An OT analysis of Syllable Structure in Punjabi

Phonology II
Auromita
H00MAENG20190259
.

1 Punjabi

Punjabi (Western, ISO-639-3 pnb) is an Indo-Aryan language (Indo-European, Indo-Iranian) spoken in Pakistan and India, and in immigrant communities in the UK, Canada, USA, and elsewhere. The lexicon of Punjabi includes loanwords from Arabic, English, Hindi-Urdu, Persian, Sanskrit, Turkish, and other contact languages. The varieties of Punjabi are broadly classified into two groups: Eastern and Western. The Eastern dialects are primarily spoken in the Indian state of Punjab, whereas the Western dialects cover the area of Punjab, Pakistan.

Punjabi is a syllable-timed language. The phonemic inventory consists of thirty two consonants, ten non-nasal vowels—seven peripheral and three central, and seven nasalized vowels (the peripheral vowels have contrastive nasalized counterparts). Vowels in Punjabi are categorized as central vs peripheral, the former being analyzed as monomoraic and the latter as bimoraic. Geminate consonants are always preceded by central vowels (Hussain et al., 2020). Since central vowels are monomoraic, this means that there is a restriction on syllable weight—- a syllable cannot have more than three moras. Thus, the language allows light (V), heavy (VC), and superheavy (VVC or CVV) syllables.

2 Data sources

Most of the data discussed here is from Lyallpuri, a Western variety of Punjabi spoken largely in the urban areas of Faisalabad (formerly Lyallpur), Pakistan, from Hussain et al. (2020). From the available studies on Eastern varieties (largely spoken in India) referenced here, syllable structure does not seem to differ across these varieties. I have also considered data from corpus-based studies of Punjabi Phonology (Singh & Lehal, 2010, 2012) and loanword phonology (Habib & Khan, 2019; Mahmood et al., 2011).

3 Analysis

3.1 Onset and Coda?

3.1.1 Onset

All consonants that appear before the nucleus of a syllable form the ‘Onset’. Typological data shows that among the world’s languages, there are those in which syllables obligatorily have an

onset (e.g. Arabic, Temiar), and those in which syllables may appear without an onset (e.g. Bangla, Japanese). However, there is no attested language in which an onset is *disallowed*. That is, if a language has onset-less syllables, then it must also have syllables with onsets. This means that cross-linguistically, the presence of an onset is less marked than the absence of an onset. This is phonetically grounded, as a vowel preceded by a consonant is acoustically more informative than a syllable-initial vowel. Thus, this tendency can be formalized as a well-formedness constraint:

(1) ONSET: Syllables must have onsets

Punjabi falls into the second category discussed above– it has both syllables with onsets and syllables without onsets. This means that in Punjabi the constraint ONSET cannot be undominated. Hypothetically, if ONSET was high-ranked in a language, then one possible repair strategy to 'fix' an imperfect syllable (say V) would be to epenthesize a consonant before the nucleus, i.e. in the onset position (to give the structure CV). This is indeed seen in languages like Axininca Campa. This strategy violates the faithfulness constraint DEP-IO:

(2) DEP-IO: Output segments must have input correspondents (no epenthesis)

Another possible repair strategy, in case of a multisyllabic structure (say V.CV) would be to delete the onset-less vowel (to give the structure CV). This would violate the faithfulness constraint MAX-IO:

(3) MAX-IO: Input segments must have output correspondents (no deletion)

Since Punjabi does not use either of these, it means that in Punjabi, DEP-IO and MAX-IO are both higher-ranked than ONSET. The following tableaux demonstrate this.

Optimal output for an input form that has an onset (/k^ha/ 'eat'):

/k^ha/	DEP-IO	MAX-IO	ONSET
☞ a. k^ha			
b. a		*!	*
c. Ck^ha	*!		

Optimal output for an input form without an onset (/əsi/ 'we') ('C' stands for any unmarked consonant):

/ə.si/	DEP-IO	MAX-IO	ONSET
☞ a. ə.si			*
b. Cə.si	*!		
c. si		*!	

These examples show that in Punjabi, the presence/absence of the onset in a syllable is dictated by the input, as the relevant faithfulness constraints are higher ranked than the corresponding markedness constraints.

