English to Urdu Transliteration System

Abbas Raza Ali and Madiha Ijaz
Center for Research in Urdu Language Processing,
National University of Computer and Emerging Sciences, Lahore, Pakistan
{abbas.raza, madiha.ijaz}@nu.edu.pk

Abstract

Urdu language processing applications encounter non-Urdu text specifically English text frequently. The accuracy of these systems e.g. machine translation, text-to-speech etc. is highly undermined as they are unable to handle English text. One possibility could be addition of multilingual language processing capabilities in Urdu language processing applications so that they may handle English text also along with Urdu but this approach is quite taxing. Another approach to handle English text is to transliterate it into Urdu automatically and then pass it on to the Urdu language processing applications.

This paper describes English to Urdu transliteration system. First the mapping rules that are used to generate Urdu text from English transcription are discussed then syllabification, manual transliteration and Urduization phase is described and finally the issues related to Out-Of-Vocabulary (OOV) are discussed.

1. Introduction

Transliteration is a method of transcribing the words from one script to phonetically equivalent words in another script. Transliteration rules provide mapping for the letters of the source script alphabet to the letter of target script alphabet based on phonetic similarity. This process is very successful for transliteration of names of people, places, companies, etc., because the translated dictionaries can never be comprehensive and are ineffective for translation of proper noun [1].

Websites, user interfaces etc. contain a lot of English text along with Urdu text. The text is read by the applications like screen reader, web page reader etc. and is passed to the Urdu language processing application e.g. machine translation for translation or text-to-speech system for speech generation. Urdu TTS or machine translation system being unable to handle English text discards it and as a result generated speech or translation lacks coherence.

English to Urdu transliteration system is being developed to eradicate this discrepancy as shown in Table 1.

Table 1: Effect of transliteration on Urdu TTS system

Urdu Text	احمد Nokia میں Alex کے ساتھ ملازمت کرتا ہے۔
With Transliteration	احمد نوکیا میں ایلکس کے ساتھ ملازمت کرتا ہے۔
Without Transliteration	احمد میں کے ساتھ کام کرتا ہے۔

In order to develop English to Urdu transliteration system, first the rule-based approach employing transliteration from English orthography to Urdu orthography was explored but soon it was realized that it would not work well as there is no one-to-one mapping between English orthography and its corresponding sound e.g. /ʃ/ sound is represented using six different letter combinations i.e. motion /moʊ.ʃən/, ocean /oʊ.ʃən/, sure /ʃʊ.ər/, she /ʃi/, admission /æd̩.mɪ.ʃən/, machine /mə.ʃin/ etc [2]. So pronunciation based transliteration was chosen as it produced better results.

An Arpabet based English pronunciation lexicon is used for acquiring pronunciation of English words. English text is converted to Urdu using English pronunciation and mapping rules. The English pronunciation lexicon is based on American accent, hence the transliteration into Urdu also depicts American accent. Frequently used English words are transliterated manually and some rules are applied for Urduization of the transliterated text in order to make it appropriate and as close as possible to the local

accent i.e. the accent that is used in Pakistan while speaking English.

Out-Of-Vocabulary problem is resolved using statistical techniques by first aligning English orthography to pronunciation sequences. Optimal pronunciation of an unknown word is computed by picking maximum probable pronunciation and then passing it for the same transliteration process.

The architecture of the English to Urdu Transliteration system is shown in figure 1.

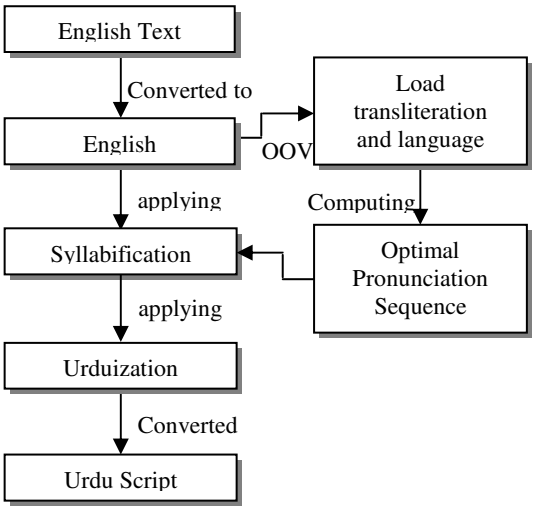


Figure 1: Architecture of English-to-Urdu transliteration

2. English to Urdu mapping

CMU pronouncing dictionary (v 0.7a) is used to acquire pronunciation of English words. The dictionary comprises of 125,000 English words and their corresponding transcription in Arpabet. The pronunciation provided is based on American accent [11].

