category. For the variable *consonant* /s/ was chosen as the reference category because this consonant shows a fairly consistent behavior in that it usually chooses /-∅-sk-/ and deletes. In addition, it is relatively common in the data. For these reasons it provides a good baseline. The analysis yields p values, which were calculated on the basis of z scores. The results of the analysis are given in Appendix A (Analysis 1). Only statistically significant results are shown. A particular suffix is deemed preferred in a given context when its coefficient is positive and significantly different from the reference level. Conversely, a suffix is considered dispreferred when it has a negative coefficient that is significantly different from the reference level.

(7)   suffix: sk, aɲsk, ijsk, ɛɲsk, tsk, V   ("sk" is the reference category)
        consonant: p, b, f, v, w, m, t, d, n, s, z, r, l, ʂ, ʐ, tʂ, dʐ, ɕ, ʑ, tɕ, dʑ, ɲ, k, g, x ("s" is the reference category)
        extrasyllabic: no, yes ("no" is the reference category)
        foreign: no, yes ("no" is the reference category)

The stacked percentage bar plot in **Figure 1** shows affix distribution and vowel insertion in LAs according to the base-final consonant. Note that vowel insertion (V), depicted as a separate category in **Figure 1**, in fact falls under the usage of the bare suffix /-∅-sk/, as none of the intermorphs is used. In other words, vowel insertion is a special case of root modification; it does not involve the usage of an intermorph. However, I use vowel insertion as a separate category, as it occurs alongside the intermorphs as a strategy to avoid extrasyllabic consonants. While there is a general preference for ∅-*sk* in the data in **Figure 1**, certain consonants stand out. For example, /ts/ predominately chooses -*tsk*- and almost half of the LAs with base-final /j/ exhibit the bare suffix and vowel insertion. /f/ mostly selects -*ij*-, while there is no clear preference for base-final /tʂ/. We return to the factor *consonant*

in the next sections, where the consonant-related preferences are examined separately in the native and foreign stratum.
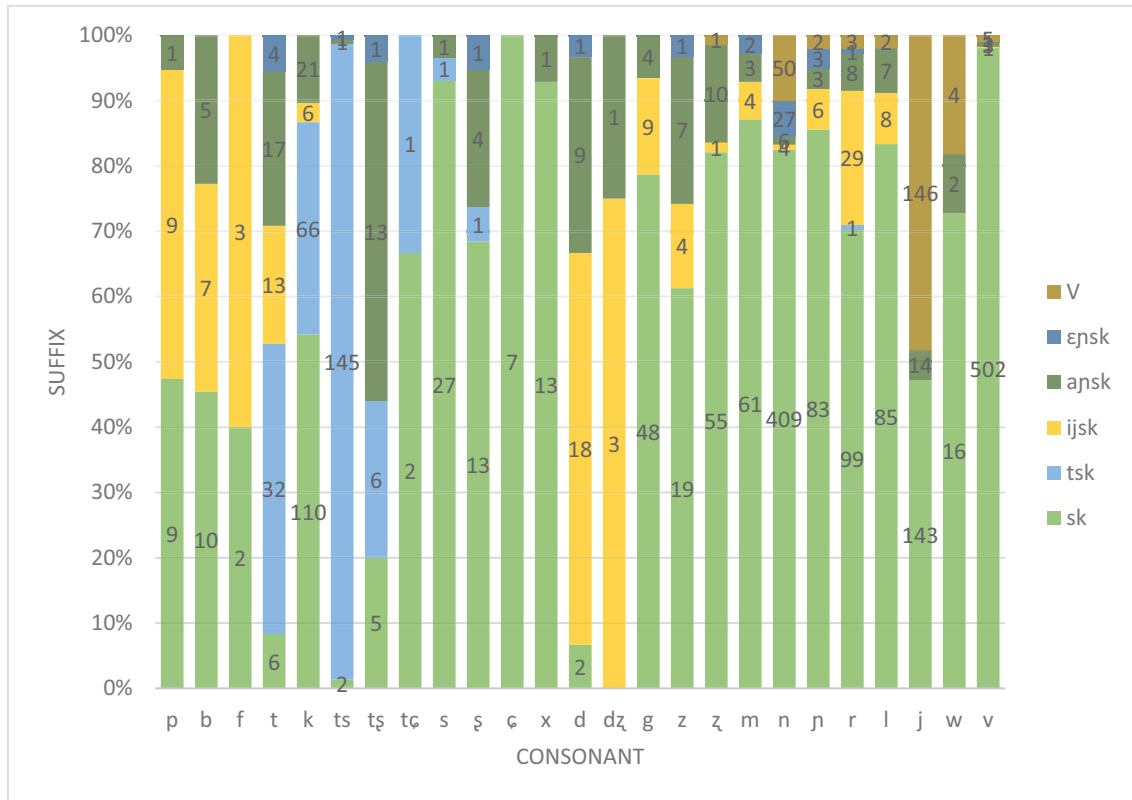

Figure 1: Observed affix and vowel insertion preferences depending on the base-final consonant.

A note about the behavior of /v/ is in order. Although /v/ is described as an obstruent in modern Polish, it shows certain traits of a sonorant in its phonological behavior, thus reflecting its historical provenance. It can be traced back to the bilabial glide /w/ (Bethin 1998: 203). Thus, in what follows /v/ is analyzed together with the sonorants in order to determine whether it functions as a sonorant or not. Evidence for the sonorant status of /v/ is given in section 5.1.

The graph in **Figure 2** shows the distribution of the affixes depending on whether there is an extrasyllabic consonant in a potential LA formed with -∅-*sk*. The results of the multinomial analysis are consistent with the proportions shown in **Figure 2**. The presence of an extrasyllabic consonant in a potential LA in -∅-*sk* significantly increases the likelihood that vowel insertion (z = 2.13, p = .03) and the intermorphs /-aɲ-/ (z = 51.06, p < .001), /-ɛɲ-/ (z = 8.09, p < .001) and /-ij-/ (z = 3.71, p < .001) are used.
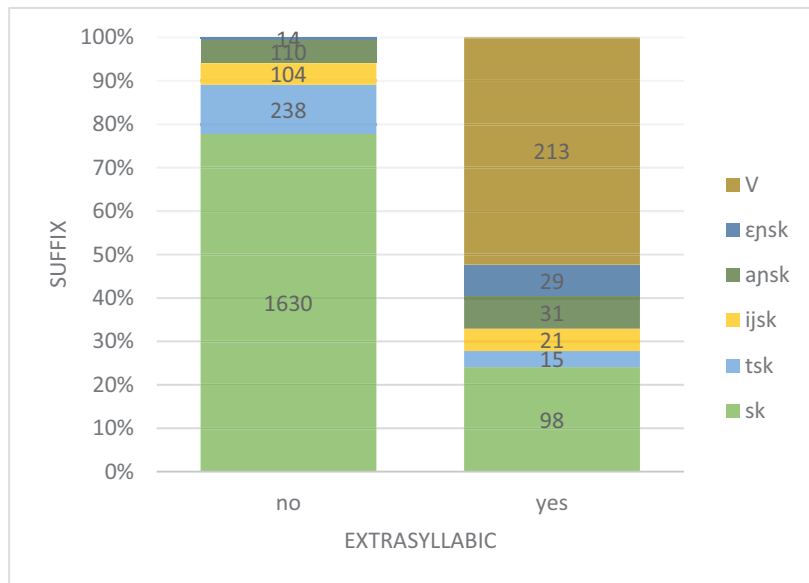
Figure 2: Observed affix and vowel insertion preferences depending on extrasyllabicity.

We turn to the factor *foreign*. The graph in **Figure 3** shows the preferences in the two strata of the lexicon: native and foreign LAs. In the native stratum vowel insertion and the intermorphs /-aɲ-/, /-ɛɲ-/ and /-ij-/ are used in 125 of 1419 LAs, which constitutes 8.8%, while in the foreign stratum these strategies are used in 397 of 1084 LAs, which amounts to 36.6%. The intermorph /-ij-/ is practically not used in the native stratum (1 case). The multinomial analysis confirms the preference for the intermorphs in the foreign stratum, as compared with the native stratum: /-aɲ-/ (z = 35.75, p < .001), /-ɛɲ-/ (z = 28.71, p < .001) and /-ij-/ (z = 5.20, p < .001).
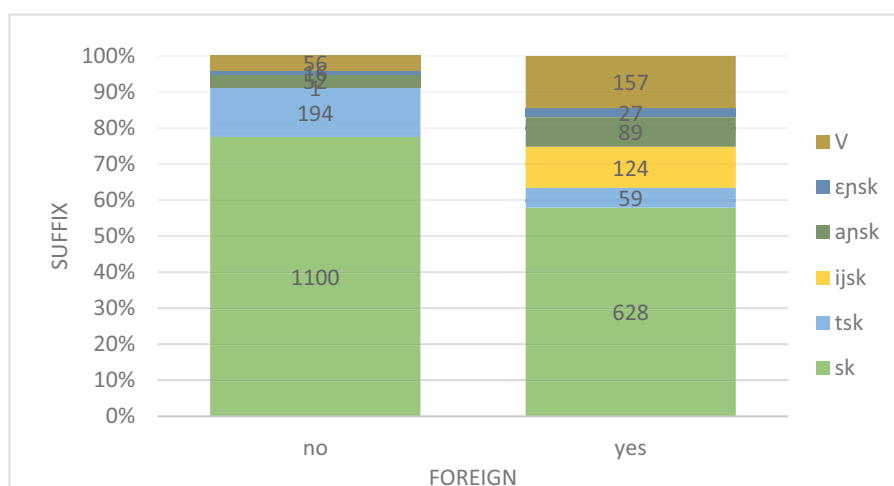

Figure 3: Observed affix and vowel insertion preferences in the two strata: native and foreign.

We now turn to the three interactions (see Appendix A for details). The results for the interaction of *consonant* and *extrasyllabic* show that a potential extrasyllabic consonant (/m n ɲ r l w v/) increases the probability of using (one or several of) the intermorphs and vowel insertion (only /j/ shows a preference for the bare suffix). The interaction of *extrasyllabic* and *foreign* has a significant effect for all the intermorphs (compared with the bare suffix): /-aɲ-/ (z = –14.22, p < .001), /-ɛɲ-/ (z = –6.33, p < .001) and /-ij-/ (z = 15.94, p < .001). This confirms that there are different preferences for the affixes used with potential extrasyllabic consonants in the native and foreign stratum. Finally, the results of the interaction of *consonant* and *foreign* indicate that for a majority of the consonants the intermorphs are used more often in the foreign than in the native stratum. Only the consonants /s ʂ ɲ w ʐ/ show a striking preference for the bare suffix in the foreign stratum. The results of the interactions suggest that the two strata: native and foreign, display different preferences for the intermorphs.

Based on these results, two multinomial analyses have been run on the data: the first one investigates the native stratum, while the second one examines the foreign stratum. In this way, we get more insight into the affix preferences in each of the two strata.

4.2 Native stratum

The multinomial analysis examines native LAs and uses the predictors in (8). In addition to the predictors *consonant* and *extrasyllabic*, the interaction of *consonant* and *extrasyllabic* is included in the model, as certain consonants can function as extrasyllabic or not. Specifically, the choice of an intermorph, vowel insertion or the bare suffix can depend on the syntagmatic context of a consonant.