3.1.2 Coda

All consonants that appear after the nucleus of a syllable form the 'Coda'. Typological data shows that the cross-linguistic pattern for codas is opposite to the pattern seen in onsets– there are languages in which syllables categorically disallow a coda (e.g. Fijian, Mazateco), and those in which syllables may appear with a coda (e.g. Arabic, English). However, there is no attested language in which a coda is *mandatory*. That is, if a language has syllables with codas, then it must also have syllables without codas. But the reverse does not hold. This means that cross-linguistically, the presence of a coda is more marked than the absence of a coda. This tendency can be formalized as a well-formedness constraint:

(4) NO-CODA: Syllables must not have a coda (syllables are open)

Hypothetically, if NO-CODA was undominated in a language, there could two potential repair strategies to 'fix' an imperfect syllable of the form CVC:

i. delete the coda consonant (giving CV). This would violate MAX-IO

ii. epenthesize a vowel after the coda consonant, so that it is syllabyfied a the onset of the following syllable (giving CV.CV). This would violate DEP-IO

Punjabi does not use either of these options, suggesting that both the faithfulness constraints are ranked above NO-CODA. The following tableaux demonstrate this.

Optimal output for an input form that has a coda (/am/ 'ordinary'):

/am/	DEP-IO	MAX-IO	*CODA
☞ a. am			*
b. a	*!		
c. a.mʊ		*!	

.

For an input form that lacks a coda (/k^ha/ 'eat'). V is any unmarked vowel:

/k^ha/	DEP-IO	MAX-IO	*CODA
☞ a. k^ha			

.

This form does not violate any of the markedness constraints considered here, and making any changes will violate additional faithfulness constraints without improving the syllable structure. This means that with respect to these four constraints, any other candidates will be intrinsically suboptimal. Thus these have not been shown here. This is an example of the principle of **Economy** in Optimality Theory: banned options are available only to avoid violations of higher-ranked constraints.

From the tableaux above, we see that the presence/absence of a coda is faithful to the input– an input form having a coda will surface with a coda, and an input without a coda will surface as such. At first this seems to suggest that the constraints ONSET and NO-CODA do not have any effect of syllable structure. However, this is not true. These markedness constraints will play a role in determining the optimal syllabification of an input form. Consider the disyllabic word /udas/ 'sad':

/udas/	ONSET	*CODA
☞ a. u.das		*
b. ud.as	*	*!*

This shows that even though the markedness constraints are low-ranked in Punjabi, they are still 'active'– they play a role in determining syllable structure. This is another important feature of an OT-analysis: lower-ranked constraints are still active, and play a role in selecting between candidates that are identical with respect to higher-ranked constraints.

To conclude the section, the constraint hierarchy developed so far for syllable structure in Punjabi is shown below:

MAX-IO, DEP-IO > ONSET, NO-CODA

3.2 Complex onset and coda?

So far, our analysis has been concerned with just the presence or absence of onset and coda consonants in the syllable. However, syllables can also differ in the nature of their onset and coda, and cross-linguistically, the phonologies of most languages make reference to these features of the syllable. Specifically, onsets and codas can be either *simple* (made up of a single consonant) or *complex* (containing more than one consonant). Typological data from a large number of languages of the world highlight the following facts:

i. the presence of syllables having complex onsets necessarily entails that the language also has syllables with simple onsets, but not vice-versa.

ii. there are languages where onsets are mandatorily simple and complex onsets are completely disallowed (e.g. Japanese, Finnish), but no attested language in which it is mandatory to have complex onsets.

iii. the presence of syllables having complex codas necessarily entails that the language also has syllables with simple codas, but not vice-versa.

iv. there are languages where the coda is mandatorily simple and complex codas are completely disallowed (e.g. Spanish, Japanese), but no attested language in which it is mandatory to have complex codas.

Together, these facts suggest that complex syllable boundaries are universally marked compared to simple boundaries. This can be translated into two markedness constraints on syllable structure:

(5) *COMPLEX (ONS): Onsets are simple

(6) *COMPLEX(CODA): Codas are simple

There is evidence to suggest that these should be treated as separate constraints, rather than an overall ban on complex boundaries: there are languages which allow complex onsets but not complex codas (e.g. Spanish), languages which allow complex codas but not complex onsets (e.g. Finnish), as well as languages which disallow both (e.g. Japanese). Thus in languages like Spanish, *COMPLEX(CODA) must be undominated and *COMPLEX (ONS) must be low-ranked, whereas the opposite holds for languages like Finnish. In languages like Japanese, both markedness constraints must be undominated. The facts about syllable complexity in Punjabi are discussed below:

3.2.1 Complex Onset

Through a large-scale corpus based analysis, Singh & Lehal (2010) identify seven types of syllables in Punjabi – V, VC, CV, VCC, CVC, CVCC and CCVC. These were extracted from a speech database created from a carefully selected balanced corpus having 104425741 total and 232565 unique words. Thus, this database represents the distribution of syllable structures in actual language use. The authors analyze how frequently each of these syllable types are attested in words of the language. They report that CCVC structures are extremely rare, having a frequency approaching 0. (The rarity of this structure was also noted previously by Singh (2002) in his reference grammar for Punjabi). Hussain et al. (2020), in their study of the Lyllpuri variety of Punjabi, do not attest to word-initial consonant clusters at all. Together, these point to the fact that complex onsets are disfavored in Punjabi phonology.