The phonemic inventory of English comprises of 24 consonants and 15 vowels. The phonemic inventory of Urdu comprises of 37 consonants and 16 vowels (Appendix B). English consonants can be easily mapped to Urdu consonants and there is one-to-one correspondence between them in all cases. There are some sounds in English e.g. dental fricatives, /Θ/ and /Ð/ which are non-existent in Urdu and hence they are mapped to their closest counterpart i.e. dental stops /ʈʰ/ and /ɖ/ respectively.

Some sounds have multiple realizations in Urdu orthography e.g. /s/ can be realized as ص، س، ث etc so in this case only one most commonly used alphabet is chosen which is س in this case. Similar is the case with /z/ as it can be realized as ض، ز، ذ، ظ and /ʈ/ which can be realized either as ت or ط.

Vowels in Urdu are represented using diacritics i.e. zair, zaber and paish and four letters alif, wao, choti yeh and bari yeh. Combination of diacritics with consonants forms short vowels while diacritics combined with alif, wao, choti yeh and bari yeh, form long vowels [3]. Same vowel is represented differently in orthography, depending on whether it exists word initially, medially or finally.

Short vowels occurring word initially use alif as place holder e.g. "urban" is transliterated to اَربَن /'ər.bən/ but when they occur word medially they are represented only by the diacritics e.g. "justly" is transliterated to جَسٹلِی /'ʤəsʈ.li/.

Short vowels when occur word finally are transformed into their corresponding long vowel i.e. zabar is converted to alif e.g. "Andorra" /Æ n d ɔ ɹ ʌ/ is transliterated to اَینڈَورَا /æn.'ɖɔ.rɑ/, similarly zair is converted to choti yeh and pesh is converted to wao.

Hence there is no one-to-one correspondence between English and Urdu vowels in most of the cases and an English vowel is transliterated using multiple Urdu characters depending on whether it occurs word initially, medially or finally as shown table 2.

Table 2: English vowels mapping to Urdu orthography

Arpabet	IPA	Urdu		
		Initial	Middle	Final
AA	ɑ	آ	◌َا	◌َا
AE	Æ	اَی	◌َی	ے
AY	Aɪ	آئِ	◌َائِ	◌َائی
AW	Aʊ	آؤ	◌َاؤ	◌َاؤ
AO	ɔ	اَو	◌َو	◌َو
OY	ɔɪ	آئِ	◌ائِ	ائے
EH	ɛ	اَی	◌َی	ے

ER	ɝ	اَر	ـَر	َر
EY	Eɪ	ای	ی	ے
IH	ɪ	اِ	ـِ	ِ
IY	I	اِی	ـِی	ِی
OW	Oʊ	او	و	و
UH	ʊ	اُ	ـُ	ُ
AH	ʌ	اَ	ـَ	ا َ

3. Syllabification

English-to-Urdu transliteration using CMU Pronunciation dictionary which is based on American accent, generates a lot of inconsistency. To improve system's accuracy; Urdu syllabification is applied on English transcription as shown in table 3.

Consonant and Vowels combine to make syllable and breaking up a word into syllables is known as syllabification. Sonority sequence principle for syllabification is commonly used in Urdu. It requires the onset to rise in sonority towards the nucleus and codas to fall in sonority from the nucleus [9].

3.1 Algorithm

Template matching technique is used to syllabify English transcription. In this technique syllabication is done by matching template of the form $C_{0,1}.V.C_n$ [4]. Urdu allows only one consonant in the onset position and multiple consonants can come in the coda position of the syllable.

1. Convert the entire phonemic transcription of the word to consonant-vowel pairs
2. Start from the end of the word, traverse backwards to find the next vowel
3. **repeat**
4. **if** there is a consonant preceding it
5. mark a syllable boundary before consonant
6. **else**
7. mark the syllable boundary before this vowel
8. **end if**
9. **until** the entire string is consumed

Table 3: Transliteration after applying syllabification

English	IPA	Urdu	
		Unsyllabified	Syllabified
Associate	ʌ . s oʊ . ʃ i . ʌ t	اَسوشیَٹ	اَسوشی ایٹ
Oblivious	ʌ b . l ɪ . v i . ʌ s	اَبلِویَس	اَب لِی وی اَس
Obedient	oʊ . b i . d i . ʌ n t	اوبیڈیَنٹ	او بِی ڈِی اَنٹ

3.2. Special case

After applying syllabification there exists problem of local accent, as transliteration is based on American accent so in order to make transliteration closer to Urdu accent some rules are applied on syllabified transliterated text.

