# ANALYSIS OF URDU SYLLABIFICATION USING MAXIMAL ONSET PRINCIPLE AND SONORITY SEQUENCE PRINCIPLE

*BILAL AKRAM*

## ABSTRACT

The paper compares the performance of MOP and SSP in forming Urdu syllables. Both the syllabification principles are applied on Urdu words and compared with the original syllabification of the words done by the native speakers. The patterns of the syllables are also extracted and it is observed that it depends upon the number of consonants in the consonant clusters; the phonotactic constraints of Urdu are also analyzed along with an epenthesis rule. In the end an algorithm is also given for the syllabification of Urdu words using these two syllabification principles.

## 1. INTRODUCTION

The Urdu language belongs to the family of the New Indo-Aryan languages that is a sub branch of Indo-European languages. Urdu is spoken by at least 50 million people in more than 10 countries (Hussain, 1997, p 39.).

Syllabification is a process that associates a linear string of segments with a syllable structure (Goldsmith, 1990, p 117). Now what is a syllable? From a descriptive point of view, words should be factorable into sequences called syllables, which should have a specifiable internal structure that is roughly constant across the language (Goldsmith, 1990, p 107). Urdu also has some specific syllable patterns or templates for its syllables.

An analysis has been done in which two major principles related to syllabification namely Maximal Onset Principle and Sonority Sequence Principle have been used. The main objective of this analysis is to determine which of the two principles can best form the syllables of Urdu, or if applied one after the other what should be there order of application. Then the syllables are analyzed to extract the “syllable templates” of Urdu language. Some times there are sequences of segments that are legal syllables with respect to these principles but not by the language, those sequences are called phonotactic constraints. These are the constraints that are only for the particular language under consideration. A list of the phonotactic constraints of Urdu has also been extracted after analysis.

## 2. LITERATURE REVIEW & PROBLEM STATEMENT

There are a number of principles related to syllabification. A few major ones are discussed below:

### 2.1 Maximal Onset Principal (MOP)

In Maximal Onset Principal the consonants are preferred in the onset and thus allowing no coda consonants except for the word final position (Goldsmith, 1990, p 128).

### 2.2 Sonority Sequence Principal (SSP)

Sounds not only differ in quality but also in *Sonority.* The sonority of a sound is determined primarily by the size of the resonance chamber through which the air stream flows or in other words it is the degree of the openness of the vocal tract apparatus during the production of sound (Goldsmith, 1990, p 110). The hierarchy of sonority is shown in figure 1 (Goldsmith, 1990, p 110) and it can also be observed by an individual that while producing a vowel the vocal tract is more open than while producing a consonant thus proving that vowels are more sonorous than consonants. According to SSP a syllable must have rising sonority till the nucleus and falling from there onwards (Goldsmith, 1990, p 110)

*Sonority Hierarchy*

- Vowels
  - Low
  - Mid
  - High
- Glides
- Liquids
- Nasals
- Obstruents
  - Fricatives
  - Affricates
  - Stops

**FIGURE1 Hierarchy of Sonority**

## 2.3 Maximal Coda Principal (MCP)

In Maximal Coda Principal the consonants are preferred in the coda and thus allowing no onset consonants except for the word initial position.

Other principles and proposals related to syllabification also exist, for example, template matching, all nuclei first, linear scanning approach etc. But in these principles and proposals prior knowledge of syllable templates is also required, which in our case were not known before the analysis.

Urdu is not a new language, but still not a lot of work has been done on the syllabification of Urdu. The first one is a book by Sohail Bukhari by the name of "Phonology of Urdu language". He has based his work on pure native words of Urdu, ignoring all the words Urdu has borrowed from other languages like Persian and Arabic. He suggests that a word is made up of at least two sounds a consonant and a long vowel, but no words begin with a long vowel nor with the consonant r, rh or η nor one ends in η. Short vowels □,□ cannot recur consecutively within a word nor can any one of them follow the middle consonant of three consonant syllable (Bokhari, 1985, p 17). The biggest Urdu word is trisyllabic hence complex words containing more than three syllables are compressed and sounds are assimilated to three syllables (Bokhari, 1985, p 19). He suggests the following possible templates for Urdu syllables as in *figure 2 (Bokhari, 1985, p 18).*

