Proceedings of the Conference on

LANGUAGE & TECHNOLOGY

2012

مرکز تحقیقات لسانیات

Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore- Pakistan

Organized by
Society for Natural Language Processing
In collaboration with
Center for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore- Pakistan

www.snlp.org.pk
ISBN: 978-969-9690-05-1

Conference Committees

Organizing Committee

General Chair:	Sarmad Hussain, *CLE, KICS-UET, Lahore, Pakistan*
Technical Committee Co-chair:	Abid Khan, *University of Peshawar, Pakistan*
Technical Committee Co-chair:	Miriam Butt, *University of Konstanz, Germany*
Publication Committee Co-chair:	Tafseer Ahmed, *University of Karachi, Pakistan*

Technical Committee

Abid Khan, University of Peshawar, Pakistan (co-chair)
Miriam Butt, University of Konstanz, Germany (co-chair)
Afaq Husain, Riphah University, Pakistan
Chai Wutiwiwatchai, NECTEC, Thailand
Douglas Clarke, Cranfield University, UK
Elena Bashir, University of Chicago, USA
Ghulam Raza, PIEAS, Pakistan
Imran Siddiqi, Bahria University Islamabad, Pakistan
Khaver Zia, Beaconhouse National University, Pakistan
Key-Sun Choi, KAIST, Korea
Muhammad Afzal, KICSIT, Pakistan
Pushpak Bhatacharyya, IIT Bombay, India
Rachel Roxas, De La Salle University, Philippines
Rajeev Sangal, IIIT Hyderabad, India
Sarmad Hussain, KICS-UET, Pakistan
Seemab Latif, NUST, Pakistan
Tafseer Ahmed, University of Karachi, Pakistan
Virach Sornlertlamvanich, NECTEC, Thailand
Yogendra Yadava, Tribhuvan University, Kathmandu, Nepal

Program Committee:

Asad Abbas Tahir, CLE, KICS-UET, Lahore, Pakistan
Atif Ali Javed, CLE, KICS-UET, Lahore, Pakistan
Ammara Amanat, CLE, KICS-UET, Lahore, Pakistan
Afia Mahmood, CLE, KICS-UET, Lahore, Pakistan
Ayesha Zafar, CLE, KICS-UET, Lahore, Pakistan
Farhat Abdullah, CLE, KICS-UET, Lahore, Pakistan
Sana Shams, CLE, KICS-UET, Lahore, Pakistan

Foreword

On behalf of the Organizing Committee, we welcome the authors and participants to the fourth Conference on Language and Technology.

Center for Language Engineering (CLE) at Al-Khawarizmi Institute Computer Science (KICS), UET, Lahore, is pleased to host the Conference on Language and Technology 2012 (CLT12). As the fourth CLT, this conference is helping mature the language technology research in Pakistan and providing the intended platform for researchers to interact.

Twenty two papers have been submitted for CLT12, of which eleven have been accepted for presentation and four for posters, through a double blind review process by an international technical committee. With two withdrawals, the current volume has thirteen papers. The papers cover Pashto, Punjabi, Sindhi, Urdu languages specifically, and a host of areas, including linguistics and computational aspects of phonetics, phonology, syntax and semantics.

This CLT also presents exciting workshops and demos, including recent research in linguistics of local languages and Nastalique optical character recognition.

On behalf of the Organizing Committee we would like to show gratitude to all who volunteered to plan and support the conference. We would like to thank the technical committee members for their diligent reviews of the research articles. We would also like to thank the conference sponsors, especially University of Konstanz and the German Academic Exchange Service (DAAD). We are grateful to the management of Al-Khawarizmi Institute of Computer Science and University of Engineering and Technology, Lahore, for their unrelenting support to hold the conference.

We wish you all a very fruitful CLT12 and a pleasant stay in Lahore.

Sarmad Hussain
On behalf of the Organizing Committee

Table of Contents

A Computational Multilingual Text Constituent Splitter and Phrasing: A Case of Pashto Language 1
Zaheer Ahmad, Mohammad Abid Khan, Jehan Zeb Khan Orakzai, Rahman Ali, Ibrar Ahmad

Wikipedia is a Practical Alternative to the Web for measuring Co-occurrence based Word Association 9
Om P. Damani, Pankhil Chedda, Dipak Chaudhari

The Pitch Contour of Declarative Sentences in Urdu Language 17
Farhat Jabeen, Sarmad Hussain

Comparing Talks, Realities and Concerns over the Climate Change: Comparing Texts with Numerical and Categorical Data 21
Yasir Mehmood, Timo Honkela

Urdu Syllable: Templates and Constraints 29
Mazhar Iqbal Ranjha

Developing a Part of Speech Tagset for Sindhi 37
Mutee U Rahman

CLE Urdu Digest Corpus 47
Saba Urooj, Sarmad Hussain, Farah Adeeba, Farhat Jabeen, Rahila Parveen

Developing Urdu WordNet Using the Merge Approach 55
Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, Asad Mustafa

An Acoustic Study of Vowel Nasalization in Punjabi 61
Saira Zahid, Sarmad Hussain

Survey of Urdu OCR: An Offline Approach 67
Naila Fareen, Mohammad Abid Khan, Attash Durrani

Poster Papers

Error Analysis of Single Speaker Urdu Speech Recognition System 73
Saad Irtza, Dr. Sarmad Hussain

A Dictionary based Urdu Word Segmentation using Dynamic Programming for Space Omission Problem 81
Rabiya Rashid, Seemab Latif

Complexities and Implementation Challenges in Offline Urdu Nastaliq OCR 85
Danish Altaf Satti, Dr. Khalid Saleem

A Computational Multilingual Text Constituent Splitter and Phrasing: A Case of Pashto Language

Zaheer Ahmad, Mohammad Abid Khan, Jehan Zeb Khan Orakzai, Rahman Ali, Ibrar Ahmad
Department of Computer Science, University of Peshawar, Khyber Pakhtunkhwa, Pakistan
ahmad.zaheer@yahoo.com, abid_khan1961@yahoo.com, janzeb@yahoo.com
rahmanali.scholar@gmail.com, toibrar@yahoo.com

Abstract

We propose a simple rules embedded matrix based method to split input sentences into their constituents and phrases. Splitting a sentence into phrases is a preprocess of machine translation for overcoming the problem of handling long sentences and improving quality of automatic translation. An effort is made to remove or at least minimize the problem of recursion that is faced during the process of phrase splitting thereby saving a lot of time. The system is dynamic in design and theoretically would work for any language that has some type of word order. However we have tested the system on Pashto language and this paper would describe the system in the perspective of Pashto language. The system can achieve more than 90% results keeping in view the Phrase Rules are carefully captured in a table.