3.2.1. Consonant Cluster. Urdu syllabification does not allow consonant cluster in onset of the syllable and in the word medial position. In this case add /ɪ/ if the second consonant is 'r' or 'l' otherwise /ʌ/ between two consonants and mark syllable boundary after it as shown in table 4.

Table 4: Examples of consonant cluster problem

English	IPA	Urdu	IPA
Treehopper	t ɹ I h ɑ p ɝ	ٹِریحَاپر	t ɪ . ɹ I . h ɑ . p ɝ
Bless	b l ɛ s	بِلَیس	b ɪ . l ɛ s
Quickly	k w ɪ k l i	کَوِکلِی	k ʌ . w ɪ k . l i

3.2.2. Urduization. If two consonants come in the onset of the syllable in the word initial position and the starting consonant is 's' then add /ɪ/ before 's' as shown in table 5.

Table 5: Examples of Urduization applied on transliterated Urdu text

English	IPA	Urdu	IPA
School	s k u l	اِسکُول	ɪ s . k u l
Skill	s k ɪ l	اِسکِل	ɪ s . k ɪ l
Special	s p ɛ ʃ ʌ l	اِسپَیشَل	ɪ s . p ɛ . ʃ ʌ l

4. Out-Of-Vocabulary problem

Out-Of-Vocabulary is a very common problem in various systems like text-to-speech, machine translation, cross language information retrieval (CLIR), etc. To resolve this problem, English phoneme to orthography alignment has to be found out probabilistically to get one-to-one mapping between them as shown in table 6, and then train those aligned sequence to get most probable pronunciation for an unknown word.

Table 6: English orthography to pronunciation alignment

English	Percentages
Pronunciation	p . er . s . eh . n . t . ih . jh . ah . z
Alignment	p(p) . er(er) . c(s) . e(eh) . n(n) . t(t) . a(ih) . g(jh) . e(ah) . s(z)

The entire procedure consists of two steps;

- English orthography to pronunciation alignment.
- Computing optimal pronunciation sequence.

After getting pronunciation of unknown text, it will be passed through the same procedure like syllabification and then Urdu transliteration. The architecture of the OOV module is shown in figure 2.

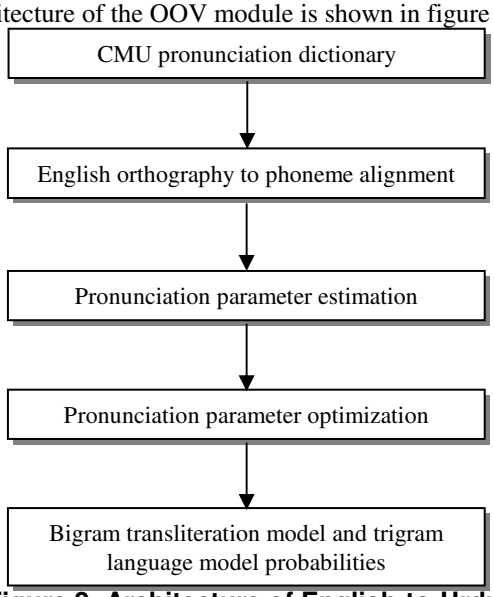


Figure 2: Architecture of English-to-Urdu transliteration

4.1. Orthography to pronunciation alignment

In this step all the valid combinations of English orthography E_n to its pronunciation sequence P_r are produced using conditional probability;

$$\hat{P}_r = \arg\max_{P_r} p(P_r \mid E_n) = \arg\max_{P_r} p(E_n \mid P_r)p(P_r) \quad (1)$$

The trigram language model $p(P_{r_{i-1}}.P_{r_i}.P_{r_{i+1}})$ and bigram transliteration model $p(E_{n_i}.P_{r_i})$ is combined to maximize the pronunciation probability P_r.