(8)   consonant: p, b, f, v, w, m, t, d, n, s, z, r, l, ʂ, ʐ, tʂ, dʐ, ɕ, ʑ, tɕ, dʑ, ɲ, k, g, x ("s" is the reference category)
         extrasyllabic: no, yes ("no" is the reference category)
         the interaction of consonant and extrasyllabic

Figure 4 shows affix and vowel insertion preferences according to the consonant.[9] **Table 3** provides the results of the multinomial analysis. It lists the statistically significant preferences and dispreferences for different strategies (based on positive vs. negative coefficients that are significantly different from the reference level). More detailed results are given in Appendix A (Analysis 2). While there is a general preference for -∅-*sk* across all the consonants, certain consonants show statistically significant preferences for the intermorphs and -∅-*tsk*. Voiceless non-labial stops and affricates predominantly choose the suffix -∅-*tsk* and less often the intermorph -*aɲ*-. Voiceless non-labial fricatives primarily select the bare suffix. The voiced velar

---

[9] There are no LAs with base-final /f/, /dz/ and /dʑ/ in the native stratum. /tɕ/ and /dʑ/ have been excluded from **Figure 4** and the statistical analysis, because they each contain less than 3 items.

stop /g/ chooses the bare suffix or the intermorph *-aɲ-*. Most sonorants and /v/ show a significant preference for the bare suffix. The only exception is /n/, which in addition shows a preference for the intermorphs. An explanation for this finding will be provided below, where the role of extrasyllabicity is elaborated. Labial stops and voiced coronal fricatives select either the bare suffix (more often) or the intermorph *-aɲ-* (less often).
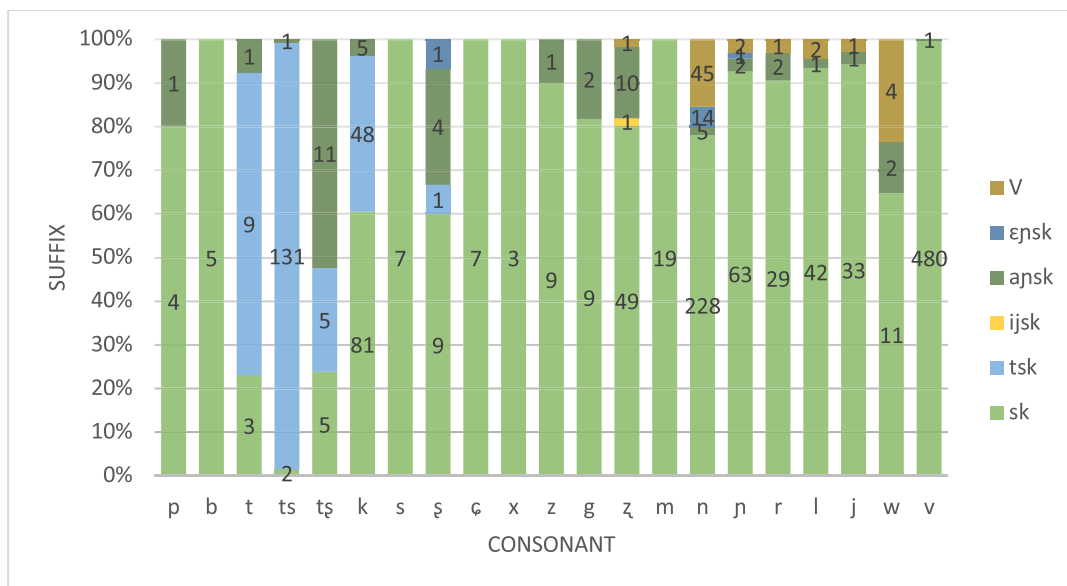
Figure 4: Native stratum. Observed affix and vowel insertion preferences depending on the base-final consonant.

cons.	preferred	dispreferred
voiceless non-labial stops and affricates		
/ts/	-tsk, -aɲ-sk	-sk, -ɛɲ-sk, -ij-sk, V
/t/	-tsk, -aɲ-sk	-sk, -ɛɲ-sk, -ij-sk, V
/k/	-tsk, -aɲ-sk	-sk, -ɛɲ-sk, -ij-sk, V
/tʂ/	-tsk, -aɲ-sk	-sk, -ɛɲ-sk, -ij-sk, V
voiceless non-labial fricatives		
/s/	-sk	-tsk, -aɲ-sk, -ij-sk, -ɛɲ-sk, V
/ɕ/	-sk	-tsk, -aɲ-sk, -ij-sk, -ɛɲ-sk, V
/ʂ/	-sk, -aɲ-sk, -tsk	-ij-sk, V
/x/	-sk	-tsk, -ij-sk, -ɛɲ-sk, V

(Contd.)

cons.	preferred	dispreferred
voiced velar stop		
/g/	-sk, -aɲ-sk	-tsk, -ɛɲ-sk, -ij-sk, V
sonorants		
/m/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/n/	-sk, -aɲ-sk, -ɛɲ-sk	-tsk
/ɲ/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/r/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/l/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/j/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/w/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
voiced labiodental fricative		
/v/	-sk	-ɛɲ-sk, -ij-sk, -tsk, V
labial stops		
/p/	-sk, -aɲ-sk	-ɛɲ-sk, -ij-sk, -tsk, V
/b/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
voiced coronal fricatives		
/z/	-sk, -aɲ-sk	-ɛɲ-sk, -ij-sk, -tsk, V
/ʐ/	-sk, -aɲ-sk	-ɛɲ-sk, -tsk, V

Table 3: Native stratum. Preferred and dispreferred ways of forming LAs according to the base-final consonant predicted by the multinomial analysis.

Included in the category of bases that end in /k/ in **Figure 4** are bases that end in /sk/. Such bases are singled out here because the base final sequence /sk/ is formally identical to the bare suffix -∅-*sk*. In LAs derived from bases in /sk/ the bare suffix is invariably chosen (in all 63 of such cases). Crucially, the LAs contain a single /sk/ sequence, which means that the resulting string /sk/ is shared between the base and the suffix, as exemplified in (9).[10] This will be analyzed as an instance of multiple correspondence in section 5.8.

(9)   sk + -sk- → -sk-
       bjɛlsk 'Bielsk' → bjɛl-sk-i

---

[10] The statistical analysis does not investigate the base-final sequence /-sk/, as it deals with single base-final consonants.

Figure 5 displays the impact of the factor *extrasyllabic*. In order to avoid an extrasyllabic consonant in a potential LA with the bare suffix -∅-*sk*, three strategies are statistically significant: the usage of vowel insertion (z = 6.11, p < .001), and two of the intermorphs: -*aɲ*- (z = 148.84, p < .001) and -*ɛɲ*- (z = 42.92, p < .001).
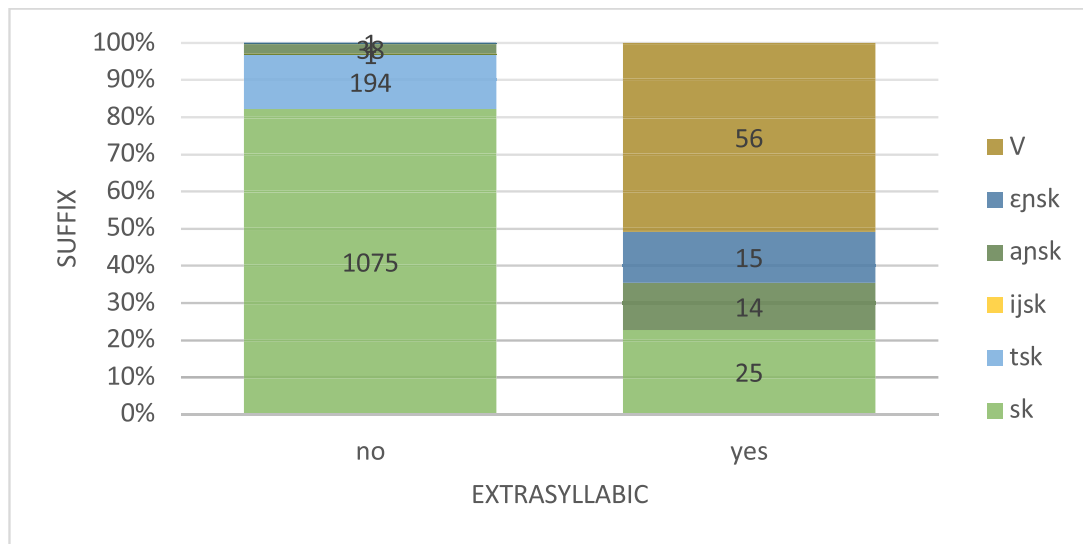

Figure 5: Native stratum. Observed affix and vowel insertion preferences depending on extrasyllabicity.

The interaction of *consonant* and *extrasyllabic* shows which suffixes are chosen when a particular consonant (sonorant) is extrasyllabic as opposed to when the consonant is not extrasyllabic in a potential LA with the bare suffix -∅-*sk*. The role of this interaction is visible in the graphs in **Figures 6** and **7**. **Figure 6** shows suffix distribution for the sonorants that do not appear as extrasyllabic in a potential LA with -∅-*sk*. These are usually sonorants that are preceded by a vowel in the base: VR. Such sonorants almost exclusively select the bare suffix -∅-*sk*, e.g. /tɔruɲ/ 'Toruń' > /tɔruɲ-sk-i/. We conclude that the sonorants that are not potentially extrasyllabic overwhelmingly select the bare suffix -∅-*sk*.

Figure 7 shows affix distribution for potentially extrasyllabic sonorants, that is, those which are preceded by an obstruent: VOR, e.g. /krɔsn-ɔ/ 'Krosno' > /krɔɕɲ-ɛɲ-sk-i/. When a LA with the bare suffix -∅-*sk* is formed, such sonorants are extrasyllabic. Compared with **Figure 6**, **Figure 7** shows no consistent preference for the bare suffix. While /j/ and /l/ exhibit a preference for the bare -∅-*sk* (more than 50% of cases), the remaining sonorants show a preference for various intermorphs or vowel insertion. The fricative /v/ is not shown, as it does not appear in "an extrasyllabic position" in native LAs.
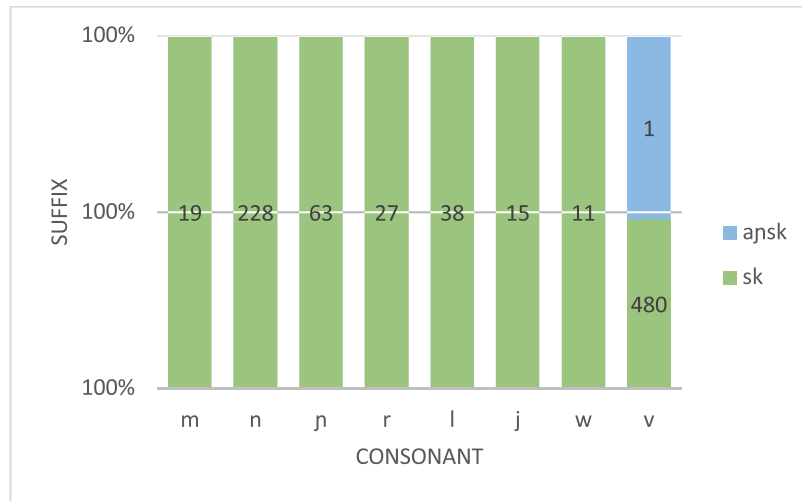
Figure 6: Native stratum. Observed affix and vowel insertion preferences depending on the base-final consonant when not extrasyllabic.
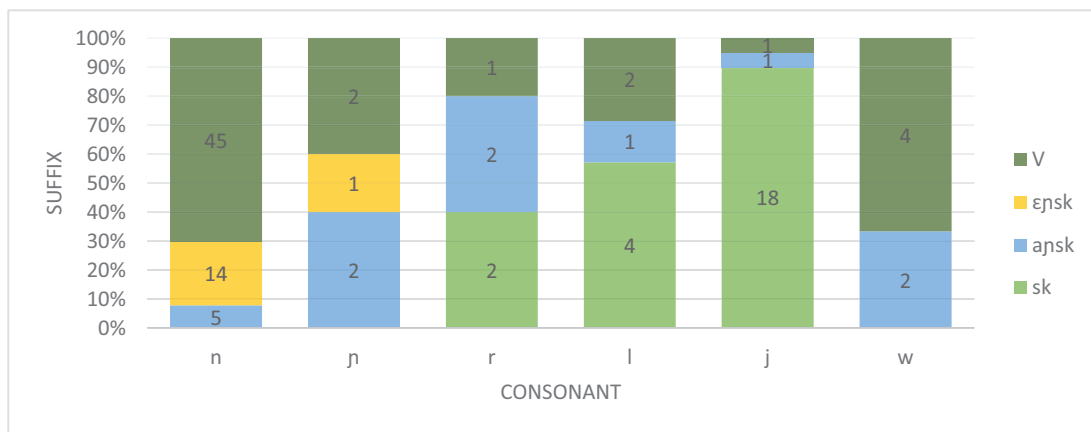
Figure 7: Native stratum. Observed affix and vowel insertion preferences depending on the base-final consonant when extrasyllabic.

Table 4 looks in more detail at which preferences and dispreferences are statistically significant for each sonorant when extrasyllabic, based on the results of the multinomial analysis. Extrasyllabic /n/ and /ɲ/ are avoided using the intermorphs *-aɲ-* and *-ɛɲ-* and vowel insertion. Extrasyllabic /r/ and /l/ are avoided using the intermorph *-aɲ-*. Extrasyllabic /w/ is avoided using the intermorph *-aɲ-* and vowel insertion. Finally, extrasyllabic /j/ is usually deleted and the bare suffix *-∅-sk* is used. These results suggest that when the sonorants /n/, /ɲ/, /r/, /l/ and /w/ are extrasyllabic in a potential LA with *-∅-sk*, specific strategies such as an intermorph or vowel insertion are recruited to preserve them. On the other hand, extrasyllabic /j/ tends to be deleted and the bare suffix is selected.

	preferred	dispreferred
when extrasyllabic		
/n/	-aɲ-sk, -ɛɲ-sk, V	-sk, -tsk, -ij-sk
/ɲ/	-aɲ-sk, -ɛɲ-sk, V	-sk, -tsk, -ij-sk
/r/	-aɲ-sk	-sk, -tsk, -ɛɲ-sk, -ij-sk, V
/l/	-aɲ-sk	-tsk, -ɛɲ-sk, -ij-sk, V
/j/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/w/	-aɲ-sk, V	-sk, -tsk, -ɛɲ-sk, -ij-sk

Table 4: Native stratum. Preferred and dispreferred ways of forming LAs according to the base-final consonant when extrasyllabic predicted by the multinomial analysis.