1. Introduction

Sentence splitting and getting constituents is a specialized area of chunking. Sentences splitting into constituents are an important preprocess in a wide variety of NLP disciplines, particularly in machine translation and sentence generation. It is not only helpful to overcome the problems of complexity faced during the analysis of long sentences and improve the quality of translation. It can also be used in the translation process directly for complete phrases. This work is based on a language neutral approach to split sentences and get their constituents. However this research paper would mainly focus on Pashto language. This section is dedicated to introduce and to elaborate sentence analysis and recursion. In section-2, related work has been discussed. Section-3 is describing some challenges associated with Pashto language with the particular focus on Unicode for Arabic script. A detail discussion has been provided in section-4 about the Pashto language syntax rules. Section-5 is about the proposed approach to split sentences into their constituents. In the last, summary of the work has been shared.

1.1. Constituent Related Sentence Analysis

A group of words normally functions as a syntactic unit/ constituent/phrases in a sentence [1], [2], [3], [4], [5]. These groupings often give meaning to a sentence and help to identify important information about the structure of constituent boundary, linear order and syntactic categories [1], [2]. These constituents and Phrases can be substituted and replaced, moved in a sentence, deleted, merged and built up by a series of merger operations to form a larger constituent [1], [2], [3]. The order of the constituent in a sentence is as important in the syntactic study of a sentence as the word order [4]. It helps to understand a sentence after removing complexities, helps in translating a sentence and representing the structure of constituent. To model and represent constituent structure, Context Free Grammar (CFG) or phrase structure grammar has been successfully used for languages such as English [6]. However, there are many disadvantages in using CFG for natural languages such as ambiguity, left-recursion and repeated parsing of sub-trees. If a sentence is structurally ambiguous, then the grammar assigns to it more than one parse tree. It will be difficult to use CFG in languages that do not follow strict word order [6].

1.2. Recursion in Syntactic Structure of Sentence

Unlike a chunk, a constituent of a particular category can be embedded inside another constituent of the same category, which, in turn, can be embedded inside another such constituent. This property or set of properties of a sentence are called recursion [7], [8]. Rules that govern these properties are called recursive rules [7]. In recursion the categories like NP or VP repeat itself on the right side of the arrow when written in Context Free Form. For Pashto language some examples are given in section 4.4. Usually recursion is graphically depicted in a tree form as shown in figure-I below. As in the figure-I, given below, when one NP or PP has another NP inside it then the first or head NP or PP is called a possessor phrase [2].

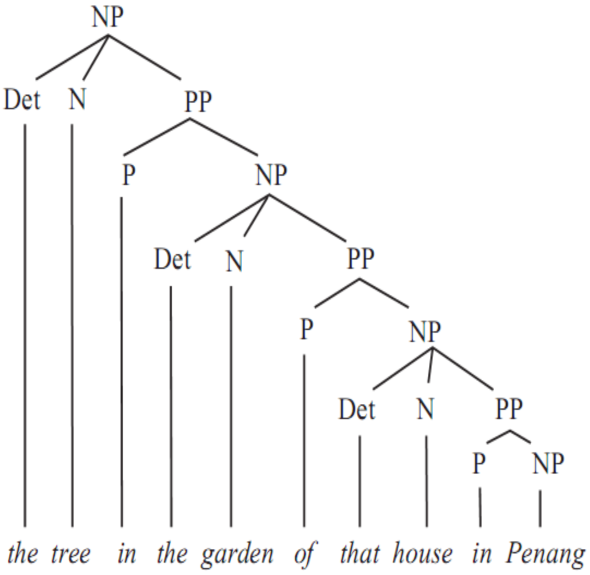


Figure I: Recursion

Pashto being a mixed word order language would not be suitable for processing by CFG because then one has to encounter and code a significant number of grammatical rules. However some special treatment of the same may produce good results. In this paper, CFG is used in the form of matrix to overcome the problems mentioned above.

2. Related Work

Sentence splitting or phrasing is usually carried out as a preprocess of machine translation to remove ambiguity from sentences and sometimes to generate sentence [9], [10], [11]. Different researchers have proposed different techniques. They are using parsing, collecting parts of trees as sentence, POS tagging and head verb or nouns to make phrases. In [10] Constraint Synchronous Grammar (CSG) has been used for this purpose. To identify NPs, Miorelli [12] proposed an ED-CER System for extracting noun phrases from Portuguese sentences based on a parser and a set of Noun grammar rules defined by Perini [13]. However, this approach needs sentences to be manually tagged. Costa [14] used LXGram to describe the grammatical properties and the meaning of Portuguese NPs. In [11], a tree based approach is used to generate phrases by identifying lexical properties of the head verb and the definiteness of arguments and their length. On the other hand, some, authors have used statistical approaches as in [15] statistical recognition of noun phrases with a chunk tagger is used, and it is presented that a part-of-speech tagger can be used for phrasing.

3. Challenges in Pashto Language

Pashto being a low resourced language [16], [17] presents many challenges. Some of them are presented here. First of all, a problem with its Unicode block would be discussed.

3.1. Unicode Problems of Pashto Language

While developing the software for the constituent splitting, it was needed to match Pashto text with words in a lexicon. During the process it was revealed that a word could not be matched in the lexicon despite it is there. On further investigation, the authors came to know that the characters in the target word, though looking similar and same are using different Unicode than the word from the corpus. It was found that different text editors were using different Unicode from the Arabic block of characters, as there are characters that look similar but have different Unicode.