4.2. Computing optimal pronunciation sequence

Expectation maximization algorithm is used to compute optimal alignment sequence. The algorithm is given below;

Initialization

For each English phoneme to orthography pair, assign equal weights to all possibilities generated from (1).

repeat

Expectation-Step

For each of the Arpabet phonemes, count up instances of its different mappings from the observations on all combinations produced in (1). Normalize the score so that the mapping probabilities sum to 1.

Maximization-Step

Recalculate the combination scores. Each combination is scored with the product of the scores of the symbol mappings it contains. Normalize the scores so that the mapping probabilities sum to 1.

until convergence

5. Results

Transliteration process becomes more accurate after applying syllabification on the pronunciation and finding probabilistic sequences of Out-Of-Vocabulary word problem as shown in figure 3.

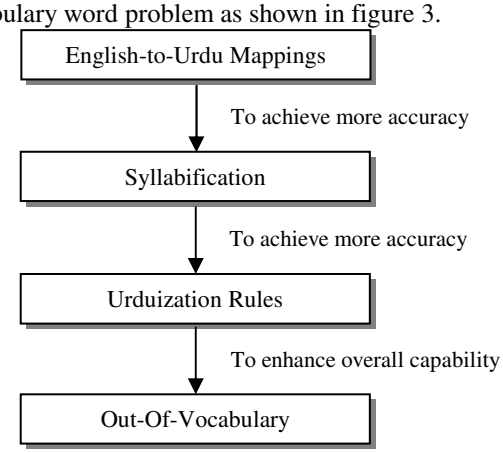


Figure 3: Modules of the system that lead it towards maturity

The System's accuracy is recorded after maturity of every independent module as mentioned in figure 3. The lexicon of most frequently used words of English (15,237 words from British national corpus (BNC)) was transliterated into Urdu using the transliteration system. Accuracy without applying syllabification and resolving unknown word problem is described in table 7 in detail. The results are generated by passing transliterated text to Urdu text-to-speech system and analyzing its output.

Table 7: English-to-Urdu mapping accuracy

Observations	Total Size
Correct Mapping (after applying rules)	12,940
Incorrect Mapping (due to Syllabification)	173
Incorrect Mapping (due to OOV)	2,124
Total	**15,237**
Accuracy (%)	**84.92**

After applying syllabification technique; out of 173 syllabication problems, 91% are resolved (manual testing). The accuracy of OOV is evaluated automatically by using automatic evaluation method Bilingual Evaluation Understudy BLEU [10] as shown in table 8.

Table 8: Overall system accuracy

Modules	Correct	Total Size	Accuracy (%)
Mapping	12,940	12,940	100.00
Syllabification	158	173	91.31
OOV	1,518	2,124	72.46
Total	**14,616**	**15,237**	**95.92**

6. Conclusion

Transliteration is a good technique which helps a system adding multi-lingual ability. It can be used in various Systems, e.g. text-to-speech, information retrieval, machine translation, English-to-Urdu parallel corpus Consistency in Proper Names etc. Overall system's accuracy is 96% which is quite promising. The System can be improved by training transliteration model on Urdu accent instead of American.

7. Acknowledgements

The work on English to Urdu transliteration system has been carried out in a project that involves development of an open-source Urdu screen reader for visually impaired people funded by National University of Computer and Emerging Sciences (NUCES), Pakistan.

8. References

[1] W. Gao., K. F. Wong and W. Lam. "Phoneme-based Transliteration of Foreign Names for OOV Problem". *In First International Joint Conference on Natural Language Processing*, Pages 374-381, 2004.

[2] Saleem, M. "Urdu Rasmulkhat ki Jaamiat". *Akhbar-i-Urdu*, Pages 6-10, Islamabad, Pakistan, 2002.

[3] S. Hussain, "Letter-to-Sound Rules for Urdu Text to Speech System". *Proceedings of Workshop on Computational Approaches to Arabic Script-based Language, COLING-2004*, Geneva, Switzerland, 2004.

[4] S. Hussain, "Phonological Processing for Urdu Text to Speech System". *Yadava, Y, Bhattarai, G, Lohani, RR, Prasain, B and Parajuli, K (eds.) Contemporary issues in Nepalese linguistics*. Katmandu, Linguistic Society of Nepal, 2005.

[5] J. Kominek, and A. W. Black, "Learning Pronunciation Dictionaries: Language Complexity and Word Selection Strategies". *In Proceedings of the Human Language Technology Conference of the NAACL*, Pages 232-239. New York City, USA, 2006.