4.3 Foreign stratum

The multinomial analysis of the foreign LAs is designed in a similar way to the analysis of the native stratum. Appendix A (Analysis 3) offers more detailed results of the analysis. The graph in **Figure 8** presents the usage of vowel insertion and various affixes depending on the base-final consonant. A cursory inspection of the graph suggests that the intermorphs are used more frequently than in the native stratum, which accords with the results of the analysis of the whole set of LAs presented in section 4.1.

Figure 8: Foreign stratum. Observed affix and vowel insertion preferences depending on the base-final consonant.

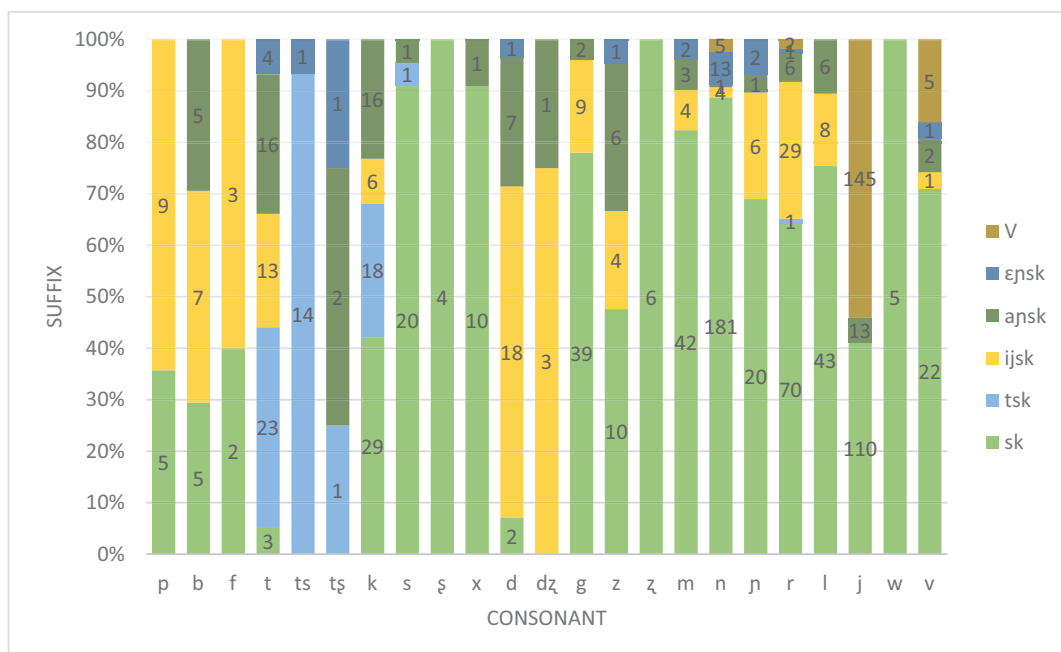The affix preferences depend on the base-final consonant, as shown in **Table 5**, where the results of the multinomial analysis are provided. Dispreference for the bare suffix -∅-*sk* is diagnosed when -∅-*sk* and vowel insertion are chosen in 40% or less of cases. Voiceless non-labial stops show a preference for -*tsk*, but the intermorphs are also used. This class shows a dispreference for -∅-*sk* and vowel insertion. Voiceless non-labial fricatives predominately select -∅-*sk*. The voiced velar stop /g/ shows a preference for -∅-*sk*, but the intermorph -*ij*- is also commonly recruited. Sonorants predominately choose -∅-*sk*, but the intermorphs and vowel insertion are also used. The two labiodental fricatives behave differently from each other. While /f/ selects mostly -*ij*-, /v/ chooses the bare suffix -∅-*sk*, the intermorphs -*ɛɲ*-, -*ij*-, or vowel insertion. We return to the usage of vowel insertion in the context of /v/ below. Labial stops tend to select the intermorphs -*ij*- and -*aɲ*-. The voiced coronal stop and affricate select various intermorphs. The voiced coronal fricatives behave differently. While both /z/ and /ʐ/ show a preference for -∅-*sk*, /z/ in addition is likely to select various intermorphs.

	preferred	dispreferred
voiceless non-labial stops and affricates		
/ts/	-tsk, -ɛɲ-sk	-sk, -aɲ-sk, -ij-sk, V
/t/	-tsk, -aɲ-sk, -ij-sk, -ɛɲ-sk	-sk, V
/k/	-tsk, -aɲ-sk, -ij-sk	-sk, -ɛɲ-sk, V
/tʂ/	-tsk, -aɲ-sk, -ɛɲ-sk	-sk, -ij-sk, V
voiceless non-labial fricatives		
/s/	-sk	-tsk, -aɲ-sk, -ij-sk, -ɛɲ-sk, V
/ʂ/	-sk	-tsk, -aɲ-sk, -ij-sk, -ɛɲ-sk, V
/x/	-sk	-tsk, -ij-sk, -ɛɲ-sk, V
voiced velar stop		
/g/	-sk, -ij-sk	-tsk, -ɛɲ-sk, V
sonorants		
/m/	-sk, -ij-sk, -ɛɲ-sk	-tsk, V
/n/	-sk, -ɛɲ-sk, -ij-sk	-aɲ-sk, -tsk, V
/ɲ/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
/r/	-sk, -ij-sk, -ɛɲ-sk, V	
/l/	-sk, -ij-sk	-ɛɲ-sk, -tsk, V

	preferred	dispreferred
/j/	-sk	-ɛɲ-sk, -ij-sk, -tsk, V
/w/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V
labiodental fricatives		
/v/	-sk, ɛɲ-sk, -ij-sk, V	-tsk
/f/	-ij-sk	-sk, -aɲ-sk, -ɛɲ-sk, -tsk, V
labial stops		
/p/	-ij-sk	-sk, -aɲ-sk, -ɛɲ-sk, -tsk, V
/b/	-aɲ-sk, -ij-sk	-sk, -ɛɲ-sk, -tsk, V
Voiced coronal stop and affricate		
/d/	-aɲ-sk, -ɛɲ-sk, -ij-sk	-sk, -tsk, V
/dʐ/	-aɲ-sk, -ij-sk	-sk, -ɛɲ-sk, -tsk, V
Voiced coronal fricatives		
/z/	-sk, -aɲ-sk, -ɛɲ-sk, -ij-sk	-tsk, V
/ʐ/	-sk	-aɲ-sk, -ɛɲ-sk, -ij-sk, -tsk, V

Table 5: Foreign stratum. Preferred and dispreferred ways of forming LAs according to the base-final consonant predicted by the multinomial analysis.

Similarly to the native stratum, in the foreign stratum when the base ends in /sk/, the bare suffix -∅-*sk* is almost always selected (27 out of 28 such cases). As shown in (10), the /sk/ string in such LAs corresponds both to the root-final sequence and to the suffix.

(10)   sk + -sk- → -sk-
         arxangjɛlsk 'Arkhangelsk' → arxangjɛl-sk-i

Just like in native LAs, in foreign LAs the factor *extrasyllabic* plays an important role. As evident in **Figure 9**, one strategy is particularly commonly used to avoid extrasyllabic consonants in LAs with -∅-*sk*: vowel insertion ($z = 103.82$, $p < .001$). The intermorphs -*aɲ*- ($z = 70.75$, $p < .001$) -*ɛɲ*- ($z = 25.27$, $p < .001$) and -*ij*- ($z = 73.43$, $p < .001$) are also used but less often.

The graphs in **Figures 10** and **11** illustrate the interaction of *consonant* and *extrasyllabic*. **Figure 10** shows sonorants that do not appear in the extrasyllabic context. While there is a general preference for the bare suffix, certain sonorants show a significant tendency to use the intermorphs or vowel insertion, e.g. /r/, /l/ and /j/. The results in **Figure 10** can be directly compared with the results in **Figure 6**, where the trends in the native stratum were depicted. Unlike in the native stratum, in the foreign stratum the usage of the intermorphs after the non-extrasyllabic sonorants is clearly detectable (except for /ɲ/ and /w/).

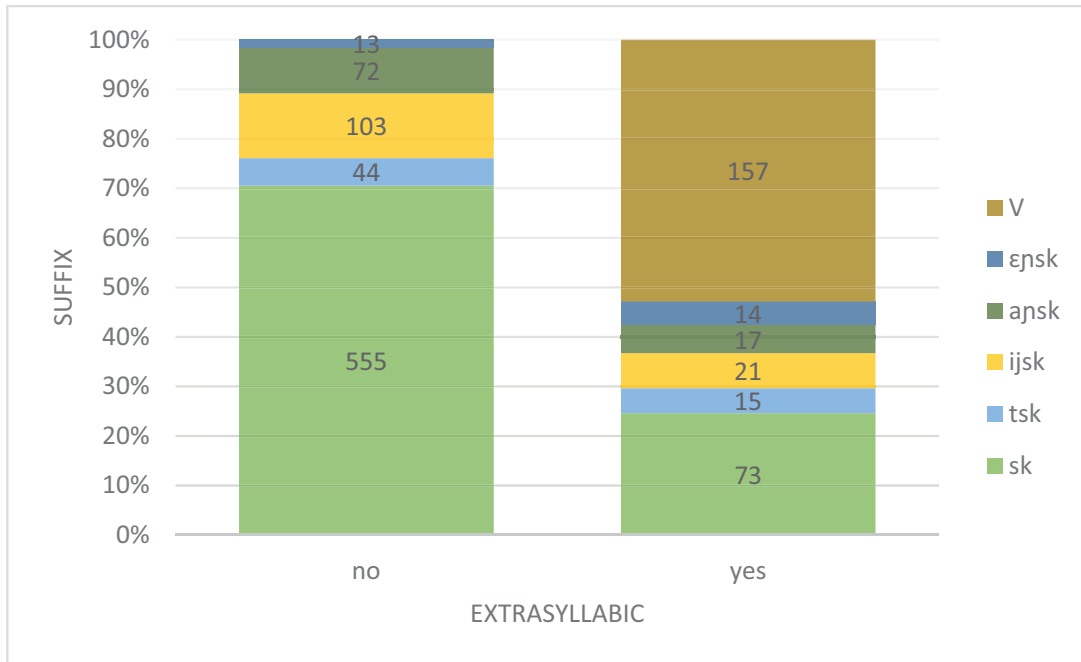**Figure 9:** Foreign stratum. Observed affix and vowel insertion preferences depending on extrasyllabicity.
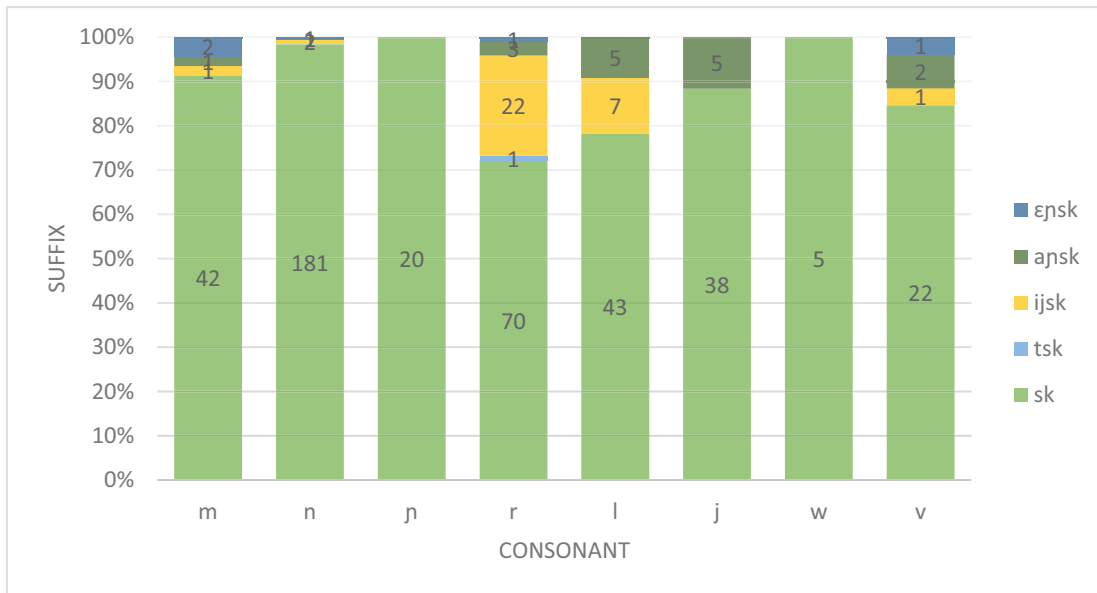

Figure 10: Foreign stratum. Observed affix and vowel insertion preferences depending on the base-final consonant when not extrasyllabic.