Pashto script comes under the Arabic block of Unicode like many other languages e.g. Urdu, Persian, Panjabi, Sindhi, Balochi and Kashmiri. In Arabic Unicode, a number of duplications are introduced as stated in [18] about Urdu Unicode that as Unicode standard has to cater to multiple existing systems and multiple languages within a script, redundancies are introduced in it. It is advised to re-standardize and short list the characters nationally before usage. However it seems that during designing and encoding, unification of characters have not been duly cared. As it is mentioned in [19] that some extended Arabic characters are typographical variants of characters already fully covered by the corresponding basic Arabic characters. In [19], it is further stated that in some cases it looks that Unicode - knowingly - confuses regional calligraphic or typographic variants for encodable characters. Some justifications are given for the existence of a number of "goal yeh" there in the Unicode but people knowing the languages like Urdu, Arabic and Pashto feel confused and some changes are needed to bring unification in the Arabic Block of Unicode, in order to make it easy for developers as well as for common users. As the Unicode standard [20] clearly states that all similar characters would be unified across languages.

Pashto is natively spoken in a number of countries such as Pakistan, Afghanistan and other parts of the world like the Gulf states, Europe, UK, India and America [21], [22], [23], [24]. Therefore, standardization for each country might differ from the other, then how data transformation would take place. Furthermore, what would be the solution for usage of these scripts on the net and which country's standard would be followed? Therefore, nation wise standardization is not feasible. It is also not a good idea to further devise standards within a standard. It would lead to the situation that existed before Unicode and too many standards for one language script would lead to no standardization at all. In works like [25], the author presented some solutions that need further development. However in the present scenario, despite the Unicode, researchers feel as if there is no standardization for the Arabic based scripts such as Pashto and Urdu. The authors believe that significant achievements can be made through hinting codes of the fonts for each language without using duplication of similar characters.

3.2. Variance in Spellings of Pashto Language

There are a number of dialects of the Pashto language. An example is the word Pashto itself as Pashto can be written and spoken as Pakhto, Pushto, Pukhto, Pashtu, Pakhtu, Pushtu, Pukhtu, Pukkhto, Pukshto, Paktu, Pooshtoo, Passtoo,Pakhtoo, Pakkhtoo and Pasto [16]. The most common characters that are used interchangeably are Kh as sh, o as u and o as oo. These characters offer many challenges while comparing strings therefore one should take care of these issues beforehand. In this paper the authors worked hard to overcome this problem by coding separate module for this.

3.3. Morphology of Pashto

Pashto has many inflectional forms in its major categories such as nouns and adjectives are differentiated for case, number, and gender [17]. However till date no complete work has been presented to capture all morphological variations. In addition, nouns are not necessarily the same class or gender for different speakers, and occasionally there is even variability within a speaker.

4. Pashto language, Syntax and Phrase Rules

Pashto is the language of over 20 million people. Some claim it to be 40-60 million. It is mainly spoken in Pakistan and Afghanistan and has more speakers in Pakistan than Afghanistan [21], [16], [24]. It is also spoken in other parts of the world like in the Gulf, Europe, UK, India and America [23], [22]. Despite that Pashto is a low resourced language [16], [17] and offers many challenges in terms of its complex syntax and phrasal rules.

Pashto is fairly rigidly head-final in NP and VP lexical categories, while several functional categories are head-initial [21]. The basic word order is SOV with some degree of word order freedom and split-ergative language [21]. These and other properties require a considerable amount of effort to capture the phrase structure rules for the language. Numerals and adjectives precede any nouns they modify, suggesting that the lexical category NP is head-final [21]. It is specifically only the lexical projections (VP, NP) that are head-final. With regard to the PP projection, the language appears to exhibit mixed headedness [21]. Some phrasal rules extracted from [26], [27], [28] are given in the subsections below one by one.

4.1. Noun Phrase

- A Noun Phrase consists of a noun or a pronoun together with modifiers that may be an adjective.
- An adjective usually precede a noun but it can appear after noun depending on the context.
- A noun can precede a preposition [phrase].
- The order of the modifiers may be like Preposition + demonstrative(that)+ quantifier + indefinite article(some, a)+descriptive adjective (big, pretty) + noun.
- Adverb that modifies adjectives (very) occurs immediately before adjectives they modify but the order can be altered if the speaker wishes to focus/stress one or the other of the modifiers.

4.2. Verb Phrase

- A verb phrase includes everything except the subject

- Verb is usually the last word in a sentence
- Usual order is a time phrase + Complement/ object+ place phrase + other modifiers + verb
- If object of a preposition is a weak pronoun, the prepositional phrase is almost always positioned just before the verb.
- In negative verb phrase, the negative article " na " occurs before the verb in the imperfective tenses
- In perfective tenses the negative article "na" occurs with simple verbs between the perfective marker and the verb stem.

4.3. Adj Phrase, Adv Phrase and PP Phrase

- Prep comes after and/or before noun or both before and after.
- Adj always comes before noun. Adj can be used as noun.

4.4. Context Free Grammar of Pashto

Based on the work in [29] and the above rules the Context Free Grammar for Pashto is given as below:

- S --> NP+VP|VP|NP+CONJ+VP
- NP --> N|PN |ADJ+N|CN|NP+PP| NP+NP | PP+CONJ+PP | PN | ADJP+ NP | N | NP+ VP | PRON | PP | Det+Adj+NNP+V | Det+Adj+N+VP | Det+Adj+N+Adv+VP
- VP --> V |VP+VP |NP+VP PP+VP|AUX|PPP+VP|PREP+VP|PP+VP|ADJ+V P|ADV+ADJ+V|ADVP+VP |V+NP| NP+V| Adv]+ PP +V
- PP --> PREP+NP |PREP+NP |PP+PP |PREP+NP POSP |PREP+NP |PREP+VP |PREP+N |NP+PP
- ADJP --> ADJ |ADJ+N |ADJ+ADJ

5. The Proposed Methodology

The proposed algorithm is robust and dynamic. It can split a sentence into constituents by taking the dictionary that has grammatical categories in separate column, a matrix having syntactic / word order rules of a language and the corpus from which each sentence has to be splitted into constituents. The table having rules is shown in section 5.3. The complete architecture of the system is given in figure-II, below.