[6] J. Lewis, , K. McGrath, and J. Reuppel, "Language Identification and Language Specific Letter-to-Sound Rules". *Colorado Research in Linguistics*, Volume 17, Issue 1, June 2004.

[7] J. Martin, , R. Mihalcea, and T. Pedersen, "Word Alignment for Languages with Scarce Resources". *In Proceedings of the ACL Workshop on Building and Exploiting Parallel Texts: Data Driven Machine Translation and Beyond*, Ann Arbor, MI, June 2005

[8] A. Sen, "Pronunciation Rules for Indian English TTS System". *Workshop on Spoken Language Processing*, Mumbai, India, January 2003

[9] R. Bokhari, and S. Pervez, "Syllabification and Re-Syllabification in Urdu". *Akhbar-i-Urdu*, Pages 63-67, Islamabad, Pakistan, 2003.

[10] K. Papineni, S. Roukos, , T. Ward, , and W. J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation". *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pages 901–904, 2002.

[11] CMU. "The CMU Pronunciation Dictionary", *www.speech.cs.cmu.edu/cgi-bin/cmudict*, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 2006.

Appendix A - English-to-Urdu Mappings

Arpabet	IPA	Urdu			Arpabet	IPA	Urdu		
		Initial	Middle	Final			Initial	Middle	Final
AA	ɑ	آ	◌َا	◌َا	**L**	L	ل	ل	ل
AE	Æ	اَی	◌َی	ے	**M**	M	م	م	م
AH	ʌ	اَ	◌َ	◌َا	**N**	N	ن	ن	ن
AO	ɔ	اَو	◌َو	◌َو	**NG**	ŋ	نگ	نگ	نگ
AW	Aʊ	آؤ	◌َاؤ	◌َاؤ	**OW**	Oʊ	او	و	و
AY	Aɪ	آئِ	◌َائِ	◌َائی	**OY**	ɔɪ	◌آئِ	وائِ	ائے
B	B	ب	ب	ب	**P**	P	پ	پ	پ
CH	ʧ	چ	چ	چ	**R**	ɹ	ر	ر	ر
D	D	ڈ	ڈ	ڈ	**S**	S	س	س	س
DH	Ð	د	د	د	**SH**	ʃ	ش	ش	ش
EH	ɛ	اَی	◌َی	ے	**T**	T	ٹ	ٹ	ٹ
ER	ɝ	اَر	◌َر	◌َر	**TH**	Θ	تھ	تھ	تھ
EY	Eɪ	ای	ی	ے	**UH**	ʊ	اُ	◌ُ	◌ُ
F	F	ف	ف	ف	**UW**	U	اُو	◌ُو	◌ُو
G	G	گ	گ	گ	**V**	V	و	و	و
HH	H	ح	ح	ح	**W**	W	و	و	و
IH	ɪ	اِ	◌ِ	◌ِ	**Y**	J	ی	ی	ی
IY	I	اِی	◌ِی	◌ِی	**Z**	Z	ز	ز	ز
JH	ʤ	ج	ج	ج	**ZH**	ʒ	ژ	ژ	ژ
K	K	ک	ک	ک	-				

Appendix B - English Phonemic Inventory

Vowels

Arpabet | IPA	Front				Central	Back		
Closed	IY | I							UW | U
			IH | ɪ			UH | ʊ		
Closed-Middle		EY | Eɪ						OW | Oʊ
Middle								OY | ɔɪ
Open-Middle			EH | ɛ		ER | ɝ	AH | ʌ		AO | ɔ
				AE | Æ				
Open				AY | Aɪ	AW | Aʊ		AA | ɑ	

Consonants

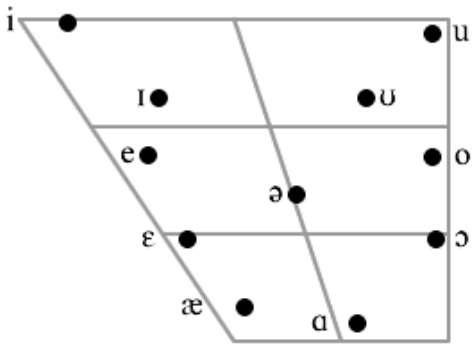
Arpabet | IPA	Bilabial	Labio-dental	Labio-velar	Dental	Al-velar	Post-alveolar	Pala-tial	Velar	Glottal
Plosive	P | p B | b			T | t D | d				K | k G | g	
Nasal	M | m			N | n				NG | ŋ	
Affricative						CH | ʧ JH | ʤ			
Fricative		F | f V | v		TH | Θ DH | Ð	S | s Z | z	SH | ʃ ZH | ʒ			HH | h
Approximant			W | w	R | ɹ			Y | j		
Lateral Approximant				L | l					