Another piece of evidence for this comes from the patterns of loanword adaptation. Mahmood et al. (2011) used recordings of collected from Punjabi stage dramas, English movies dubbed into Punjabi, Punjabi news channels in Pakistan to create a 19:14:26 hour speech corpus. From this spoken corpus, 292 English words were selected and analyzed to find general adaptation strategies. At the syllable level, the most persistent pattern of adaptation was the simplifying of onset clusters through vowel epenthesis. Habib & Khan (2019) reports a similar pattern of cluster simpification. Some examples:

(7) a. /skul/ –> /səkul/ ‘school’
b. /træk/ –> /təræk/ ‘track’
c. /plæn/ –> /pəlæn/ ‘plan’

Translating this into OT constraints, this means that in Punjabi, the markedness constraint *COMPLEX (ONS) is high-ranked and undominated. There are two potential repair strategies to ‘fix’ a syllable with a complex onset (say CCV):

i. epenthesize a vowel, to give CV.CV (position of epenthesized vowel is discussed later). This violates the faithfulness constraint DEP-IO

ii. delete one of the consonants in the cluster, to give CV. This violates MAX-IO

We find that Punjabi always makes use of strategy (i), but not (ii). This means that in Punjabi, MAX-IO mut be higher ranked than DEP-IO. This constraint heirarchy is demonstrated in the following tableau:

/skul/	*CC(ONS)	MAX-IO	DEP-IO
a. skul	*!		
☞ b. səkul			*
c. kul		*!	
d. sul		*!	

.

Thus, the hierarchy of constraints for onset complexity so far: *COMPLEX(ONSET), MAX-IO > DEP-IO.

This accounts for the observation that epenthesis is preferred over deletion. However, there are some question still unanswered:

Which vowel is epenthesized? An epenthesized vowel has no correspondent in the input. Therefore, faithfulness constraints do not apply to it. In this situation, the OT principle of minimal violation predicts that the chosen segment will be such that it violates the minimum number of markedness constraints. As a result, it is generally found that the epenthesized segment is either unmarked (minimal violation of context-free markedness constraints), or contextually colored (minimal violation of context-sensitive markedness constraints). Habib & Khan (2019) and Mahmood et al. (2011) both note that the most common epenthetic vowel in Punjabi is /ə/, which is an unmarked segment ([-high], [-low], [-back], [-round], [-nasal]). In certain phonological environments, /e/ is used (alernation discussed by Habib & Khan (2019)).

Where is the epenthesized vowel placed? In the constraint hierarchy shown above, the the optimal candidate contains an epenthesized vowel between the two consonants of the complex onset. However, there is another potential candidate that respects *COMPLEX(ONS), and could thus be 'optimal' in the tableau: *[əs.kul]. However, this form is not attested in Punjabi. What prevents this from emerging as an optimal candidate? We can account for this once we add these constraints into our existing hierarchy:

*COMPLEX(ONS), MAX-IO > DEP-IO > ONSET, *CODA

The following tableau demonstrates this:

/skul/	*CC(ONS)	MAX-IO	DEP-IO	ONSET	*CODA
a. skul	∗!				∗
☞ b. sə.kul			∗		∗
c. kul		∗!			∗
d. sul		∗!			∗
e. əs.kul			∗	∗!	∗∗

.

We see that the unattested candidate *[əs.kul] is ruled out because it violates additional markedness constraints: it creates a syllable without an onset and having a coda. Once again, we see that the low-ranked markedness constraints are active, and interact with phonological processes.

3.2.2 Complex Coda

While Punjabi unambiguously disallows complex onsets, the case of complex codas is slightly more...complex. Punjabi does not fall clearly into any of the categories of languages described in sec. 3.2. In their description of Lyllpuri Punjabi, Hussain et al. (2020) note that CVCC syllables do not appear to exist in the native Punjabi lexicon, but are largely found only in Arabic, Persian, or Urdu loanwords (e.g. /sust/ 'lazy', /gəʃt/ 'meat'). However, they also point out that these loanwords have largely been integrated into the phonology of the language. For all purposes, native speakers of Punjabi today identify them as Punjabi words. Therefore, whether modern Punjabi 'disallows' complex codas seems to depend on a researcher's theoretical position on the status of these words. Further evidence for this gradience comes from the fact that:

i The realization of final consonant clusters by Punjabi speakers depends not only on speech rate, but also speaker literacy and bilingualism (Gill & Gleason, 1969)

ii In a corpus-based analysis of Punjabi words (Singh & Lehal, 2010), syllables with final consonant clusters (VCC and CVCC) were present, but among the least frequent syllable types.