CVC
CVCC
CVVC
CVCVC
CVCVV
CVCVCC
CVCVVC

**FIGURE 2 Syllable templates by Bokhari**

Another work done is by Sarmad Hussain in his PhD thesis. He has talked about the structure of the syllable and to some extent the phonotactic constraints as well. He suggests that open syllables with short vowels do not occur at word final position. There can be complex codas and complex onsets, however there are some limitations on the formation of these complex codas and onsets. First of all the Sonority Sequence Principal should be satisfied. Secondly these complex codas and onsets can contain at most only two consonants. Where there are two consonants in the onset the second consonant is limited to glides or may be a /h/. When there are two consonants in the coda the first is consonant in the coda is limited to a voiceless fricative or nasal and the second consonant is limited to a stop. The alveolar flap cannot occur in the onset position. And he further goes on to add that there may be more restrictions on these onset and codas that need to be discovered (Hussain, 1997, p 41). He has also given some templates for the Urdu syllables shown in *figure 3*

*Simple Onset*

- *CV*
- *CVC*
- *CVCC*
- *CVV*
- *CVVC*
- *CVVCC*

*Complex Onset*

- *CCV*
- *CCVC*
- *CCVCC*
- *CCVV*
- *CCVVC*
- *CCVVCC*

**FIGURE 3 Syllable Templates by Sarmad**

## 3. METHODOLOGY

First of all a good data set of about five thousand words were gathered for analysis and transcribed in IPA. All the words used for the analysis were collected from dictionary (Qureshi, 1992,p 1-496). In this dictionary the syllable boundaries in words were also marked but some times native speaker's help was also taken to find out the syllabification of some of the words.

### 3.1. No Consonant Cluster

The analysis began with the words that have simple CVCV structure i.e. there were no consonant clusters. In these words the syllabification was very simple and easy.

### 3.2. With Consonant Clusters

The main problem of syllabification comes when there is a consonant cluster in the word. In this scenario the two principles are applied on those words one by one. Now there are two possibilities:

- The two principles produce the same syllabification

There are another two cases in it i.e. the syllabification is the correct syllabification for that word or not. If it's correct, than good enough. But if it not correct than look for some phonotactic constraints. Now again syllabify it keeping in view that phonotactic constraint and try to come up with the correct syllabification.

- The two principles produce different syllabification

In this case the results of both the principles were compared with the correct syllabification that was taken from the dictionary. The principle that resulted in correct syllabification was noted down to help arrive at a conclusion at the end of the entire analysis. But if none of the principles gave the correct syllabification than again we looked for phonotactic constraints but if there were none than its one of the very few exceptions. Usually when the two principles produced different results, one of the results was correct.

### 3.3. Order of Application

After this the order of the application of principles is changed and again we have the same two possibilities. The results with the change order are compared with those before the change to arrive at the conclusion about the correct order of application.
The correct syllables formed were then also analyzed to find out maximum number of consonants allowed in the onset and coda of the syllable and also to find out the templates of syllables for Urdu language.

## 4. RESULTS

### 4.1. Words With No Consonant Clusters

If there are no consonant clusters in the word but only a consonant followed by a vowel i.e. CVCV than Urdu always prefer the consonant in the onset.

**TABLE 1 Syllabification with no C Cluster**

| Words | Syllabification | Syl using MOP |
|---|---|---|
| azar | a.zar | a.zar |
| asɪja | a.sɪ.ja | a.sɪ.ja |
| burak | bu.rak | bu.rak |
| bʊzʊrg | bʊ.zʊrg | bʊ.zʊrg |

In this case the coda comes only at the word final positions as in the examples table 1. If there is a word having the pattern VCVCVC than the first vowel will form a syllable and the rest will be same as the previous case an example of it can be seen in the first row of Table 1.

### 4.2. Words With Two Consonants Together.

In case of two consonants together in the middle of the word, one goes into the coda of the previous syllable and the other one goes into the onset of the next syllable.

**TABLE 2 Words with 2 Consonants**

| Words | Syllabification |
|---|---|
| abdoz | ab.doz |
| ɪbtɪda | ɪb.tɪ.da |
| xʊrʃɪd | xʊr.ʃɪd |
| xuʃbu | xuʃ.bu |
| rəftar | rəf.tar |

### 4.3 Words With Three Consonants Together

This case is mostly avoided in Urdu by epenthesis and phonotactic constraints. But if it happens than Urdu prefers to have more codas than onsets. It means it would have two codas and one onset of the three consonants together.

**TABLE 3 Words with 3 Consonants**

| Words | Syllabification |
|---|---|
| bənd$^h$na | bənd$^h$.na |
| gondni | gond.ni |
| lʊndmʊnd | lʊnd.mʊnd |
| læhnga | læhn.ga |
| bʊrdbar | bʊrd.bar |

### 4.4. Order of Application

The order of application of the principles in each of the above results was observed to be Maximal Onset Principal first and than Sonority Sequence Principal so as to check if the syllables formed by MOP violate SSP or not. If they do they are changed so as to be following the SSP.