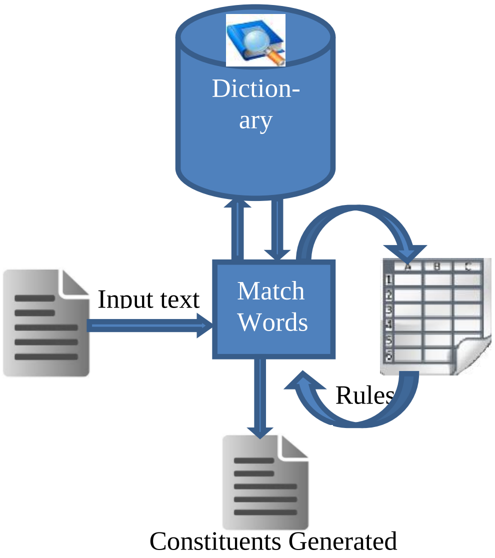


Figure-II. System Diagram

All the rules were developed in a table in MS Excel, lexicon was saved in a database format in MS Access, where as the corpus was in plain text format. C# (Visual Studio) was used to develop and test the system.

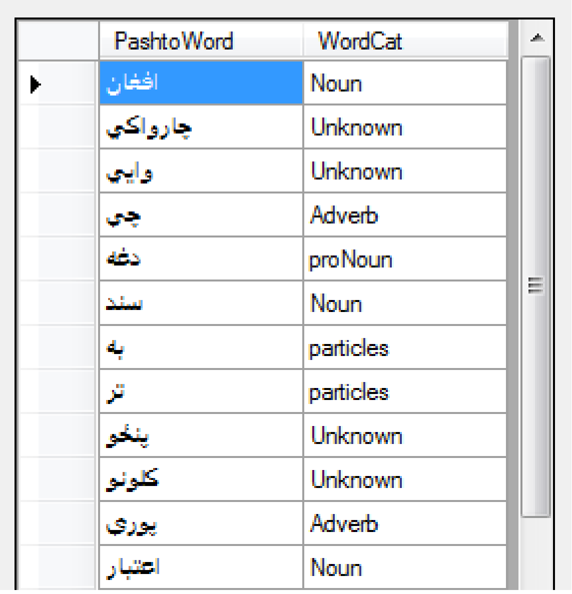
To understand the working of the proposed system, a step wise flow of the system is given in the following subsections.

5.1. Detecting words and the end of sentences

A sentence is read character by character from a text file, and wherever a white space is found a word is marked there. A dot (.) or (-) or (?) is marked as the end of the sentence. The system is developed for singe words only and multiwords are not considered because of the unavailability of a lexicon for the same.

5.2. POS Tagging of the Words

A lexicon of Pashto language having more than 14000 words is used by the software. The dictionary is made of a table, having Pashto word, POS and meaning of the word in English in separate columns. On reading the text from the text file, each word is matched in the dictionary with its list of words. Matching a Pashto word with the dictionary suffers badly from the Unicode problems as discussed earlier in this work. Some extra lines of code have been written to solve this problem and the problem of the different dialects of the Pashto language. To overcome the Unicode problem, all similar characters with different Unicode are stored in a two dimensional matrix to check alternate Unicode for a character if a word is not found in the lexicon attached with the software. During the process when a word is found in the lexicon, its lexical category is read from the relevant POS column. Otherwise the module for Unicode and dialect is called to execute and find the relevant word in the lexicon. If still no word is matched in the lexicon, the word is marked as 'Unknown'. Below figure-III, shows a screen shot of the POS tagged list of words taken from the software.



PashtoWord	WordCat
افغان	Noun
چارواکي	Unknown
وايي	Unknown
چې	Adverb
دغه	proNoun
سند	Noun
به	particles
تر	particles
پنځو	Unknown
کلونو	Unknown
پورې	Adverb
اعتبار	Noun

Figure III: System Generated POS Sample

5.3. Rules embedded in a table

Rules described in section-4 of this paper are embedded in tabular format in order to get constituent of a sentence. A partial list of these rules IS given in Table-I.

Rules from table-I are read from top to bottom, and left to right, starting from titled column P1 of the table. Column under P1 are read, lexical categories for its relevent rows in the POS titled column are counted and read. The relevent entries in the POS columns are matched against the tagged text in figure-III. If the POS of the read text are matched with the lexical elements of the POS column for the P1 column, it means the combination of words against the tagged POS is a valid phrase or constituent. The process is rpeated for each column of the Table-1 untill S titled column is reached. This matching process is linear in order however the constituents occurs recursively in natural text. Therefore the text is read linearly however each found constituents is marked with its position in the sentence. Whenever another constitutent is found in the same position, these constitutnets are placed in order , as they were before. This way the problem of recursion is sloved and the shorcoming of CFG is overcomed.

Table 1: Syntax and Phrase Rules

S#	PoS	P1	P2	P4	P5	S
1	Noun	NPP	DAN	X2	Noun Phrase	S
2	*PP*					
3	Noun	NAP				
4	Adv		NA			
5	proN					
6	Noun		proN	X3		
7	Part	PA	PAVP		Verb Phrase	
8	Part					
9	Unkn	Unkn				
10	Unkn		ADP			
11	Adv	Adv				

5.4. How Recursion is tackled

As discussed earlier in section 1.3, recursion is the property of a constituent to contain another constituten. A sentence may have many levels of constituent within constituent. Opening up this layer by layer is not an easy process. Parsing and syntactic tree are mostly used to catch these layers. However, here in this work, the phenomena of recursion is tackled as an iterative process to simplify the complexities of recursion. As given in the algorithm below, during the process of matching of rules against the tagged words/sentence, the rules check only for constituents without looking for constituents within constituents. However a log is maintained of the location of each constituent with in the sentence or possessor constituent by keeping a count for each word. This log is used in the end of sentence completion to put all the constituents in an order.

5.5. Algorithm

The proposed algorithm to read the un-tagged corpus, tag the words of the corpus using a lexicon and make phrases based on rules embedded in a table is given below.

1. **Read text** from un-tagged Corpus word by word
2. **Search Dictionary** for each word read in step-1
3. **If Match Found** read its relevant grammatical category from the dictionary attached

4. **ElseIf no match found**, repeat step-3 with different Unicode for the same word (as some characters repeat in the Arabic Block with different Unicode)
5. **ElseIf No Match Found** repeat step-3 with alternate characters(form) for the same word(as in Pashto some words have the same meaning with different form in terms of spellings e.g. Pashto is written both as Pashto and Pakhto)
6. **Match Found**, then TAG each matched word with its grammatical category
7. **No Matched Found**, TAG the word as ‘Unknown’
8. **Read Rules** attached, from the matrix
9. **loop step 9-13** to fire rules from top to bottom and left to right
10. **Apply Rules** on the list of words tagged with grammatical category
11. **If Passed The Rule** by the set of words then,
12. **Look for old location,** write the new phrase and old phrase together and mark the position of new phrase
13. **Increment** to change set of words or rules
14. End

The main advantage of the proposed algorithm is its speed and ease of use. The table has been used to work like “if-then-else” clauses or rules. Coding and firing of rules is not only a complex and tedious job but also suffers from the recursion when dealing with analysis of sentence structure. Whereas rules embedded in a table like in the proposed algorithm makes the whole process recursion free, faster, and easy to build, change and improve rules.