Appendix C - Urdu Phonemic Inventory

Vowels

IPA	Initial	Middle	Final
/ə/	◌ا	◌	◌
/ɑ/	◌آ	◌ا	◌ا
/ɪ/	◌ا	◌ِ	◌ِ
/i/	◌ا	◌ِی	◌ی
/ʊ/	◌ا	◌ُ	◌ُ
/u/	◌او	◌ُو	◌و
/e/	◌ا	◌ِی	◌ے
/ɛ/ | /ɑɪ/	◌ا	◌ِی	◌ے
/o/	◌او	◌ُو	◌و
/ɑu/	◌او	◌ُو	◌و

IPA	Letter	Description
/ə/	◌َ	Zabar
/ɪ/	◌ِ	Zer
/ʊ/	◌ُ	Paish



Consonants

IPA	Bilabial		Labio-Dental		Dental		Alveolar		Retroflex		Post-Alveolar		Velar		Uvular| Glottal	
Plosive | Stops	p p^h	b $b^ɦ$			t̪ $t̪^h$	d̪ $d̪^ɦ$			ʈ $ʈ^h$	ɖ $ɖ^ɦ$			k k^h	g $g^ɦ$	q	ʔ
Nasal	m	$m^ɦ$					n	$n^ɦ$					ŋ	$ŋ^ɦ$		
Affricate											tʃ $tʃ^h$	dʒ $dʒ^ɦ$				
Fricative			f	v			s	z			ʃ	ʒ	x	ɣ	h	
Trill							r	$r^ɦ$								
Lateral							l	$l^ɦ$								
Flap									ɽ	$ɽ^ɦ$						
Approximant											j					

Appendix D - Transliteration

English	Urdu	English	Urdu	English	Urdu
abandon	اَبینڈَن	daniel	ڈَینیَل	sponsorships	سپانسَرشِپس
abilities	اَبِلَٹِیز	expressly	اَیکسپرَیسلِی	strategies	سٹرَیٹَجِیز
abuse	اَبیُوس	financial	فَنَینشَل	syria	سِرِیَا
accelerated	اَیکسَیلَریٹَڈ	flourishing	فلَرِشنگ	systematic	سِسٹَمیٹِک
acknowledgment	اَیکنَالِجمَنٹ	gradually	گرَیجُوَلِی	thriving	تھرائِونگ
ascent	اَسَینٹ	handling	حَینڈلِنگ	turkey	ٹَرکِی
aspiration	اَیسپرَیشَن	interview	اِنٹرویُو	urban	اَربَن
authoritative	اَتھورَٹیٹِو	iran	اِرَان	urgently	اَرجَنٹلِی
babies	بیبِیز	justifications	جَسٹِفِکیشَنز	veterinary	وَیٹرَنَیرِی
banned	بَینڈ	kennedy	کَینَڈِی	visually	وِژوَلِی
belgium	بَیلجَم	lithuania	لِتھوینیَا	vulgar	وَلگَر
boxing	بَاکسِنگ	luxembourg	لَکسَمبَورگ	wellington	وَیلِنگٹَن
brackets	بریکَٹس	mathematics	مَیتھَمیٹِکس	williams	وِلیَمز
bradford	بریڈفَرڈ	maxwell	مَیکسوَیل	workshops	وَرکشَاپس
bravery	بریوَری	morphological	مَورفَلَاجِکَل	wrapping	رَیپِنگ
bribes	برائِبز	neighbouring	نیبَرِنگ	yorkshire	یَورکشَر
chilly	چِلِی	nominating	نَامَنیٹِنگ	youth	یُوتھ
chocolate	چوکلَٹ	outstanding	آؤٹسٹَینڈِنگ	yugoslavia	یُوگوسلَاوِیَا
computer	کَمپیُوٹَر	physiological	فِزِیَلَاجِکَل	zimbabwe	زِمبَابوے
dangerously	ڈینجَرَسلِی	projects	پرَاجَیکٹس	zoology	زوَالَجِی