Thus, it could be said that Punjabi generally *disfavors* complex codas. But in cases where the input contains a complex coda, it surfaces in the output (e.g. /sust/). In terms of constraints, this means that the markedness constraint *COMPLEX(CODA) cannot be undominated. It must be dominated by faithfulness constraints that prevent the loss of the input coda cluster. Hypothetically, if *COMPLEX(CODA) was high-ranked, two potential repair strategies could be used by the language to 'fix' an imperfect syllable of the form CVCC:

i delete one of the coda consonants to get the structure CVC. This would violate MAX-IO

ii epenthesize a vowel to break the cluster, giving the structure CV.CVC or CVC.CV. This would violate DEP-IO

Since Punjabi does not appear to employ any of these strategies, we can infer that both the faithfulness constraints MAX-IO and DEP-IO are higher ranked than *COMPLEX(CODA). This

accounts for the examples discussed above, as shown in the following tableau (V is any unmarked vowel):

/sust/	DEP-IO	MAX-IO	*CC(CODA)
☞ a. sust			*
b. su.sVt	*!		
c. sus.tV	*!		
d. sut		*!	
e. sus		*!	

.

Adding this to the existing constraint hierarchy:
*COMPLEX(ONS), MAX-IO > DEP-IO > ONSET, *CODA, *COMPLEX(CODA)
While these constraints account for the pattern seen in Punjabi words, the phonological adaptation of unambiguous loanwords (words that are not considered to be part of the Punjabi lexicon) does not follow this pattern. For example, when more recent borrowings from English are adapted into Punjabi phonology, the language's tendency to disfavor complex codas takes precedence, and the input cluster does not surface. The following examples are noted in (Mahmood et al., 2011):

(8) a. /film/ –> /filəm/ 'film'
b. /bʌtn/ –> /bʌtən/ 'button'
c. /dʌbl/ –> /dʌbəl/ 'double'
d. /levl/ –> /levəl/ 'level'

In (i), the coda cluster is broken up by the epenthesis of [ə]. In (ii), (iii) and (iii), the native English pronunciation contains a syllabic consonant in the second syllable. Since Punjabi does not allow syllabic consonants, these are adapted into the corresponding [-syllabic] segments [n] and [l]. This gives a consonant cluster in the coda, which is then broken up by [ə]. These examples suggest that there is a more specific markedness constraint that disallows complex codas in loanwords. This seems to be reasonable, as many languages show evidence of phonological constraints that access lexical/morphological information. Proposing the following markedness constraint:

*COMPLEX(CODA)-LW: No complex coda in loanwords

This has to be a high-ranked constraint. As the examples above show, the repair strategy that is used to avoid a violation of this constraint is epenthesis, which violates DEP-IO. Another possible repair strategy would be to delete one of the coda consonants (e.g. to give /levl/ –> */lev/), which would violate MAX-IO. However, this is not attested. This means that the new constraint *COMPLEX(CODA)-LW must be ranked higher than DEP-IO but lower than MAX-IO. The following tableau shows the adaptation of the input form /film/ 'film':

/film/	MAX-IO	*CC(CODA)-LW	DEP-IO
a. film		∗!	
b. fil	∗!		
c. fim	∗!		
☞ d. filəm			∗

.

Adding this to the hierarchy developed so far:
*COMPLEX(ONS), MAX-IO > *COMPLEX(CODA)-LW > DEP-IO > ONSET, *CODA, *COMPLEX(CODA)

This hierarchy accounts for the onset and coda complexity in Punjabi. However, there is still one unaddressed question: in the tableau above, we considered a candidate in which epenthetic segment is inserted between the two consonants of the coda cluster. However, this is not the only possible pattern for epenthesis. It is possible to insert the segment after the cluster, giving the candidate /fil.mə/. This also avoids the *COMPLEX(CODA)-LW violation, and does not appear to violate any additional constraints in the present hierarchy. Yet it is not attested in the language. What prevents this form from being the optimal candidate? This is discussed in the next section.