Figure 11 shows affix distribution in the context of various extrasyllabic sonorants in potential LAs with -∅-*sk*. All the sonorants select various intermorphs or vowel insertion in the majority of cases. The usage of the bare suffix -∅-*sk* is found for /j/ only, but even in this case the usage of vowel insertion and one of the intermorphs (-*aɲ*-) predominates.
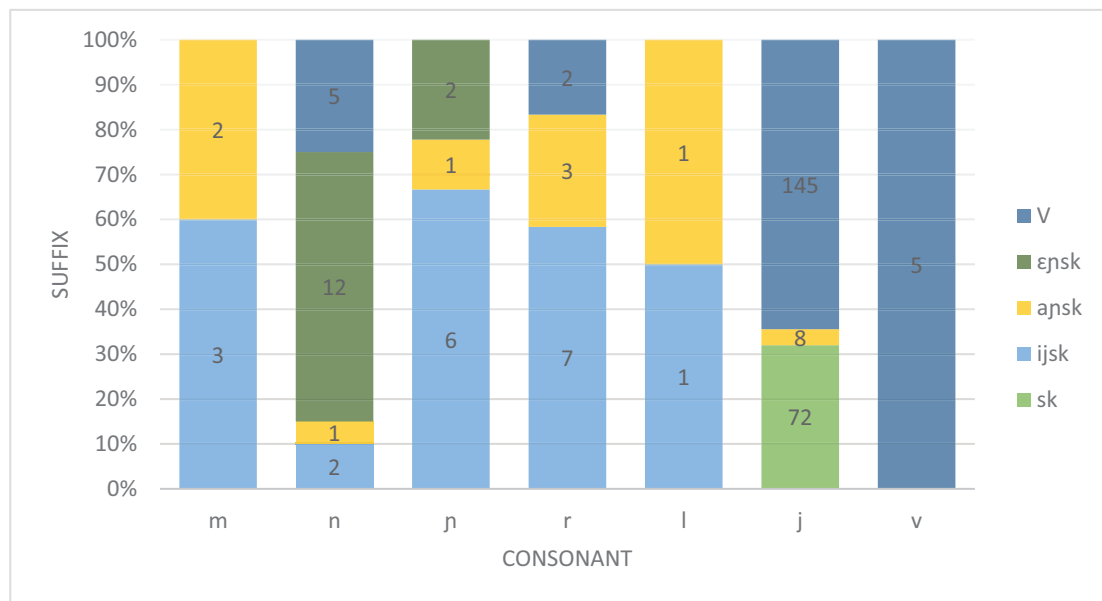
Figure 11: Foreign stratum. Observed affix and vowel insertion preferences depending on the base-final consonant when extrasyllabic.

Table 6 details which preferences and dispreferences in affix selection are statistically significant for each potentially extrasyllabic consonant. Extrasyllabic /m/, /n/, /ɲ/ and /r/ are avoided using various intermorphs and vowel insertion. Certain strategies exhibit specialization. The intermorph -*ɛɲ*- is used exclusively in the context of the coronal nasals /n/ and /ɲ/. The glide /j/ merits a closer analysis. Judging by the graph in **Figure 11**, vowel insertion occurs in the majority of cases. However, vowel insertion comes out as not statistically significant in the results shown in **Table 6**. This is caused by a considerable number of instances in which the potentially extrasyllabic /j/ selects the bare suffix -∅-*sk* and is deleted. Avoidance of "an extrasyllabic /v/" leads to the usage of vowel insertion. The observed productive use of vowel insertion as a strategy to avoid sequences where /v/ is flanked by two obstruents, resembling a trapped sonorant, may thus be a remnant of its past status as a sonorant, e.g. /padv-a/ 'Padova' > /padɛf-sk-i/, */patf-sk-i/. These results provide confirmation that intermorph and vowel insertion are strategies aimed at avoiding extrasyllabic sonorants and "trapped" /v/.

	preferred	dispreferred
when extrasyllabic		
/m/	-aɲ-sk, -ij-sk	-sk, -tsk, -ɛɲ-sk, V
/n/	V, -ɛɲ-sk, -aɲ-sk, -ij-sk	-sk, -tsk
/ɲ/	-aɲ-sk, -ɛɲ-sk, -ij-sk	-sk, -tsk, V
/r/	V, -aɲ-sk, -ij-sk	-sk, -tsk, -ɛɲ-sk
/l/	-aɲ-sk, -ij-sk	-sk, -tsk, -ɛɲ-sk, V
/j/	-sk	-tsk, -aɲ-sk, -ɛɲ-sk, -ij-sk, V
/v/	V	-sk, -tsk, -aɲ-sk, -ɛɲ-sk, -ij-sk

Table 6: Foreign stratum. Preferred and dispreferred ways of forming LAs according to the base-final consonant when extrasyllabic predicted by the multinomial analysis.

4.4 Summary

The analysis has identified important generalizations pertaining to the selection of affixes and vowel insertion in the two strata. They are given in **Table 7**.

Both strata		
	a.	Sonorants function as a class in that they show a preference for -∅-*sk*-.
	b.	The voiced labial fricative /v/ functions as a sonorant for two purposes: preference for -∅-*sk*- and the resolution of extrasyllabic sequences (vowel insertion).
	c.	Voiceless non-labial fricatives select -∅-*sk*-; the fricatives are deleted.`
	d.	The voiced velar stop /g/ selects -∅-*sk*-; the stop is deleted.
	e.	Voiceless non-labial stops and affricates select -*tsk*-; the stops/affricates are deleted (in fact, they undergo coalescence, see below).
	f.	Velar stops /k/ and /g/ are not treated uniformly: /k/ selects -*tsk*, while /g/ selects -∅-*sk*- and is deleted.
	g.	Extrasyllabic sonorants are avoided using vowel insertion or an intermorph.
	h.	Non-uniform treatment of the same inputs: extrasyllabic consonants are avoided using vowel insertion or various intermorphs.
	i.	The choice of an intermorph depends on the specific extrasyllabic consonant.
	j.	There is a categorical avoidance of repeated /sk/. -∅-*sk*- is selected in the great majority of cases.

(Contd.)

The native and foreign strata compared		
	k.	Most consonants (except /s/, /ʂ/, /ʐ/ /ɲ/ and /w/) are significantly more likely to select the intermorphs in the foreign than in the native stratum.
	l.	Non-extrasyllabic sonorants practically do not select the intermorphs in the native stratum. In the foreign stratum such sonorants show an affinity for various intermorphs. This is a categorical difference.
	m.	The intermorph -*ij*- is used almost exclusively in the foreign stratum. This is a categorical difference.

Table 7: Main generalizations in the two strata.

5 Constraint-based analysis

The following formal analysis rests on the assumption that the input contains all the possible intermorphs along with the adjectivizing suffix, as shown in (11) on the example of a LA derived from /alask-a/ 'Alaska'. First, the preferences identified in the statistical analysis are formalized as constraints which are responsible for the morphological and phonological composition of the generated outputs. Then, a probabilistic framework, Noisy Harmonic Grammar (Boersma & Pater 2016), is used to assess the relative importance of the constraints for the corpus data.

(11)   alask + {ij, aɲ, ɛɲ} + sk

At the outset, it should be noted that the constraints and schemas used in this analysis are construction-specific, that is, they apply in the grammar of LAs in derived environments. I do not claim that they have identical effects within morphemes or in the grammars of other constructions (see Inkelas 2014, esp. Chapter 8 for evidence for construction-specific effects). For example, the sequence /t/ + /s/ undergoes coalescence to yield the affricate /tʂ/ in LAs, e.g. /rabat/ 'Rabat' > /raba-tʂk-i/ but not in monomorphemic words, e.g. /dʑудʑitsu/ 'Ju-jitsu', where /t/ and /s/ are two distinct segments.

5.1 Product-oriented schemas

The proposed analysis refers to schemas and subcategorization frames, both of which have been used in previous studies. Paster (2006) argues that generalizations are phonologically arbitrary and include morphological information. For example, in Kaititj the ergative suffix appears as /-ŋ/ after disyllabic stems and as /-l/ after trisyllabic stems. Paster (2006) and Embick (2010) use subcategorization frames which make reference to phonological and lexical information. The Kaititj pattern is represented in terms of the subcategorization frames in (12), which state that disyllabic bases correspond to ergatives formed using /-ŋ/, while trisyllabic bases correspond to ergatives with the suffix /-l/ ("σ" stands for syllable.). Although the formula that captures

the generalization makes reference to phonological vocabulary, in this case the syllable, this instance of allomorph selection is phonologically arbitrary in the sense that it does not in any way improve phonological well-formedness (Paster 2006).

(12) Kaititj ergative

σσ ↔ -ŋ $_{ergative}$

σσσ ↔ -l $_{ergative}$

Product-oriented schemas (Bybee 2001; Booij 2010) are useful in capturing the distribution of allomorphs. Product-oriented schemas define the shape of the output without reference to the input. The observed preference for -∅-*sk* after sonorants (especially in the native stratum) can be explained once we notice that all the intermorphs show the structure V + R (V stands for a vowel and R for a sonorant). Based on the results of the quantitative analysis, it is proposed that the product-oriented schema R-sk regulates the shape of LAs. The two subcategorization frames in (13) capture the impact of the schema. When the base ends in a sonorant (R), the bare suffix is selected in a LA, as represented in (13a). On the other hand, when the base does not end in a R (the elsewhere condition), an intermorph is selected, (13b). As the three available intermorphs /-ij-/, /-aɲ-/ and /-ɛɲ-/ have the structure VR, the output of both subcategorization frames, complies with the R-sk schema.

(13) a. [R]$_{Base}$ ↔ -sk-$_{LA}$

b. elsewhere: […]$_{Base}$ ↔ [-VR-]$_{Intermorph}$-sk-$_{LA}$

There is good reason to believe that the class of sonorants (R) in the R-sk schema should be extended to include /v/. First, the analysis of the data in the previous sections demonstrates that the voiced labial fricative /v/ functions as a sonorant for two purposes: preference for ∅-*sk*- and the resolution of extrasyllabic sequences (vowel insertion). Second, Polish /v/ originated from a sonorant in Common Slavic (Bethin 1998: 203). In several Slavic languages it retains its sonorant properties. In Russian, it is pronounced as an obstruent and undergoes voice assimilation. However, unlike obstruents, it does not trigger voice assimilation. Lightner (1965) posits that the Russian /v/ is underlyingly a sonorant, /w/. The same pattern occurs in Czech (Short 1993). In many dialects of Ukrainian, it is both pronounced as a sonorant, [ʊ] or [w], and patterns as one with respect to voice assimilation (Czaplicki 2003). Although in Polish [v] both triggers and undergoes voice assimilation, it is possible that at least for some patterns [v] functions together with sonorants, as it is often the case that morphology and phonology lag behind phonetic change. For example, modern Polish /ʂ ʐ tʂ dʐ/ are described in phonetic studies as retroflexes (Hamann 2002) but function together with palatal consonants for certain morphophonological patterns (Gussmann 2007), the latter fact reflecting their historical origin. In the revised formulation of the schema in (14) "R" stands for "consonantal sonorants + /v/". The formulation is consistent

with Mielke (2008), who provides evidence that classes of sounds that function together in phonology can be language specific, as opposed to universal. For example, in Evenki (Nedjalkov 1997: 320, 175) suffix-initial /v s g/ change to nasals when they follow nasal consonants, but other consonants do not nasalize in this position. Notably, /g/ and /s/ undergo the change, but /d/ does not. Therefore, it is not possible to say that alveolars (or stops) undergo the change, but velars (or fricatives) do not, or vice versa. Such phonetically unnatural patterns cannot be insightfully described using the familiar set of distinctive features and must instead refer to some arbitrary groupings of segments. Further instances of phonetically arbitrary generalizations will be identified later in this paper.

(14)   R-sk-$_{LA}$ (final version)
       R = /m ɲ r l j w v/

LAs that comply with the schema in (14) emerge in three different ways. In adjectives like the one in (15a), the sequence *R-sk* results from the use of an intermorph, while in the adjectives instantiated in (15b), a sonorant appears at the end of the base. Therefore, the latter adjective does not require an intermorph to comply with the schema. *R-sk* strings in LAs like the ones in (15c–e) emerge through phonological operations such as insertion, deletion or mutation.

(15)   *Ways of forming LAs that are compliant with the R-sk schema*
       a.   marɔk-ɔ 'Morocco'   marɔk-aɲ-sk-i   intermorph
       b.   ukrain-a 'Ukraine'   ukraiɲ-sk-i   no intermorph
       c.   gvinɛ-a 'Guinea'   gvinɛj-sk-i   insertion
       d.   sandɔmjɛʐ-a 'Sandomierza' gen.sg.   sandɔmjɛr-sk-i   mutation
       e.   nɔv-ɨ jɔrk 'New York'   nɔvɔjɔr-sk-i   deletion

5.2 Avoidance of the intermorphs

In the formal analysis of affix selection in LAs a general constraint against intermorphs will be used. *STRUCT penalizes phonological structure (Kager 1999: 404). It is based on the idea that structures without a specific morphosyntactic meaning should be avoided. Intermorphs, which are often described as suffix extensions (Kreja 1989), fall into this category. Suffixes such as *-sk-* are shielded from *STRUCT by means of higher ranked constraints (omitted from the tableaux) that require realization of morphosyntactic features.[11] Evaluations of LAs derived from bases

---

[11] This analysis is based on the idea that the input to an evaluation is in fact a set of morphosyntactic and semantic features that the speaker wishes to express, called the *intent* by Zuraw (2010). High-ranked constraints enforce morphosyntactic and semantic identity between the intent and output (see Zuraw 2010 for details). Insofar as intermorphs do not carry morphosyntactic meaning, they are not protected by such constraints. It follows that intermorphs should not be selected unless violations of higher ranked constraints are at stake. However, we cannot rule out a possibility that the intermorphs have acquired some specific meanings through language use. For example,

ending in a VO (vowel + obstruent) sequence and a VR (vowel + sonorant) sequence are given in (16) and (17). "IM" stands for an intermorph, + + stands for a strong boundary and + signals a weak boundary (following the distinction defined in section 2). MAX penalizes consonant deletion. In (16) the schema requiring R-sk in a LA enforces the selection of an intermorph. In (17) the final R is present in the base, which means that using an intermorph is not required to satisfy the schema.