6. Summary

The system is proposed to split input sentences into their constituents and phrases. A simple knowledge based system having all rules in a table is presented in this paper. The splitting process is quite encouraging with more than 90% results for any language. The algorithm used is mainly tested on Pashto language. The algorithm is designed to minimize the pitfall of CFG and overcome the complexity arises because of lengthy sentences.

References

[1] Paul R. Kroeger, *Analyzing Syntax: A Lexical-Functional Approach,* Cambridge University Press, New Yoark, 2004.

[2] Paul R. Kroeger, *Analyzing Grammar: An Introduction*, Cambridge University Press, New York, 2005.

[3] Andrew Radford, *Analyzing English Sentences a minimal approach*, Cambridge University Press, New York, 2009.

[4] Maggie Tallerman, Hodder Education, *Understanding Syntax (third Ed)*, UK, 2011.

[5] Bas Aarts, *Syntactic Gradient the nature of grammatical indeterminacy,* Oxford University Press, New Yoark, 2007.

[6] Model Selvam M, Natarajan. A M, and Thangarajan R, “Structural Parsing of Natural Language Text in Tamil Using Phrase Structure Hybrid Language”, International Journal of Computer and Information Engineering 2:4 2008.

[7] Steven Bird, Ewan Klein, and Edward Loper , *Natural Language Processing with Python,* O’Reilly Media, Inc, 2009.

[8] Santorini, Beatrice, and Anthony Kroch, “The syntax of natural language: An online introduction using the Trees program”, 2007. Available: http://www.ling.upenn.edu /~beatrice/syntax-textbook.

[9] Takao Doi, Eiichiro Sumitta , “Input Sentence Splitting and Translating: ATR Spoken Language Translation Research Laboratories Hikaridai, Kansai Science City, Kyoto, Japan.

[10] Francisco Oliveira , “Systematic Noun Phrase Chunking by Parsing Constraint Synchronous Grammar in application to Portuguese Chinese Machine Translation” , in proc. ICITA, 2009.

[11] G. Kempen and K. Harbusch, *“Generating Natural Word Order in a Semi-free word order Langauge: Treebank-based linearization preferances for German”,* in proc.. *of the Fifth CICLING, Seoul, Korea 2004*.

[12] S. T. Miorelli, “ED-CER: Extração do Sintagma Nominal. em Sentenças em Português”, Ph. D. Thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

[13] M. Perini, “A Gramática descritiva de Português”, São Paulo: Editora Ática, 1995.

[14] F. N. Q. M. C. Costa, “Deep Linguistic Processing of Portuguese Noun Phrases”, Master Thesis, Universidade de Lisboa, Portugal.

[15] Wojciech, S. and Thorsten, B. “Chunk Tagger - Statistical Recognition of Noun Phrases”, ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing, Saarbrucken, 1998.

[16] Craig Korpis, “Computing in Pashto An overview of a major language in Afghanistan and Pakistan*”,* in proc. *Multilingual Computing & Technology*.

[17] Andreas Kathol, Kristin Precoda, Dimitra Vergyri, Wen Wang, Susanne Riehemann , *“Speech Translation for Low-Resource Languages: The Case of Pashto”* in Proc. Interspeech, 2005.

[18] Sarmad Hussain, , Sana Gul, Afifah Waseem ,”Urdu Encoding and Collation Sequence for Localization” , Center for Research in Urdu Language Processing National University of Computer and Emerging Sciences.

[19] Tom Milo, *Some comments on the Arabic block in Unicode*, DecoType.

[20] Unicode Standard 6.1, Date, Retrieved 06/06/2012, Available:www.unicode.org/ versions/Unicode6.1.0/References.pdf

[21] Taylor Roberts, *Clitics and Agreement*, Phd thesis for Doctor of Philosophy in Linguistics at the Massachusetts Institute of Technology ,June 2000

[22] Hamza A Shierani, *Building bilingual Anglo-Pashto Proper noun Lexicon from the we*b, Thesis, Department of Computer Science, The University of Shefield

[23] Lewis, M. Paul, Pashto, Southern: A language of Pakistan**,** Date, Retrieved: 6/6/2012, Available: http://www.ethnologue.com/ show_language.asp?code=pbt

[24] Penzl, Herbert; Ismail Sloan, *A Grammar of Pashto a Descriptive Study of the Dialect of Kandahar, Afghanistan*. Ishi Press International. 2009.

[25] Jonathan Kew, *Notes on some Unicode Arabic characters: recommendations for usage* (Draft), Retrieved:6/6/2012, Available: http://scripts.sil.org/cms /scripts/render_download.php?format=file&media_id=arabicletterusagenotes&filename=ArabicLetterUsageNotes.pdf

[26] Tegey, Habibullah, Robson, Barbara, *A Reference Grammar of Pashto,* Washington, Center for Applied Linguistics, 1996.

[27] Zarghona Rekhten Zewar, Pakhtu Nahoa (Grammar), Da Sapye Da Pakhto Serono ao Prekhtia Zaye Markaz,Peshawar, 2003.

[28] Sidiq Ullah Reshteen, Da Pakhto Ishtiqaqona ao Tarkebona. Peshawar.

[29] Rahman Ali, Muhammad Amir, Mohammad AbidKhan “Developing Pashto Treebank”, in Proc: International Conference on Networks & Information Technology (ICCNIT), University of Peshawar, Pakistan, 2011.

Wikipedia is a Practical Alternative to the Web for measuring Co-occurrence based Word Association

Om P. Damani, Pankhil Chedda, Dipak Chaudhari
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{damani,pankhil,dipakc}@cse.iitb.ac.in

Abstract

While the World Wide Web is an attractive resource, few researchers can access or manage a Web-scale corpus. Instead they use search-hit counts as a substitute for direct measurements on a web corpus. In contrast, one can download a small high quality corpus like Wikipedia and carry out exact measurements. By extensive experiments with multiple word-association measures and several public datasets, we show that for exploring document level co-occurrence based word associations, despite being three orders of magnitude smaller in size, the Wikipedia is a reasonable alternative to a web corpus that can only be accessed using search engines.