3.3 Alignment

In the previous section, we noted a problem: out of two possible output candidates for the input /film/, only one (/fi.ləm/) is attested, even though the other (/fil.mə/) appears to be equally good in all respects. A similar problem with the position of the epenthetic segment was noted in 3.2.1, where /sə.kul/ and /əs.kul/ were the competing candidates for the input /skul/. We proposed that /əs.kul/ is rejected because it violates additional markedness constraints, namely *CODA and ONSET. The relevant tableau is reproduced below:

/skul/	*CC(ONS)	MAX-IO	DEP-IO	ONSET	*CODA
a. skul	∗!				∗
☞ b. sə.kul			∗		∗
c. əs.kul			∗	∗!	∗∗

.

However, in the case of coda epenthesis, the two candidates do not seem to differ in their violations of *CODA and ONSET:

/film/	MAX-IO	*CC(CODA)-LW	DEP-IO	ONSET	*CODA
a. film		∗!			∗
b. fi.ləm			∗		∗
c. fil.mə			∗		∗

.

This tableau is inconclusive. This suggests that there is some other difference between the candidates that is not captured by these constraints. Looking at the candidates, one difference is that of alignment: in /fi.ləm/, the right edge of the output corresponds to te right edge of the input, whereas in /fil.mə/, epenthesis disrupts the edge. The tendency for phonological processes to respect word edges is a cross-linguistically robust phenomenon. This is grounded in phonetics– since the edges of a prosodic word are maximally salient, preserving the edges is a good strategy to ensure that acoustic information is used optimally by the language to maintain lexical contrasts. Thus, we can propose a constraint for Punjabi as follows:

ALIGN-R: The right edge of the Grammatical Word coincides with the right edge of the Prosodic Word

This can account for the selection of /fi.ləm/ over /fil.mə/. Since these two candidates do not differ in any other respect, the position of the constraint, is irrelevant– placing it anywhere will give the required result:

/film/	MAX-IO	*CC(CODA)-LW	DEP-IO	ONSET	*CODA	ALIGN-R
a. film		∗!			∗	
☞ b. fi.ləm			∗		∗	
c. fil.mə			∗		∗	∗!

.

In fact, it is possible to reanalyze the /sə.kul/ vs /əs.kul/ difference on the basis of this, by extending the notion of alignment to include the left edge. Proposing a general constraint on alignment that subsumes both edges:

ALIGN: The edge of the Grammatical Word coincides with the corresponding edge of the Prosodic Word

We can now say that /sə.kul/ is the optimal candidate over /əs.kul/ because it does not violate ALIGN. This would be a desirable explanation, as it would give a uniform analysis for epenthesis in the onset and the coda. For this, it must be ranked higher than ONSET and *CODA. Considering data from other phonological processes will narrow down its exact position in the constraint hierarchy.

3.4 Conclusion

This paper proposes an OT constraint hierarchy to account for the syllable structure of Punjabi. The final proposed hierarchy is shown below (the position o ALIGN is tentative):

*COMPLEX(ONS), MAX-IO > *COMPLEX(CODA)-LW > DEP-IO, ALIGN > ONSET, *CODA, *COMPLEX(CODA)

This accounts for the syllable structures observed in native Punjabi, as well as observed patterns of phonological adaptation in loanwords. It also gives some indication about the possible factors that explain the frequency distribution of the different syllable types seen in large-scale corpus studies.

References

Gill, H. S., & Gleason, H. A. (1969). *A reference grammar of punjabi*. Department of Linguistics, Punjabi University Patiala, India.

Habib, M. A., & Khan, A. A. (2019). Vowel epenthesis in loanword integration: A study of english consonant cluster at onset. *International Journal of English Linguistics*, *9*(4), 332–347.

Hussain, Q., Proctor, M., Harvey, M., & Demuth, K. (2020). Punjabi (lyallpuri variety). *Journal of the International Phonetic Association*, *50*(2), 282–297.

Mahmood, R., Hussain, Q., & Mahmood, A. (2011). Phonological adaptations of english words borrowed into punjabi. *European Journal of Social Sciences*, *22*(2), 234–245.

Singh, P. (2002).
In *Sidhantik bhasha vigeyan* (4th ed., p. 371–372). Madan Publications.

Singh, P., & Lehal, G. S. (2010). Corpus based statistical analysis of punjabi syllables for preparation of punjabi speech database. *International Journal of Intelligent Computing Research (IJICR)*, *1*(3).

Singh, P., & Lehal, G. S. (2012). Punjabi text-to-speech synthesis system. In *Proceedings of coling 2012: Demonstration papers* (pp. 409–416).