(16)	VO+ + IM + sk	MAX	R-sk	*STRUCT
	a. VO--sk-		*!	
☞	b. VO--IM-sk-			*
	c. V--sk-	*!	*	

(17)	VR+ + IM + sk	MAX	R-sk	*STRUCT
☞	a. VR--sk-			
	b. VR--IM-sk-			*!
	c. V--sk-	*!	*	

5.3 Extrasyllabic consonants

The corpus data show three commonly attested strategies to avoid extrasyllabic sonorants in potential LAs with the suffix -∅-*sk*: vowel insertion, usage of an intermorph or sonorant deletion as illustrated in (18). This can be viewed as an instance of conspiracy.

(18) vigr-ɨ 'Wigry'   vigjɛr-sk-i
      pilzn-ɔ 'Pilzno'   pilzn-ɛɲ-sk-i
      rumuɲj-a 'Romania'   rumuɲ-sk-i

Sonority Sequencing Principle (SSP) will be used to represent the dispreference for extrasyllabic sonorants. Vowel insertion is penalized by the constraint DEP. MAX penalizes consonant deletion. The evaluation in (19) shows the ranking necessary to generate two of the three attested ways to avoid extrasyllabic sonorants: candidate (b) shows vowel insertion and candidate (d) uses an intermorph. The evaluation of the same input in (20) produces the third of the attested outputs: a candidate with sonorant deletion, candidate (c). SSP dominates all the other constraints in (19)

Endersen (2015) fit a distributional semantic model to seemingly semantically empty morphs in Russian and discovered that they in fact have meanings. It should be noted though that the intermorphs differ from other morphs by having restricted distribution. They are extensions of the adjectivizing suffix -*sk*. I leave for future research to determine whether the intermorphs have acquired meanings.

and (20). To generate non-uniformity of the outputs the ranking of the constraints MAX, R-sk, *STRUCT and DEP must be probabilistic rather than strict. This issue is addressed in section 5.10, where I use a model that allows for assigning probability distributions to outputs using weighted constraints.[12] None of the strategies are used in (21), where there is no potential extrasyllabic sonorant and SSP is moot.

(19)	VOR + + IM + sk	SSP	MAX	R-sk	*STRUCT	DEP
	a. VOR--sk-	*!				
☞	b. VO<u>V</u>R--sk-					*
	c. VO--sk-		*!	*		
☞	d. VOR--IM-sk				*	

(20)	VOR + + IM + sk	SSP	*STRUCT	DEP	MAX	R-sk
	a. VOR--sk-	*!				
	b. VO<u>V</u>R--sk-			*!		
☞	c. VO--sk-				*	*
	d. VOR--IM-sk		*!			

(21)	VR + + IM + sk	SSP	MAX	R-sk	*STRUCT	DEP
☞	a. VR--sk-					
	b. VR<u>V</u>--sk-			*!		*
	c. V--sk-		*!	*		
	d. VR--IM-sk				*!	

How do we account for the fact that different consonants use different strategies to avoid violations of SSP? In addition to markedness constraints such as SSP, consonant-specific schemas specifying which strategy is used are necessary, as exemplified in (22). The schemas are supported by the results of the statistical analysis in section 4. The schemas in (a) and (b) define the possible resolutions of an extrasyllabic /n/ (marked as "Cn"). The schema in (c) refers to an extrasyllabic /r/ (marked as "Cr"). Their ranking (or weighting) is probabilistic, reflecting the

---

[12] In addition, the preferred strategies for dealing with extrasyllabic consonants differ in the two lexical strata: native and foreign.

attested variation in the outputs.[13] The idea that schemas (subcategorization frames) are ranked (or weighted) just like OT constraints is not new.[14]

(22) a. Cn → -ɛɲ-
       b. Cn → -aɲ-
       c. Cr → -aɲ-

5.4 Voiceless non-labial fricatives

For the remaining factors, there is also variability in the selection of the output. However, a clear preference for one of them can be identified. In the analysis, unless indicated otherwise, we consider the dominant output, as reflected in the type frequency data in the native stratum (provided in section 4.2). The hypothesized preferences are verified in the next section, where I model the grammars of LAs using a probabilistic framework. First, I investigate base-final voiceless non-labial fricatives (/s ɕ ʂ x/), as this factor promotes the selection of /-∅-sk/. As exemplified in (23), such fricatives tend to fuse with the following /s/ of /-sk-/. It is important to observe that the conditioning context is absent from the output but present in the input. That is, affix distribution in the output cannot be predicted from the output. Information from the input is crucial. We return to this issue in section 7.

(23) nɨs-a 'Nysa'   nɨ-sk-i
       pʂasnɨʂ 'Przasnysz'   pʂasnɨ-sk-i

The schema and the constraints responsible for this outcome are given in (24). The schema in (a) mandates that the bare suffix is selected after S, which stands for a voiceless non-labial fricative. It is proposed that such positive schemas are useful, as certain consonants or groups of consonants show an affinity for a particular way of forming LAs (see also the various consonant-specific resolutions of extrasyllabic consonants mentioned above).[15] The constraints *Ss and IDENT[+anter] enforce coalescence and are jointly labeled COALS. *Ss penalizes a voiceless

---

[13] Determining the relative impact of such consonant-specific schemas is beyond the scope of this paper.

[14] For example, Burzio (2002) proposes a theory of surface-to-surface relations in which representations are clusters of entailments that directly condition other representations. The entailments are conceptualized as subcategorization frames that reference both the input and the output. Such subcategorization frames are ranked, which reflects their relative importance.

[15] Kapatsinski (2021) adopts a different approach. He uses exclusively product-oriented schemas to derive morphophonological patterns. In his approach, product-oriented schemas are positive and negative. Extrapolating Kapatsinski's (2021) approach to handle the Polish LAs, a positive schema would activate the output /-sk-/, while negative schemas would suppress outputs such as /S-aɲ-sk-/ and /S-ij-sk-/. However, Kapatsinski's product-oriented schemas cannot accommodate generalizations whose conditioning context is absent from the output (i.e. source-oriented generalizations), see section 7.

non-labial fricative followed by /s/. IDENT[+anter] makes sure that the two fricatives fuse to /s/, as opposed to /ɕ/, /ʂ/ or /x/. UNIFORM penalizes coalescence (McCarthy & Prince 1995).

(24)   S→sk: [-voice, -labial, +cont] → ∅-sk

       COALS:

          *Ss: Assign one violation for any S + + s sequence, where S is a [-labial, +continuant] consonant and + + is a strong morpheme boundary within a word.

          IDENT[+anter]:   Assign one violation if an input segment that is [+coronal, +anterior] does not correspond to an output segment that is [+coronal, +anterior].

          UNIFORM: No element in the output has multiple correspondents in the input.

The tableau in (25) shows an evaluation of a LA like /pʂasnɨ-sk-i/ derived from /pʂasniʂ/ 'Przasnysz'. Two non-identical indices on one segment mean that two segments have been fused. Candidate (a) selects the wrong affix and fatally violates the relevant schema. Candidates (b) and (c) fail to apply coalescence and violate *Ss. Candidate (d) shows deletion, which is penalized by MAX. Candidates (e) and (f) show coalescence, but in candidate (f) the output of coalescence is /ʂ/, which violates IDENT[+anter]. Candidate (e) wins. The evaluation shows that UNIFORM must be ranked below the other constraints.

(25)	Sᵢ + + IM + sⱼk	S→sk	*Ss	IDENT[+anter]	MAX	R-sk	*STRUCT	UNIFORM
	a. ʂᵢ--IM-sⱼk-	*!					*	
	b. ʂᵢ--sⱼk-		*!			*		
	c. sᵢ--sⱼk-		*!			*		
	d. --sⱼk-				*!	*		
☞	e. --sᵢ,ⱼk-					*		*
	f. --ʂᵢ,ⱼk-			*!		*		*

5.5 Voiceless non-labial non-continuants

Base-final voiceless non-labial stops and affricates /t ts tʂ k/, represented as [-voice, -labial, -continuant], regularly choose the allomorph -tsk of the suffix -sk. The input sequence: a base-final stop or affricate plus the suffix initial /s/ yields the affricate /ts/, as schematized and illustrated in (26). This process will be conceptualized as an instance of coalescence. Recall from section 2 that /g/ does not participate in the coalescence, hence the restriction to [-voice]. Similarly, labial stops fail to coalesce with the following suffix, e.g. /wɛb-a/ 'Łeba' – /wɛp-sk-i/. Therefore, base-final voiceless non-labial non-continuants, /t ts tʂ tɕ k/, need to be considered

as a context favoring the allomorph /-tsk-/ of the adjectivizing suffix. Following (Jakobson et al. 1952), I assume that affricates are represented as strident stops. The fused [ts] acquires the feature [-continuant] from the non-labial stop or affricate and the remaining features, i.e. [-voice, +coronal, +anterior, +strident], from the fricative /s/.

(26) a. t ts tʂ k ([-voice, -labial, -continuant]) + -sk- → ∅-tsk-
      b. rabat 'Rabat'   raba-tsk-i
          dalmatsj-a 'Dalmatia' dalma-tsk-i
          tɨlitʂ 'Tylicz'   tɨli-tsk-i
          irak 'Iraq'   ira-tsk-i

The schema in (27) compels the usage of the bare suffix *-∅-sk* after a /T/, which stands for [-voice, -labial, -continuant]. The two constraints *Ts and IDENTT[-contin] coerce the coalescence /T/ + /s/ → /ts/. IDENTT[-contin] is restricted to voiceless stops and affricates, as /g/ does not participate in coalescence. IDENT[+anter], defined in (24) above, makes sure that the fused affricate is /ts/, rather than /tʂ/.

(27) T → ∅-sk   [-voice, -labial, -continuant] → ∅-sk
      COALT:
        *Ts: Assign one violation mark for any T + + s sequence, where T is a [-labial, -continuant] consonant and + + is a strong morpheme boundary within a word.
        IDENTT[-contin]:   Assign one violation mark if an input segment that is [-voice, -labial, -continuant] does not correspond to an output segment that is [-continuant].

In the evaluation of a LA with a base-final T in (28), candidate (a) is eliminated because it selects an intermorph. Candidate (b) fails, as it forgoes coalescence and as a consequence violates *Ts. Candidate (c) shows deletion, which brings about a violation of MAX. The remaining candidates show coalescence, but candidates (e) and (f) choose suboptimal outputs of coalescence and violate the relevant IDENT constraints. Candidate (d), which shows fusion to /ts/, comes out victorious.

(28)	T_i + + IM + s_jk	T→sk	*Ts	IDENTT [-contin]	IDENT [+anter]	MAX	R-sk	*STRUCT	UNIFORM
	a. T_i--IM-s_jk-	*!						*	
	b. T_i--s_jk-		*!				*		
	c. --s_jk-					*!	*		
☞	d. --ts$_{i,j}$k-						*		*
	e. --s$_{i,j}$k-		*!				*		*
	f. --tʂ$_{i,j}$k-				*!		*		*

5.6 Voiced non-labial obstruents

Voiced non-labial obstruents are less likely to fuse with the suffix *-sk* than their voiceless counterparts and instead are more likely than their voiceless counterparts to choose an intermorph. Another common strategy is to delete the voiced obstruent. Both strategies are exemplified in (29). However, the output is not uniform. For instance, /z/ selects all the intermorphs as well as the bare suffix in the foreign stratum (see section 4.3).

(29) piz-a 'Pisa'   piz-aɲ-sk-i
      xag-a 'the Hague'   xa-sk-i

The constraint IDENT[+voice] in (30) coerces the preservation of [+voice]. In the case at hand, it is violated when a fused voiceless segment in the output [s$_{1,2}$] corresponds to two input segments with differing specifications for voicing, e.g. /z/ and /s/. This dispreference can be observed in the data, e.g. /s/ most often fuses with /s/, but /z/ is more likely to appear with an intermorph in the foreign stratum, and /k/ fuses with /s/, but /g/ does not.[16] Voice Assimilation (VA) makes sure that obstruent clusters agree in voicing.

(30) IDENT[+voice]: Assign one violation if an input segment that is [+voice]
                      corresponds to an output segment that is not [+voice].
      VA: Obstruent clusters must agree in [voice].