### 4.5. Clash of Principles

If there is a clash in the syllabification of the two principles than SSP always wins over MOP in forming the correct syllable.

**TABLE 4 Clash of Principles**

| Original Syllabification | Syllabification Using MOP | Syllabification Using SSP |
|---|---|---|
| ab.doz | a.bdoz | ab.doz |
| ab.kari | a.bka.ri | ab.ka.ri |
| ɪb.tɪ.da | ɪ.btɪ.da | ɪb.tɪ.da |
| xuʃ.bu | xu.ʃbu | xuʃ.bu |
| dəs.ta.vez | də.sta.vez | dəs.ta.vəz |

Meaning that the syllabification of MOP is altered so as to satisfy SSP. Given no phonotactic constraint it would result in correct syllabification.

### 4.6. Phonotactic Constraints

There are two types of phonotactic constraints in Urdu.

1. The first one is of consonant patterns that have an increasing order of sonority. This constraint makes it possible to avoid multiple onsets in Urdu. Some examples of phonotactic constraints are given in table 5.

**TABLE 5 Examples of Phonotactic Constraints**

| Original | With MOP | With SSP | Ph.Const |
|---|---|---|---|
| hʊl.ja | hʊ.lja | hʊ.lja | lj |
| dʊk$^h$.ri | dʊ.k$^h$ri | dʊ.k$^h$ri | k$^h$r |
| səb.zi | sə.bzi | sə.bzi | bz |
| kat.na | ka.tna | ka.tna | tn |

2. Secondly Urdu doesn't allow affricates to follow or precede fricatives or stops in a syllable.

All other combinations of consonants other than these two constraints are allowed in Urdu.

### 4.7. Complex Onsets And Codas

Urdu doesn't haves any complex onsets. The first phonotactic constraint in sec 4.6 and epenthesis discussed in sec 5.4.1 stop it from having multiple onsets. Urdu has complex Codas and allows two consonants at the maximum in the coda. These complex codas can be in the middle or at the word final positions.

**TABLE 6 Complex Codas**

| Words | Syllabification |
|---|---|
| tɪlɪsm | tɪ.lɪsm |
| gond$^h$ni | gond$^h$.ni |
| ʃəxs | ʃəxs |

### 4.8. Templates for Urdu Syllables

From the analysis the following templates of syllables were observed. These templates along with examples of the syllables are shown in table 7.

**TABLE 7 Syllable templates for Urdu**

| Templates | Syllables | Words |
|---|---|---|
| V | ə | ə.tʃa.nək |
| VV | a | a.xi.rət |
| VC | ɪb | ɪb.rət |
| VVC | ab | ab.pa.ʃi |
| VCC | ərdʒ | ərdʒ.mənd |
| CV | ʃɪ | ʃɪ.hab |
| CVC | dət | ʃə.ha.dət |
| CVCC | bənd$^{h}$ | bənd$^{h}$.na |
| CVV | ba | ba.ri |
| CVVC | rat | rat |
| CVVCC | saxt | saxt |

## 5. DISCUSSION

### 5.1. Words With No Consonant Clusters

From the analysis it was found that Urdu is an onset loving language. Therefore when there is only one consonant in between two vowels than it prefers it in the onset rather than in the coda (table 1). An extreme case is one in which the word starts with a vowel for example ”a.xi.rət” instead of the first vowel forming a coda with the next consonant it forms a syllable of only one vowel ”a” and the consonant goes into the onset of the next vowel “xi”. When the pattern is simple that is there are no consonant clusters we don’t need to apply SSP after MOP because SSP would never be violated, actually the reason for applying SSP after MOP is to check that the syllables formed by MOP don’t violate SSP, and that in this case would never be violated as there is only one consonant before the vowel.

### 5.2. Words With Two Consonants Together

If there are two consonants together at the end of the word than they both will go in the onset or the coda respectively (table 6).
But if these two consonants are in the middle of the word than one of them goes to the coda and the other one goes to the onset (table 2). That coda than helps in determining the stress on the syllable by making it a heavy syllable, it thus explains the sensitivity of the Urdu language.

### 5.3. Order of Application & Clash of Principles

The order of application of the two principles is to apply MOP first and SSP afterwards. If the two principles give the same result and it is not the correct syllabification than it is always due to the phonotactic constraints. But if the syllabification of MOP doesn’t satisfy SSP than the colliding syllable is modified with respect to sonority and it results in correct syllabification of the word (table 4). Thus it shows that SSP has an upper hand in the syllabification of the Urdu.