Further, with Wikipedia, one can carry out measurements at a granularity finer than document scale. Instead of document level co-occurrence, one can consider a word-pair occurrence significant, only if the two words occur within a certain threshold distance of each-other. In general, such fine-grained information cannot be obtained from search engines. Our experiments show that the word level co-occurrence measures perform better than the document level measures. This indicates another practical advantage of the Wikipedia, or any other downloadable corpus, over a Web corpus which can only be accessed using search engines.

1. Introduction

The World Wide Web is an attractive resource for carrying out the NLP research. If one does not need the entire document contents and can just work with the frequency information of certain document types, then using the APIs provided by various search engines, one can use the Web as a corpus and need not collect and manage a corpus. An example area where one can take advantage of these APIs is the measurement of word association based on lexical co-occurrence [1].

The notion of *word association* is important for numerous NLP applications, like, information retrieval, question-answering, word sense disambiguation, optical character recognition, speech recognition, parsing, lexicography, text summarization, natural language generation, and machine translation. In [2] word association is motivated as *the basis for a statistical description of a variety of interesting linguistic phenomena*.

While the traditional co-occurrence based word association measures are formulated in terms of the word frequencies, it is straight-forward to reformulate them as working with the document frequencies. As an example, consider the the popular word association measure PMI [2]. It is defined as:

$$PMI(x,y) = log\frac{p(x,y)}{p(x)p(y)}$$

where $p(x)$ and $p(y)$ are unigram probabilities and $p(x,y)$ is bigram probabilities. These probabilities are obtained by dividing $f(x), f(y), f(x,y)$ by corpus-size in words. $f(x), f(y)$are the number of occurrences of words x and y in the corpus, i.e. the unigram frequencies of x and y, and $f(x,y)$ is the number of occurrences of the word-pair (x,y) in the corpus, i.e. the bigram frequency of (x,y).

To work with web corpus, one can simply replace word-frequencies with document frequencies, provided one knows the number of documents in the corpus, an information that is generally not available when using search engines. Instead, note that we need not work with probabilities. We can directly work with document frequencies since we are only interested in relative rankings of word-pairs and not in their absolute PMI values. Hence, as discussed later in Section 3.1, ignoring the corpus size does not affect any of our ranking based results. Therefore, we redefine PMI as:

$$PMI(x,y) = log\frac{n(x,y)}{n(x)n(y)}$$

where $n(x,y), n(x), n(y)$ are the counts of documents containing both words x and y, only x, and only y respectively. In the same way, we can redefine most other word-association measures in terms of document frequencies.

Though the large number of documents available on the Web are an attractive resource, Kilgarriff argues in [3] that ”Googleology is bad science”. One of the reasons cited there is the unreliability of the document counts obtained. After giving that warning, Kilgarriff accepts that ”With enormous data, you get better results”, and exhorts the readers to ”make resources on this scale available”.

Given that very few researchers can afford to access or manage a Web-scale corpus, only alternative they are left with is to use search-hit counts as a substitute for doing direct measurements on a Web-scale corpus. However, as argued in [3] and other places, for various performance and cost reasons, search-hit counts provided by search-engines are only crude approximations and poor substitute for actual Web statistics.

Given these limitations of working with Web, in this work, we argue that for applications like determining the word association, the quality of the data is much more important than the quantity. We find that using a Wikipedia dump containing 2.7 million documents gives better word association results than using the Yahoo and Bing search engines which indexed roughly 3.5 billion and 12 billion pages respectively[1] at the time of our experiments. Hence if a researcher cannot afford a Web-scale corpus, then it is better to work with a Wikipedia dump, than to use search-hit counts, at least for measuring word association.

Further, with Wikipedia, one can carry out measurements at a granularity finer than document scale. Instead of document level co-occurrence, one can consider a word-pair occurrence significant, only if the two words occur within a certain threshold distance of each-other. In general, such fine-grained information cannot be obtained from search engines. Our experiments show that the word level co-occurrence measures perform better than the document level measures. This indicates another practical advantage of the Wikipedia, or any other downloadable corpus, over a Web corpus which can only be accessed using search engines.

2. Related Work

The existing word association measures can be divided into three broad categories:

Frequency based measures rely on co-occurrence frequencies of both words in a corpus in addition to the individual unigram frequencies.

Distributional Similarity based measures based on Firth's ”You shall know a word by the company it keeps” [4], these measures characterize a word by the distribution of other words around it and compare two words for distributional similarity [5, 6, 7, 8].

Knowledge-based measures rely on knowledge-sources like thesauri, semantic networks, or taxonomies [9, 10, 11, 12, 13, 14].

In this work, our focus is on choice of resources for frequency based co-occurrence measures, and we do not discuss the details of the distributional similarity and knowledge based measures.

Chklovski and Pantel [15] have mined the web for fine-grained semantic relations such as similarity, strength, antonymy, enablement, and temporal happens-before relations between a pair of verbs. Mihalcea et. al. [16] measure the semantic similarity of short texts using several knowledge based and corpus based measures. They use the Microsoft paraphrase corpus [17], which was constructed by automatically collecting potential paraphrases from thousands of news sources on the Web over a period of 18 months. In [18], a new co-occurrence measure called Co-occurrence Significance Ratio is introduced and it is compared with a host of other measures using a Wikipedia corpus.

Although previously mentioned researchers have used the Web [15, 16] or the Wikipedia [18] for computing co-occurrence measures, to our knowledge no-one has performed a comparative study of the Web vs. the Wikipedia.

Information from the Wikipedia, such as its link structure [9], its concepts [11, 12], and its category trees [13] has earlier been used for knowledge-based word association measures. It is not surprising that the Wikipedia has been found useful for the knowledge-based measures. What is somewhat surprising is that the accurate measurements over Wikipedia give better results than the crude search hit counts from the Web for exploring even lexical co-occurrence based word associations, where one would expect that the much bigger corpus would always give better results due to the law of large numbers.