The evaluation of a LA with a base-final /D/ in (31) shows the importance of IDENT[+voice]. Candidate (b) incurs a fatal violation of a phonotactic constraint, *Ts (it is applicable to voiceless and voiced non-continuants). Candidate (c) is eliminated because it shows fusion of /D/ and /s/ and the fused /ts/ lacks [+voice]. Candidate (d) preserves [+voice] in the fused segment but violates VA, as two adjacent obstruents differ in voicing. In accordance with the trends identified in the data, the second possible output is candidate (e), with deletion. For this candidate to win, *STRUCT must be promoted over MAX (the tableau is not shown for reasons of space). This again justifies the need for a probabilistic ranking of the constraints. In sum, the usage of an intermorph (candidate a) and deletion (candidate e) are preferred over coalescence for voiced obstruents.

(31)	D$_i$+ + IM +s$_j$k	*Ts	VA	IDENT[+voice]	MAX	R-sk	*STRUCT	UNIFORM
☞	a. D$_i$--IM-s$_j$k-						*	
	b. D$_i$--s$_j$k-	*!	*			*		
	c. --ts$_{i,j}$k-			*!		*		*
	d. --dz$_{i,j}$k-		*!			*		*
	e. --s$_j$k-				*!	*		

---

[16] There is a historical explanation for the different behavior of /k/ and /g/ in this context. /g/ underwent spirantization and the spirant was deleted before *-sk*: g → ɣ → ∅.

5.7 Labials

Labial stops /p b/ and the voiceless fricative /f/ (/f/ does not appear in the native stratum) are also likely to select an intermorph, though this preference is visible only in the foreign stratum, e.g. /tɛb-ɨ/ 'Thebes' > /tɛb-aɲ-sk-i/. Here coalescence is prevented by the constraint IDENT[+labial], which forbids changes in the [+labial] specification. In the evaluation in (32), IDENT[+labial] and IDENT[+anterior] militate against coalescence with a labial. Note that IDENT[+anterior] as formulated in (24) applies to coronals only. The change of a coronal to a labial in candidate (d) violates IDENT[+anter], as the coronal articulator, of which [+anterior] is a dependent (Hume & Clements 1995), has been deleted. Candidates with deletion, (e) and (f), incur fatal violations of MAX. The ranking of R-sk over *STRUCT compels the selection of the output with an intermorph (candidate a).

(32)	P_i + + IM + s_jk	IDENT[+labial]	IDENT[+anter]	MAX	R-sk	*STRUCT	UNIFORM
☞	a. P_i--IM-s_jk-					*	
	b. P_i--s_jk-				*!		
	c. --ts$_{i,j}$k-	*!			*		*
	d. --P$_{i,j}$k-		*!		*		*
	e. --s_jk-			*!	*		
	f. P_i--k-			*!	*		

5.8 Multiple correspondence

Multiple correspondence occurs when a given output string corresponds to two distinct input strings. For example, the string /sk/ serves both as part of the root and as an adjectival suffix in (33). The preference for the multiple correspondence of /sk/ is almost categorical in the data: 90 of 91 LAs with base-final /sk/ select the bare suffix -∅-*sk*.[17]

(33) sk + -sk- → -∅-sk-
      alask-a + -sk-i → ala-sk-i

The schema in (34) enforces the selection of -∅-*sk* after base-final /sk/. Multiple correspondence is often coerced by the avoidance of identical sounds or morphemes (Czaplicki 2022).[18] In OT formalization, a constraint against repeated elements, OCP (Goldsmith 1976), is ranked above a constraint penalizing multiple correspondence, UNIFORM, defined above. In accordance with the evidence suggesting that similarity avoidance is stronger in morphologically derived

---

[17] The factor base-final /sk/ was not included in the statistical analysis, as the latter was not concerned with base-final clusters. However, the observed near categorical behavior indicates that this factor plays an important role and for that reason it is included in the morpho-phonological analysis.

[18] Avoidance of identical elements is termed haplology.

environments than within single morphemes (Jurgec 2016; Czaplicki 2022), the OCP in (34) is morphologically restricted.

(34)   sk → -sk-
         OCP: No identical elements across a strong morphological boundary.[19]

In the evaluation of an LA derived from a base ending in /-sk/ in (35), indices are attached to the relevant /sk/ sequences to indicate their non-identity. The schema ensuring the selection of the bare suffix in this context is responsible for eliminating the candidate with an intermorph, (a). The presence of two /sk/ sequences across a strong morphological boundary in candidates (a) and (b) incurs a violation of the OCP. Candidate (c) with multiple correspondence wins as long as the OCP is ranked over UNIFORM.

(35)	sk_i + + IM + sk_j	sk → sk	OCP	UNIFORM
	a. sk_i--IM-sk_j-	*!	*	
	b. sk_i--sk_j-		*!	
☞	c. --$sk_{i,j}$-			*

5.9 Lexical strata

How can we account for the more common usage of the intermorphs in the foreign than in the native stratum? In the analyzed data, the intermorphs and vowel insertion are used in 36.6% of LAs of foreign origin, compared with 8.8% of LAs of native origin. Moreover, the intermorph *-ij-* is used exclusively in the foreign stratum. The different preference for the intermorphs can be derived from different constraint ranking (or weighting) in the two grammars: foreign and native. Two possibilities are considered. According to the first hypothesis, there is a dedicated constraint that requires that a LA contains an intermorph, INTERMORPH. This constraint is based on the idea that the intermorphs have become a marker of the foreign status of LAs. The constraint is a cover-term for constraints that refer to the usage of specific intermorphs in the context of particular consonants, e.g. *r* → *-ij-, t* → *-aɲ-*. INTERMORPH dominates *STRUCT in the foreign grammar. The second possibility is that there is no dedicated constraint requiring intermorphs and the difference between the two grammars rests on the different ranking of *STRUCT with respect to the other constraints regulating allomorph selection.[20] The hypothesized rankings of the constraints (both options) in the two grammars are given in (36).

---

[19] The adjacency of /sk/ in /sk-IM-sk/ rests on the idea that the morphological boundary between the intermorph and the suffix is weak, as the intermorphs always occur together with the *-sk* suffix and are thus a type of suffix extensions (Kreja 1989): /sk + +IM+sk/. As mentioned in section 2, there is evidence that affixes bind with a range of strengths to stems and other affixes. This shows that there is ranking in the decomposability of morphologically complex words. Intermorphs are less decomposable than other affixes.

[20] The relative importance of these constraints will be determined using a probabilistic framework in the next section.

(36)   a.   G_{Foreign}: INTERMORPH/other constraints > *STRUCT
       b.   G_{Native}: *STRUCT > INTERMORPH/other constraints

LAs derived from /saxar-a/ 'Sahara' (foreign) and /mazur-i/ 'Mazury' (native) serve to illustrate the observed discrepancies in the preference for the intermorphs in foreign and native words. Indices show whether a word belongs to the foreign or native stratum. In the evaluation of the foreign word in (37), the candidate with an intermorph wins, as LAs with the intermorphs are preferred in G_{Foreign}, reflecting the ranking of *STRUCT below the other relevant constraints. In the evaluation of the native word in (38), the candidate without an intermorph comes out victorious, as the ranking in G_{Native} inhibits the selection of the intermorphs. Thus, different rankings in the two grammars result in different outputs. As demonstrated by the results of the statistical analysis, the preferences in the affix distribution in LAs are gradient, rather than categorical. In the next section, I attempt to assess the importance of each constraint for affix distribution using a probabilistic framework. G_{Native} and G_{Foreign} are considered separately.

(37)	saxar-a_{Foreign} + + {ij, aɲ, ɛɲ} + sk	INTERMORPH/other constr.	*STRUCT
	a. saxar--sk-i	*!	
☞	b. saxar--ij-sk-i		

(38)	mazur-i_{Native} + {ij, aɲ, ɛɲ} + sk	*STRUCT	INTERMORPH/other constr.
☞	a. mazur--sk-i		*
	b. mazur--ij-sk-i	*!	

5.10 Grammar modeling

Classic OT models, which assume strict constraint ranking, are not sufficient to generate the attested variation in the selection of outputs of LAs. In an attempt to determine the gradient impact of various pressures I use a model that allows for assigning probability distributions to outputs, rather than predicting a single winner. Such probabilistic frameworks come in two flavors: they refer to either constraint ranking or constraint weighting. A Stochastic OT grammar (Boersma 1998) is a probability distribution over strict-ranking OT grammars. Each constraint is assigned a ranking value, which is used to determine the probability of a given constraint ranking after adding noise. When the ranking values of two constraints are far apart from each other, the ranking of the constraints is fixed. When the ranking values of two constraints are closer to each other, the ranking of the two constraints is more likely to be reversed and variation occurs. In contrast, Harmonic Grammar (Legendre et al. 1990) uses constraint weighting. Each constraint bears a real number that determines its importance in the selection of the winning candidates. In particular, the weighted sum of the constraint violations of each candidate, termed the harmony of the candidate, is used to select the winning candidates. Both frameworks can in principle be used to model the distribution of LAs. However, recent evidence indicates that Harmonic Grammar is more effective in modeling variation than Stochastic OT (Zuraw & Hayes

2017). For that reason, we use Harmonic Grammar, and more specifically, its implementation called Noisy Harmonic Grammar (NHG; Boersma & Pater 2016). What makes NHG probabilistic is that at each evaluation time, some Gaussian noise is added to each constraint's weight, which might lead to a change of the winning candidate.

The tableaux used in the simulations of the Polish data are given in Appendix B. The key constraints formulated in the previous sections have been used. As regards the constraints coercing coalescence, COALS is a shorthand for *Ss and IDENT[+anter], and COALT stands for *Ts and IDENT[-cont]. The tableaux list the inputs, output candidates and all the constraint violations (marked by "1"). Frequency values reflect the type frequency of the specific pattern in the data (how often a winning candidate occurs for each pattern). OTSoft (Hayes et al. 2013) was used to run two simulations: one for the native grammar and one for the foreign grammar.[21] The weighting of the constraints derived from the simulations is given in **Table 8**.[22]

Native grammar		Foreign grammar	
constraint	**weight**	**constraint**	**weight**
SSP	9.57151	SSP	6.102073
COALT	6.930417	COALT	4.814077
IDENT[+voice]	6.811049	IDENT[+voice]	4.63582
*STRUCT	**6.708513**	COALS	4.10475
DEP	5.244751	**S → sk**	**2.888474**
OCP	5.002984	IDENT[+labial]	2.884004
COALS	4.630497	OCP	2.880539
R-sk	**3.264369**	***STRUCT**	**2.564468**
MAX	2.571597	DEP	1.642409
IDENT[+labial]	2.312655	**R-sk**	**1.622213**
sk → sk	**0.559263**	MAX	0.836828
IM	0.341487	IM	0.335532
UNIFORM	0.046	sk → sk	0.005
S → sk	0.003	T → sk	0
T → sk	0	UNIFORM	0

Table 8: Constraint weighting in the native and foreign grammar.

---

[21] Default initial weights: all 0. Initial plasticity: 0.01; final plasticity: 0.001. Number of learning trials: 10,000,000. The only bias imposed is that weights may not be negative.

[22] As demonstrated in section 4, the usage of the intermorphs, particularly -ij-, is a likely cue to foreignness. Also, most consonants are significantly more likely to select the intermorphs in the foreign than in the native stratum. This is not evident in the simulations, as they use more general constraints, such as *STRUCT. What is evident, however, is that there are different preferences for the intermorphs in the two grammars.

The SSP is the constraint with the highest weight in both grammars, which means that extrasyllabic sonorants are strongly avoided. The most striking difference between the two grammars is the relative weight of *STRUCT. In the native grammar, *STRUCT is weighted among the first four most important constraints. In contrast, in the foreign grammar it belongs to a group of constraints with a medium weight. This difference suggests that the intermorphs are less preferred (i.e. more dispreferred) in the native than in the foreign grammar. The constraint coercing the usage of an intermorph, IM, is not particularly active in the two grammars. The constraint R-sk exhibits a medium weight in both grammars (though it is slightly more important in the native grammar, as compared with the foreign grammar, where the constraint is counterbalanced by a general preference for the intermorphs). This result provides evidence for the importance of the R-sk schema for allomorph selection in LAs. Further evidence for the impact of R-sk is provided below.