### 5.4. Words With Three Consonants Together

Now comes the case when there are three consonants together in a cluster. It has been observed that there are very few words in Urdu that have three consonants in a single cluster. This ratio is as low as 8 out of approx 5000 words analyzed. And mostly these words are formed as a result of affixation e.g. “dard-mand”. Now when the syllabification principles are applied on theses words they result in two codas and one onset (table 3). And in this case also the order of application of principles is MOP first and SSP following it. But if we follow this order we would end up with two onsets rather than two codas, for example in case of “dard-mand” the above sequence will result in the following syllabification” dar-dmand”. Here again the phonotactic constraints come into play and prevent the two-onset syllabification.

#### 5.4.1. Epenthesis to prevent multiple Onsets

Some times the occurrence of two onsets is prevented by epenthesis. An example of it is in case of “ih.tram” all of the native speakers syllabify it like this. But in Urdu dictionary it is syllabified as ”ih.ti.ram” thus preventing the two-onset case by the epenthesis of a vowel ‘i’. There are other words also that support the fact that Urdu uses epenthesis to prevent multiple onsets.

### 5.5. Phonotactic Constraints

Now come the phonotactic constraints in Urdu. In general it has been seen that Urdu prefers only one onset, as it repels the case of two onsets by applying epenthesis and other phonotactic constraints. The phonotactic constraints observed in this analysis are all of increasing sonority. Some examples are given in figure 5.

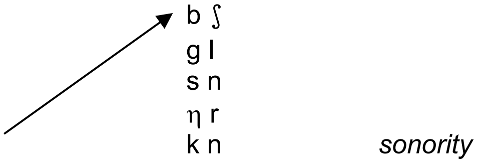

**FIGURE 4 Increase in Sonority for Phonotactic Constraints**

From this we can very easily see that these phonotactic constraints are to avoid multiple onsets, because if these combinations are allowed before a vowel than the MOP and SSP will result in the same syllabification and that syllabification will have two onsets (table 5). At this point these phonotactic constraints come into play to avoid the two onsets. During the entire analysis of about 5000 words no words with affricates following or preceding fricatives or stops were found. Thus it was inferred from this observation that it is also a phonotactic constraint in Urdu.

### 5.6. Syllable Templates

After all this analysis the patterns of the syllables were observed and noted down these templates along with examples are shown in table 7.

## 6. ALGORITHM FOR URDU SYLLABIFICATION USING MOP AND SSP

Based on the entire analysis an algorithm is proposed for the syllabification of Urdu words using only MOP and SSP. Following are the steps for that principle. The direction is from Left to Right.

1. First apply MOP, followed by SSP on the transcription of the target word.
2. If syllabification of the two principles is same than go to step 4.
3. But if the two syllabifications are different than consider the one resulting after SSP application and move on to step 4.

Now look for multiple onsets other than the first syllable. If there is none than it is the correct syllabification. But if there are multiple onsets than transfer the first onset to the coda of the previous syllable.

5. After removing all the multiple onsets u will have the original syllabification.

### 6.1. Examples of Syllabification Using the Algorithm

**TABLE 8: 1st Example for syllabification**

| | | |
|---|---|---|
| Step 1 | a.bxo.rəh<br>**(MOP)** | a.bxo.rəh<br>. . **(SSP)** |
| Step 2 | Same no clashes | |
| Step 4 | Moving b to previous syllable<br>ab.xo. rəh | |
| Step 5 | The correct syllabification<br>ab.xo. rəh | |

**TABLE 9: 2ND Example for syllabification**

| | | |
|---|---|---|
| Step1 | ə.rdZmənd<br>(MOP) | ər.dZmənd<br>(SSP) . . . . |
| Step 2 | Not applicable | |
| Step 3 | selecting the one after SSP<br>ər.dZmənd | |
| Step 4 | ərdZ.mənd | |
| Step 5 | Original Syllabification ərdZ.mənd | |

## 7. REFERENCES

Bokhari, S. 1985. Phonology of Urdu Language. Royal Book Company, Karachi.

Goldsmith, A. 1990. "Auto segmental and Metrical Phonology", Massachususs Basil Blackwell LTD

Hussain, S. 1997. "Phonetic Correlates of Lexical Stress in Urdu." Unpublished Ph.D. dissertation, Northwestern University, IL, USA.

Mamoru, Shibayama and Satoshi, Hoshino 2001 "Thai Morphological Analyses Based on the Syllable Formation Rules" http://www.ipsj.or.jp/members/JInfP/Eng/index.html
Site of the Information Processing Society Japan.

Masica,C Cambridge 1991."The Indo Aryan" languages Cambridge University

Qureshi, B, 1992. "Standard Twentieth Century Dictionary: Urdu into English". Delhi Photo Offset Printers