3. Wikipedia vs. Web

The advantage of the Web as a corpus is that it takes very little effort to work with it. However, while it is easy to replicate experiments on traditional corpora, the Web content keeps changing. In addition, the indexing and search strategies of the commercial search engines also change over time. Hence, it is hard to rerun the Web based experiments for reproducibility. Still, given the advantage of size and the ease of effort, it is worth exploring whether co-occurrence measures performs better

[1] Source: http://www.worldwidewebsize.com/

Table 1: Definition of Co-occurrence based word association measures.

Measure	Document Count	Word Count
Dice	$\frac{2n(x,y)}{n(x)+n(y)}$	$\frac{2\hat{f}(x,y)}{f(x)+f(y)}$
Jaccard	$\frac{n(x,y)}{n(x)+n(y)-n(x,y)}$	$\frac{\hat{f}(x,y)}{f(x)+f(y)-\hat{f}(x,y)}$
Ochiai	$\frac{n(x,y)}{\sqrt{n(x)n(y)}}$	$\frac{\hat{f}(x,y)}{\sqrt{f(x)f(y)}}$
PMI	$log\frac{n(x,y)}{n(x)n(y)}$	$log\frac{\hat{p}(x,y)}{p(x)p(y)}$
SCI	$\frac{n(x,y)}{n(x)\sqrt{n(y)}}$	$\frac{\hat{p}(x,y)}{p(x)\sqrt{p(y)}}$

$n(x,y)$	Total number of documents in the corpus having at-least one occurrence of (x,y)
$n(x),n(y)$	the number of documents in the corpus containing at least one occurrence of x and y respectively
$f(x),f(y)$	unigram frequencies of x,y in the corpus
$\hat{f}(x,y)$	span-constrained bigram frequency of x,y in the corpus
N	Total number of tokens in the corpus
$\hat{p}(x,y),p(x),p(y)$	$\hat{f}(x,y)/N,f(x)/N,f(y)/N$

or worse with the Web than with a much smaller corpus like Wikipedia.

3.1. Co-occurrence based Association Measures

To compare the performance of the Web and the Wikipedia, we experiment with six different co-occurrence based word association measures: Dice [19], Jaccard [20], Ochiai [21], Pointwise Mutual Information - PMI [2], and Semi-Conditional Information - SCI [22]. Their definitions are given in Table 1. Except SCI, all other measures are well-established and besides language processing, have been used in several domains like ecology, psychology, and medicine.

The word count based definitions are discussed later in Section 4. In this section, we are concerned with document count based definitions only. It is important to note that the word-count based versions count span-constrained bigram occurrences while the document based versions do not take span into account, since for the Web we do not have access to the span information.

Our results show that the Jaccard and the Dice have almost identical performance, since $[n(x,y) \ll n(x)]$ and $[n(x,y) \ll n(y)]$ for most word pairs. Hence we do not distinguish between these measures when presenting our results.

We have not experimented with other popular measures like the Log Likelihood Ratio - LLR [23], and the T-test, since their definitions require knowing the total number of documents in the corpus. Technically, knowledge of the corpus size in documents is needed even for our chosen measures, but we can ignore the corpus size by working with a scaled versions of these measures. For example, in the definition of PMI given earlier, technically all three terms $n(x,y),n(x)$ and $n(y)$ should be divided by the corpus size, but ignoring the corpus size does not affect any of our ranking based results, while the same cannot be said of the LLR and the T-Test.

As explained later, we evaluate a measure on a given dataset by the Spearman rank correlation coefficient between the word-associations produced by the measure and the gold-standard ratings for the dataset. The Spearman rank correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. Since any monotonic transformation of the word association scores produced by a measure leaves the rankings unchanged, the modified scores obtained by ignoring the corpus size leaves the rankings unchanged.

3.2. Datasets and Resources

The two major types of word associations discussed in literature are *free association* and *semantic relatedness*.

Free association refers to the first response given by a subject on being given a stimulus word [24]. The standard methodology for collecting *free association* data is explained at [24]: ”Native speakers are presented with stimulus words and are asked to write down the first word that comes to mind for each stimulus. The degree of free association between a stimulus (S) and response (R) is then quantified by the percentage of test subjects who produced R when presented with S.”

We use five different publicly available datasets for measuring free association: Kent [26], Minnesota [27], White-Abrams [28], Goldfarb-Halpern [25], and

Table 2: Characteristics of data sets used. 'Respondents' is the number of individuals who were asked to respond to a given set of stimulus words. 'Word Pairs' is the total number of unique (stimulus,response) pairs generated. 'Filtered Word Pairs' is the size of the subset of the corresponding dataset used in our experiments.

Aspect	Data Set	No. of Respondents	No. of Word-Pairs	No. of Filtered Word-Pairs
Semantic relatedness	wordsim353 [14]	16	353	351
Free-Association	Essli [24]	100	272	272
	Goldfarb-Halpern [25]	316	410	384
	Kent [26]	1,000	14,576	14,086
	Minnesota [27]	1,007	10,447	9,649
	White-Abrams [28]	440	745	652

Essli [24].

The *semantic relatedness* encompasses relations like synonymy, meronymy, antonymy, and functional association [29]. We use the publicly available Wordsim [14] dataset to measure the semantic relatedness. The word association scores in this dataset are the average of the values on a scale from 0 to 10 given by the respondents when they were asked to estimate the relatedness of the words in a given pair.

One could say that free association datasets are asymmetric datasets where one stimulus words occur with multiple response words. In contrast, semantic relatedness datasets are symmetric in that both words in a pair have the same status.

Many of these datasets contain multi-word expressions. We removed word-pairs containing multiword expressions. For data sets with more than 10,000 word-pairs, we filtered out pairs that contain stop words listed in [30]. The details of the dataset after filtering is given in Table 2.

3.3. Corpus

We use the a Wikipedia dump with 2.7 million documents and of size 1.24 Gigawords. We used Lucene[2] APIs to obtain various statistics from the corpus. No function-word removal, lemmatization or any other preprocessing was performed on the corpus by us other than whatever preprocessing is done by default by Lucene.

For the web search, we use Yahoo and Bing search services. For Yahoo, we use BOSS API[3]. For Bing, we issue simple search requests and parse the response pages to obtain the hit count. In both cases, we use boolean conjunctive queries to get the count of documents containing both words in the pair. We could not use Google Search since it allows only 1000 queries a day.