As regards the source-oriented schemas requiring the selection of the bare suffix, they have a small impact, with the notable exception of S → sk in the foreign grammar and perhaps sk → sk in the native grammar. More specifically, in the foreign grammar voiceless non-labial fricatives show a preference for the bare suffix (with coalescence), in spite of the general affinity of the other consonants for the intermorphs. The affinity of base-final /sk/ for the bare suffix is weighted higher in the native than in the foreign grammar, where its impact is small. This means that certain classes of segments may show different preferences in affix choice than others. Phonotactic constraints regulating coalescence (COALS and COALT) and similarity avoidance (OCP) are weighted high in both grammars. The fact that IDENT[+voice] and IDENT[+labial] are weighted high suggests that coalescence tends to be blocked for voiced obstruents and labial obstruents.[23] The relative weighting of DEP over MAX in the two grammars indicates that segment deletion is preferable over segment insertion when satisfaction of higher ranked constraints is at stake. In addition, DEP and MAX are weighted higher in the native than foreign grammar, which indicates that segmental faithfulness is respected more in the native than foreign grammar.[24]

The product-oriented schema R-sk turned out to be important in the simulations. Additional evidence for the relevance of the schema R-sk is based on the comparison of the relative frequency of final Rs in LAs (i.e. before the -sk suffix) and their base nouns. The latter frequency provides a baseline useful in determining whether final sonorants are overrepresented in LAs. The data in **Table 9** show that final Rs are more common in LAs than in their base nouns. The proportions of final Rs are 0.83 for LAs and 0.697 for their base nouns, with the difference between them amounting to 0.133. A binomial test in R (binom.test(2078, 2503, p = 0.697)) yields $p < .001$

---

[23] The high weight of phonotactic constraints in relation to R-sk is understandable once we consider the fact that the former target illicit sequences, while the latter captures a gradient pressure.

[24] The low weight of MAX is mainly due to the frequent deletion of potentially extrasyllabic /j/ in both grammars.

and a 95% confidence interval from 0.815 to 0.845. This provides additional support for the impact of the R-sk schema in LAs.[25]

	Yes	No	Total	Proportion of final Rs
LA	2,078	425	2,503	0.83
Base Noun	1,744	759	2,503	0.697

Table 9: Frequency of final Rs in LAs and their base nouns.

6 A probable role of lexical frequency in predicting variation

Why are some adjectives of foreign origin stable, e.g. /anglj-a/ 'England' → /angjɛl-sk-i/ and /kanad-a/ 'Canada' → /kanad-ɨj-sk-i/, while others appear in doublets or even triplets, e.g. /alask-a/ 'Alaska' → /ala-sk-i/, /alask-ɨj-sk-i/ and /alask-aɲ-sk-i/? It is claimed that lexical frequency plays an important role in morphological stability. LAs of high lexical frequency tend to be stored whole, while less frequent LAs are generated on the fly. I offer a preliminary analysis of the role of lexical frequency in predicting variation in LAs.

Variation in the data can be elucidated with the use of the dual-route model of lexical access (McQueen & Cutler 1998; Hay 2003; Plag 2012). According to this model, high-frequency words are stored whole, which implies that they should be more morphologically stable. In contrast, low-frequency words are derived on the fly from their component morphemes, which means that they are predicted to show more variability. Among LAs showing high token frequency in the corpus are /angjɛl-sk-i/ (845,200), /kanad-ɨj-sk-i/ (70,500) and /kɔpɛnxa-sk-i/ (9,211). These LAs are stable. On the other hand, multiple LAs showing lower token frequencies vacillate. **Table 10** shows the results of a search in the corpus for the LAs which show variability in the choice of a suffix. TkF stands for token frequency. It appears that all such LAs except one (i.e. /ira-tsk-i/) show a relatively low frequency. Several place names have two attested adjectives, while for others three adjectives are found, though often with a considerably different token frequency. All the LAs in **Table 10** refer to places of foreign origin.

In order to compare the token frequency of LAs with and without variation, combined frequencies of all LAs derived from a single noun were used. That is, if several LAs were attested for one noun, their token frequencies were combined. The median of the (combined) token frequency of the adjectives in **Table 10** (with variation) is Mdn = 321. For comparison, the median of the token frequency of LAs without variation is Mdn = 1,325. A Shapiro-Wilk normality test (the *shapiro.test ()* function from the *stats* package in R) confirms that the data is not normally

---

[25] We need to entertain the possibility that the reason for the difference between the two grammars with respect to the importance of the R-sk schema is that base nouns in the native stratum exhibit different proportions of final Rs than base nouns in the foreign stratum (the baseline is different). This is not the case. The proportions are almost exactly the same: 0.698 in the native and 0.695 in the foreign stratum.

distributed. A non-parametric Mann-Whitney U test, available in R as the *wilcox.test()* function from the *stats* package, with *token frequency* as the dependent variable and *variation* (no/yes) as the categorical independent variable was performed. It confirms that adjectives exhibiting variation have lower values of token frequency than adjectives that do not show variation, W = 3702, p = .049. The Hodges-Lehmann estimate more precisely indicates that the difference between the medians of token frequency for the two groups (no variation/variation) is 881.

The finding that it is predominantly low-frequency adjectives (all of them derived from place names of foreign origin) that show variation is consistent with the predictions of the dual-route model of lexical access. Given that low-frequency words are more likely to be processed on the fly, competition between various ways of forming LAs is to be expected for such words.[26] However, as noted above, a more systematic approach would be necessary to confirm the hypothesis that lexical frequency plays an important role in the data. I leave such an analysis for future research. Here I point to a possible link between lexical frequency and variation (i.e. the lack of morphological stability).

Base	*-∅-sk-*	*TkF*	*-ij-sk-*	*TkF*	*-aɲ-sk-*	*TkF*
alask-a 'Alaska'	ala-sk-i	656	alask-ij-sk-i	17	alask-aɲ-sk-i	554
angɔl-a 'Angola'	angɔl-sk-i	805	angɔl-ij-sk-i	18	angɔl-aɲ-sk-i	202
bangladɛʂ 'Bangladesh'	bangladɛ-sk-i	235	bangl-ij-sk-i	25		
dʑibuti 'Djibouti'	dʑibu-tsk-i	3	dʑibut-ɨj-sk-i	13	dʑibut-aɲ-sk-i	6
gan-a 'Ghana'	gaɲ-sk-i	303	gaɲ-ij-sk-i	18		
irak 'Iraq'	ira-tsk-i	31,048	irak-ij-sk-i	487		
mɔzambik 'Mozambi-que'	mɔzambi-tsk-i	345	mɔzamb-ij-sk-i	80		
ɔsak-a 'Osaka'			ɔsak-ij-sk-i	7	ɔsak-aɲ-ski	6
tanzaɲj-a 'Tanzania'	tanzaɲ-sk-i	681	tanzaɲ-ij-sk-i	7		
tɔng-ɔ 'Tongo'			tɔng-ij-sk-i	68	tɔng-aɲ-sk-i	11

Table 10: Vacillating LAs.

---

[26] A similar tradeoff has been reported in Morgan & Levy's (2016) study of binomial expressions, which found that the processing of novel expressions relies on abstract knowledge, while reliance on stored multi-word representations increases with increased exposure to an expression.

7 Implications

The above corpus-based analysis has provided evidence for the relevance of a product-oriented schema: the presence of a final sonorant defines the preferred morphological structure of LAs, repeated in (39). An intermorph is more likely to be selected when the base does not end in a sonorant. Conversely, when the base ends in a sonorant, a LA is more likely to surface without an intermorph. One of the findings of this study is that the product-oriented schema plays a role in both the native and the foreign stratum.

(39)   R-sk-$_{LA}$

The observed selectional restrictions identified in the corpus-based analysis depend on the input and are to some extent phonologically arbitrary in the sense that they are not fully predictable from phonetic properties. For example, voiceless non-labial fricatives (S) show an affinity for the bare suffix, while most other consonants prefer the intermorphs. Similarly, /g/ and /sk/ in both strata show a preference for the bare suffix. What these generalizations share is that the conditioning context is absent from the output (due to deletion, coalescence or multiple correspondence) but present in the source, as shown in (40a–c). The application of such patterns justifies the need for source-oriented generalizations (contra Bybee 2001), as the distribution of suffixes in the output depends on the segmental structure of the input and cannot be insightfully determined on the basis of the output alone. For comparison, the generalization in (40d) can be expressed either with a source- or product-oriented schema, as the context is present in the output.

(40)   a. S→ -sk-   nɨs-a 'Nysa' > nɨ-sk-i   S : ∅
         b. g → -sk-   xag-a 'the Hague' > xa-sk-i   g : ∅
         c. sk → -sk-   alask-a 'Alaska' > ala-sk-i   sk : ∅
         d. b → -aɲ-sk- *or* b-aɲ-sk-   tɛb-ɨ 'Thebes' > tɛb-aɲ-sk-i   b : b

Moreover, the outputs are probabilistic. For example, extrasyllabic consonants are avoided using vowel insertion, an intermorph or segment deletion. Further, as exemplified in (41), different intermorphs can be used to avoid one and the same extrasyllabic consonant. The attested variation in the output in the same phonological context calls for frameworks that allow for probabilistic weighting or ranking of constraints. Crucially, the proposed analysis using ranked or weighted schemas is suitable to derive such generalizations, as it assumes that the output can be conditioned by the context in both the output and the input and can be non-uniform.

(41)   Cn → -ɛɲ-sk-, Cn → -aɲ-sk-
         Cr → -aɲ-sk-

The selection of an intermorph to a large extent depends on whether a LA belongs to the native or foreign lexical stratum. The intermorphs are significantly more likely to be selected when a word is of foreign than of native origin. The grammar of LAs is composed of two subgrammars, G_{Native}

and $G_{Foreign}$, with different phonological properties. The intermorphs have become a marker of foreign status of LAs. This is especially true for the intermorph *-ij-*, which is practically not used in the native stratum. Also, various intermorphs are used after non-extrasyllabic sonorants in the foreign stratum; in the native stratum the intermorphs are not used at all after such sonorants. Such categorical differences between the two strata indicate that the two subgrammars differ in weighting the relevant constraints.

Alternatively, source-oriented generalizations can be formalized without using source-oriented schemas. Becker & Gouskova (2016) analyze vowel ~ zero alternations in Russian and propose a model where a source-oriented generalization is captured by a phonotactic grammar that functions as a grammar inference mechanism, assigning words to the grammar that is best suited to derive them. To this aim, Becker & Gouskova (2016) propose that in addition to grammars proper, there are phonotactic gatekeeper grammars. Each new word must be fed to a gatekeeper grammar, which determines which grammar proper it should be sent to. This inference mechanism allows the speaker to extend source-oriented generalizations from the real words of the language to novel words. In the case at hand, each way of forming Polish LAs (*-sk-, -tsk-, -ąsk-, -ęsk-, -ijsk-*) would have both its grammar proper and a phonotactic gatekeeper grammar. Moreover, each of these construction-specific grammars would have distinct analogs in the native and foreign sublexicon. The key claim shared by Becker & Gouskova's (2016) analysis and the present analysis is that speakers encode product- and source-oriented generalizations.

8 Conclusion

This corpus-based quantitative analysis has provided evidence for a product-oriented schema which defines a preferred segmental structure of locative adjectives and, in this way, regulates the selection of intermorphs and final suffixes. Locative adjectives referring to places of both foreign and native origin are more likely to choose an intermorph when the base does not end in a sonorant. The existence of generalizations that crucially refer to the input justifies the need for source-oriented schemas, which function alongside product-oriented schemas. Certain base-final consonants or groups of consonants show an idiosyncratic affinity for particular affixes. The context that determines affix preference cannot be deduced from the output. Allomorph distribution also depends on markedness principles such as sonority sequencing, similarity avoidance, phonotactic restrictions and base-derivative identity relations. The identified selectional restrictions show varying degrees of arbitrariness, which supports the view that morphological patterns are learnable even if they are governed by phonologically arbitrary generalizations. A schema-based approach is well-suited to derive patterns that are phonologically arbitrary and dependent on the structures present in the input but absent from the output.

The gradiently different behavior of foreign and native words with respect to the propensity to use the intermorphs points to the existence of two subgrammars: $G_{Foreign}$ and G_{Native}. An intermorph is more likely to be selected in $G_{Foreign}$ than in G_{Native}. Thus, foreign and native locative adjectives differ in their preferred morphological composition.

This account has relied on statistical analyses of data extracted from a corpus and in this way has demonstrated the importance of quantitative data in measuring the gradient effects of a preferred segmental composition, source-oriented generalizations, markedness principles, and base-affix identity pressures. It complements analyses that use data from dictionaries to study similar phenomena. The gradience of the identified patterns implies that linguistic knowledge is not categorical but relies on statistical distributions. An important finding of corpus-based analyses, including the current one, is that exceptionless generalizations are less common than previously assumed in studies that relied mainly on the intuitions of idealized native speakers.

Supplementary file

Appendices. Results of the statistical analyses. Tableaux for the NHG simulations. DOI: https://doi.org/10.16995/glossa.10177.s1

Acknowledgements

I would like to thank two anonymous reviewers and an associate editor for helpful remarks that significantly improved this paper.

Competing interests

The author has no competing interests to declare.

References

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language* 78. 684–709. DOI: https://doi.org/10.1353/lan.2003.0002

Albright, Adam & Hayes, Bruce. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90. 119–161. DOI: https://doi.org/10.1016/S0010-0277(03)00146-X

Alegre, Maria & Gordon, Peter. 1999. Rule-based versus associative processes in derivational morphology. *Brain and Language* 68. 347–354. DOI: https://doi.org/10.1006/brln.1999.2066

Baayen, R. Harald. 1993. On frequency, transparency and productivity. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer. DOI: https://doi.org/10.1007/978-94-017-3710-4_7

Baayen, R. Harald & Lieber, Rochelle. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29. 801–843. DOI: https://doi.org/10.1515/ling.1991.29.5.801

Baayen, R. Harald & McQueen, James M. & Dijkstra, Ton & Schreuder, Robert. 2003. Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In Baayen, Harald R. & Schreuder, Robert (eds.), *Morphological structure in language processing*, 355–390. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110910186.355

Becker, Michael & Gouskova, Maria. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry* 47(3). 391–425. DOI: https://doi.org/10.1162/LING_a_00217

Bethin, Christina Y. 1998. *Slavic prosody: Language change and phonological theory*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511519765

Boersma, Paul. 1998. *Functional Phonology: Formalizing the interaction between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, Paul & Pater, Joe. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In McCarthy, John & Pater, Joe (eds.), *Harmonic Serialism and Harmonic Grammar*, 389–434. Sheffield: Equinox.

Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Booij, Geert & Audring, Jenny. 2017. Construction morphology and the parallel architecture of grammar. *Cognitive Science* 41(S2). 277–302. DOI: https://doi.org/10.1111/cogs.12323

Burzio, Luigi. 2002. Surface-to-surface morphology: When your representations turn into constraints. In Boucher, Paul (ed.), *Many morphologies*, 142–177. Somerville: Cascadilla Press.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form* (Vol. 9). Amsterdam: John Benjamins Publishing. DOI: https://doi.org/10.1075/tsl.9

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455. DOI: https://doi.org/10.1080/01690969508407111

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511612886

Carstairs, Andrew. 1988. Some implications of phonologically conditioned suppletion. In Booij, Geert & van Marle, Jaap (eds.), *Yearbook of morphology* 1988, 68–94. Dordrecht: Foris.

Carstairs, Andrew. 1990. Phonologically conditioned suppletion. In Dressler, Wolfgang U. & Luschützky, Hans C. & Pfeiffer, Oskar E. & Rennison, John R. (eds.), *Contemporary morphology*, 17–23. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110874082.17

Chomsky, Noam & Halle, Morris. 1968. *The sound pattern of English*. New York: Harper & Row.

Czaplicki, Bartłomiej. 2003. *Syllabification in current phonological theories: Issues in the phonology of English and Ukrainian*. Warsaw: University of Warsaw dissertation.

Czaplicki, Bartłomiej. 2013a. Arbitrariness in grammar: Palatalization effects in Polish. *Lingua* 123. 31–57. DOI: https://doi.org/10.1016/j.lingua.2012.10.002

Czaplicki, Bartłomiej. 2013b. R-metathesis in English: An account based on perception and frequency of use. *Lingua* 137. 172–192. DOI: https://doi.org/10.1016/j.lingua.2013.09.008

Czaplicki, Bartłomiej. 2014a. *Lexicon based phonology: Arbitrariness in grammar*. Munich: Lincom Europa.

Czaplicki, Bartłomiej. 2014b. Frequency of use and expressive palatalization: Polish diminutives. In Cyran, Eugeniusz & Szpyra-Kozłowska, Jolanta (eds.), *Crossing phonetics-phonology lines*, 141–160. Newcastle upon Tyne: Cambridge Scholars Publishing.

Czaplicki, Bartłomiej. 2019. Measuring the phonological (un)naturalness of selected alternation patterns in Polish. *Language Sciences* 72. 160–187. DOI: https://doi.org/10.1016/j.langsci.2018.10.002

Czaplicki, Bartłomiej. 2020. Construction-specific phonology: Evidence from Polish vowel-zero alternations. In Jaskuła, Krzysztof (ed.), *Phonological and phonetic explorations,* 77–93. Lublin: Wydawnictwo KUL.

Czaplicki, Bartłomiej. 2021. The strength of morphophonological schemas: Consonant mutations in Polish. *Glossa: A Journal of General Linguistics* 6(1): 25. 1–34. DOI: https://doi.org/10.5334/gjgl.1255

Czaplicki, Bartłomiej. 2022. Construction-specific effects of phonological similarity avoidance. *Poznań Studies in Contemporary Linguistics* 58(2). 159–204. DOI: https://doi.org/10.1515/psicl-2022-0010

Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58. 931–951. DOI: https://doi.org/10.1016/j.jml.2007.11.005

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition* 24. 143–188. DOI: https://doi.org/10.1017/S0272263102002024

Embick, David. 2010. *Localism versus globalism in morphology and phonology*. (Linguistic Inquiry Monographs 60). Cambridge, MA: MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262014229.001.0001

Endersen, Anna. 2015. *Non-standard allomorphy in Russian prefixes: Corpus, experimental, and statistical exploration*. UiT The Arctic University of Norway dissertation.

Fabb, Nigel. 1988. English Suffixation is Constrained only by Selectional Restrictions. *Natural Language and Linguistic Theory* 6. 527–539. DOI: https://doi.org/10.1007/BF00134491

Goldsmith, John. 1976. *Autosegmental Phonology*. Cambridge, MA: MIT dissertation.

Gussmann, Edmund. 2007. *The phonology of Polish*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780199267477.001.0001

Hamann, Silke. 2002. Postalveolar fricatives in Slavic languages as retroflexes. In Baauw, Sergio & Huiskes, Mike & Schoorlemmer, Maaike (eds.), *OTS Yearbook 2002,* 105–127. Utrecht: Utrecht Institute of Linguistics.

Harris, John. 1994. *English Sound Structure*. Oxford: Basil Blackwell.

Haspelmath, Martin & Sims, Andrea D. 2010. *Understanding morphology* (Understanding Language Series), 2nd edn. London: Hodder Education.

Hay, Jennifer. 2003. *Causes and consequences of word structure*. London: Routledge. DOI: https://doi.org/10.4324/9780203495131

Hayes, Bruce & Siptár, Péter & Zuraw, Kie & Londe, Zsuzsa. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). 822–863. DOI: https://doi.org/10.1353/lan.0.0169

Hayes, Bruce & Tesar, Bruce & Zuraw, Kie. 2013. "OTSoft 2.5" software package, http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Hume, Elizabeth & Clements, George N. 1995. The Internal Organization of Speech Sounds. In Goldsmith, John (ed.), *Handbook of phonological theory*, 245–306. Oxford: Basil Blackwell.

Inkelas, Sharon. 1999. Exceptional stress-attracting suffixes in Turkish: representations vs. the grammar. Presented at the Workshop on Prosodic Morphology, Utrecht University, 1994. DOI: https://doi.org/10.1017/CBO9780511627729.006

Inkelas, Sharon. 2014. *The interplay of morphology and phonology*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199280476.001.0001

Itô, Junko & Mester, Armin. 1995. The core-periphery structure of the lexicon and constraints on reranking. In Beckman, Jill & Urbanczyk, Suzanne & Walsh, Laura (eds.), *University of Massachusetts Occasional Papers in Linguistics Vol. 18: Papers in Optimality Theory,* 181–209. University of Massachusetts, Amherst: GLSA.

Itô, Junko & Mester, Armin. 1999. The phonological lexicon. In Tsujimura, Natsuko (ed.), *The handbook of Japanese linguistics*, 62–100. Oxford: Blackwell. DOI: https://doi.org/10.1002/9781405166225.ch3

Jakobson, Roman & Fant, Gunar M. & Halle, Morris. 1952. *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates.* Cambridge: MIT Press.

Jurgec, Peter. 2016. Velar palatalization in Slovenian: Local and long-distance interactions in a derived environment effect. *Glossa: A Journal of General Linguistics* 1(1). 24. DOI: https://doi.org/10.5334/gjgl.129

Kager, René. 1999. *Optimality Theory*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511812408

Kallas, Krystyna. 1999. Przymiotnik. In Grzegorczykowa, Renata & Laskowski, Roman & Wróbel, Henryk (eds.), *Gramatyka współczesnego języka polskiego: Morfologia.* Warsaw: Wydawnictwo Naukowe PWN.

Kapatsinski, Vsevolod. 2010. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1(2). 361–393. DOI: https://doi.org/10.1515/labphon.2010.019

Kapatsinski, Vsevolod. 2013. Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* 89(1). 110–148. DOI: https://doi.org/10.1353/lan.2013.0003

Kapatsinski, Vsevolod. 2021. What are constructions, and what else is out there? An associationist perspective. *Frontiers in Communication* 5. 134. DOI: https://doi.org/10.3389/fcomm.2020.575242

Kenstowicz, Michael. 1996. Base identity and uniform exponence: Alternatives to cyclicity. In Durand, Jacques & Laks, Bernard (eds.), *Current trends in phonology: Models and methods* 1. 363–394. Salford: University of Salford.

Kowalik, Krystyna. 1997. *Struktura morfonologiczna współczesnej polszczyzny*. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Kreja, Bogusław. 1989. *Z morfonologii i morfonotaktyki współczesnej polszczyzny*. Wrocław: Zakład Narodowy im. Ossolinskich.

Legendre, Geraldine & Miyata, Yoshiro & Smolensky, Paul. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Lightner, Theodore. 1965. *Segmental phonology of modern standard Russian*. MIT dissertation.

MacWhinney, Brian. 1978. *The acquisition of morphophonology. Monographs of the society for research in child development* 43. 1–122. DOI: https://doi.org/10.2307/1166047

Mańczak, Witold. 1980. Laws of analogy. In Fisiak, Jacek (ed.), *Historical morphology*, 283–288. The Hague: Mouton. DOI: https://doi.org/10.1515/9783110823127.283

McCarthy, John & Prince, Alan. 1995. Faithfulness and reduplicative identity. In Beckman, Jill & Dickey, Laura & Urbanczyk, Suzanne (eds.), *University of Massachusetts occasional papers in linguistics 18: Papers in Optimality Theory,* 249–384. Amherst, MA: GLSA.

McQueen, James & Cutler, Ann. 1998. Morphology in word recognition. In Spencer, Andrew & Zwicky, Arnold A. (eds.), *The handbook of morphology*, 406–427. Oxford: Blackwell. DOI: https://doi.org/10.1002/9781405166348.ch21

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780199207916.001.0001

Morgan, Emily & Levy, Roger. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition* 157. 384–402. DOI: https://doi.org/10.1016/j.cognition.2016.09.011

Nedjalkov, Igor. 1997. *Evenki*. New York: Routledge.

Nevins, Andrew. 2011. Phonologically Conditioned Allomorph Selection. In van Oostendorp, Marc & Ewen, Colin J. & Hume, Elizabeth & Rice, Keren (eds.), *The Blackwell companion to phonology,* 2357–2382. Malden: Wiley-Blackwell. DOI: https://doi.org/10.1002/9781444335262.wbctp0099

Paster, Mary. 2006. *Phonological conditions on affixation*. University of California at Berkeley dissertation.

Pater, Joe. 2010. Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Parker, Steve (ed.), *Phonological argumentation: Essays on evidence and motivation,* 123–154. London: Equinox.

Pierrehumbert, Janet. 2006. The statistical basis of an unnatural alternation. In Goldstein, Louis & Whalen, Douglas H. & Best, Catherine (eds.), *Laboratory Phonology 8: Varieties of phonological competence*, 81–107. Berlin: Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110197211.1.81

Plag, Ingo. 1996. Selectional Restrictions in English Suffixation Revisited. A Reply to Fabb (1988). *Linguistics* 34. 769–798. DOI: https://doi.org/10.1515/ling.1996.34.4.769

Plag, Ingo. 2012. *Word-formation in English*. Cambridge: Cambridge University Press.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Satkiewicz, Halina. 1969. *Produktywne typy słowotwórcze współczesnego języka ogólnopolskiego*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Selkirk, Elisabeth. 1982. *The Syntax of Words*. Cambridge, MA: MIT Press.

Short, David. 1993. Czech. In Comrie, Bernard & Corbett, Greville G. (eds.), *The Slavonic languages*, 455–532. New York: Routledge.

Stemberger, Joseph P. & MacWhinney, Brian. 1988. Are inflected forms stored in the lexicon? In Hammond, Michael & Noonan, Michael (eds.), *Theoretical morphology: Approaches in modern linguistics*, 101–116. San Diego, CA: Academic Press. DOI: https://doi.org/10.1163/9789004454101_009

Steriade, Donca. 2000. Paradigm uniformity and the phonetics-phonology boundary. In Broe, Michael & Pierrehumbert, Janet (eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*, 313–334. Cambridge: Cambridge University Press.

Szymanek, Bogdan. 2010. *A panorama of Polish word-formation*. Lublin: Wydawnictwo KUL.

Venables, William N. & Ripley, Brian D. 2002. *Modern Applied Statistics with S*, Fourth edition. New York: Springer. DOI: https://doi.org/10.1007/978-0-387-21706-2

Wierzchowska, Bożena. 1960. Z badań eksperymentalnych polskich głosek nosowych. *Biuletyn fonograficzny* 3. 67–87.

Zuraw, Kie. 2010. Model of lexical variation and the grammar: With application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28. 417–472. DOI: https://doi.org/10.1007/s11049-010-9095-z

Zuraw, Kie & Hayes, Bruce. 2017. Intersecting constraint families: An argument for Harmonic Grammar. *Language* 93. 497–548. DOI: https://doi.org/10.1353/lan.2017.0035