3.4. Evaluation Methodology

For the word-pairs in each dataset, each measure under consideration produces a ranked list of the word association scores. We also have the gold-standard human judgment ranking available for each dataset. We follow the standard methodology of evaluating a word-association measure on a given dataset by the Spearman rank correlation coefficient between the word-associations produced by the measure and the gold-standard ratings for the dataset.

3.5. Results

The purpose of our experiments is not to compare the word association measures but to compare a downloadable Wikipedia corpus with a Web corpus that can be accessed using only search-engine interfaces. That is, we wish to find out how the performance of a given measure on a given dataset changes as we move from the Web to the Wikipedia. The results of our experiments are shown in Table 3. For completeness, we also show the results from [13], the only known document level co-occurrence result from the literature for these datasets.

We can see that for all measures and for five out of the six datasets, the performance always improves as we move from the Web to the Wikipedia. Even for the sixth dataset Goldfarb-Halpern, the Web does not perform better than the Wikipedia, except when the correlations are close to zero, i.e. when things are pretty random.

4. Further Improvement

Another advantage of using the Wikipedia is that unlike the Web, in case of the Wikipedia, a researcher can download the entire Wikipedia and can carry out measurements at a granularity finer than the document scale. In fact, traditionally co-occurrence based measures are defined in terms of the span constrained word-pair oc-

[2] http://lucene.apache.org/

[3] http://developer.yahoo.com/search/boss/

Table 3: Performance of various document based co-occurrence measures while using Wikipedia and the Web. For each measure and each dataset, the *best performing* version has been *highlighted.* Results for the Web-Google [13] are available only for the *wordsim353* dataset. Also, note that we have filtered out the multi-word expressions from each dataset. Hence, for example, we work with only 351 of the 353 pairs in the *w*ordsim353 dataset.

Measure	Corpus	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	words353 (351)	Esslli (272)
PMI	Wikipedia	**0.20**	**0.12**	**0.19**	**0.18**	**0.58**	**0.33**
	Web-Yahoo	0.18	0.04	0.11	0.18	0.35	0.17
	Web-Bing	0.08	0.02	0.07	0.18	0.20	0.10
SCI	Wikipedia	**0.36**	**0.24**	**0.28**	**0.17**	**0.48**	**0.50**
	Web-Yahoo	0.26	0.18	0.14	0.07	0.28	0.27
	Web-Bing	0.21	0.10	0.15	**0.17**	0.23	0.26
Ochiai	Wikipedia	**0.31**	**0.20**	**0.20**	-0.02	**0.41**	**0.31**
	Web-Yahoo	0.24	0.18	0.11	-0.03	0.18	0.14
	Web-Bing	0.22	0.12	0.12	**0.04**	0.29	0.11
Jaccard/ Dice	Wikipedia	**0.31**	**0.20**	**0.18**	-0.01	**0.36**	**0.21**
	Web-Yahoo	0.24	0.14	0.11	-0.01	0.14	0.09
	Web-Bing	0.22	0.12	0.09	**0.04**	0.25	0.05
	Web-Google [13]	-	-	-	-	0.18	-

currences[4]. By span we mean the inter word distance. When querying the Web, we get the counts of documents containing the word pair regardless of the distance between them in the documents. By span constrained occurrence we mean that a word-pair occurrence is counted only if the words occur close enough, that is if their span is less than a given threshold. That is, with every measure, a span-threshold parameter is attached.

4.1. Span Constrained Word Count Performance

We compare the performance of word based and document based version of each measure as given in Table 1. Our methodology of computing ranked correlation for a measure on a dataset remains the same (as described in Section 3.4). Only difference is that the word based version of each measure has span threshold as a parameter.

We follow the standard methodology of evaluating parametrized measures by cross validation. Each dataset is divided into five random partitions, four of which are used for training and one for testings. The span threshold is varied between 5 and 50 words for each measure and the span value that performs best on four training folds is used for the remaining one testing fold. The performance of a measure on a dataset is its average Spearman rank correlation over 5 runs with 5 different test folds.

4.2. Comparison

The comparison of document based and word based measures are shown in Table 4. From the results, we can see that with Wikipedia, further performance gain is obtained by moving from document counts to span-constrained word counts. We have four measures and six datasets for a total of twenty-four combinations. For eighteen out of the twenty-four combinations, such a performance gain is observed.

As an aside, it is interesting to note that in Tables 3 and 4, regardless of the corpus, performance of the Dice measure is virtually identical to that of the Ochiai measure on the Kent and Minnesota - two largest datasets. This is interesting because the Dice is the harmonic mean while the Ochiai is the geometric mean of the Conditional Probabilities $\frac{n(x,y)}{n(x)}$ and $\frac{n(x,y)}{n(y)}$.

5. Conclusions

By performing extensive experiments with various measures and multiple datasets of varying size, we demonstrate that despite being three orders of magnitude smaller in size, the Wikipedia is a reasonable alternative to the Web for measuring the document level co-occurrence based word association.

Another practical advantage of Wikipedia compared to web is that most researchers can exploit the span

[4] Note how Church and Hanks define joint probability in their seminal paper [2] that introduced PMI: *Joint probabilities, $P(x, y)$, are estimated by counting the number of times that x is followed by y in a window of w words, $f_w(x, y)$, and normalizing by N.*

Table 4: Performance comparison of the word based and document based version of each measure on Wikipedia. For each measure and each dataset, the *better performing* version has been *highlighted*. All standard deviations across 5 cross-validation runs for Kent and Minnesota are between 0.01 and 0.02, for White-Abrams were between 0.05 and 0.07, for Goldfarb-Halpern between 0.05 and 0.14, for Wordsim were between 0.02 and 0.11, and for Esslli were between .09 and .17. Note that the word-count based versions count span-constrained bigram occurrences while the document based versions do not take the span information into account.

Measure	Kent (14,086)	Minnesota (9,649)	White-Abrams (652)	Goldfarb-Halpern (384)	words351 (351)	Esslli (272)
PMI-doc	0.20	0.12	0.19	**0.18**	0.58	**0.33**
PMI-word	**0.36**	**0.26**	**0.22**	0.11	**0.69**	0.32
SCI-doc	0.36	0.24	**0.28**	**0.17**	**0.48**	**0.50**
SCI-word	**0.38**	**0.27**	0.23	0.06	0.37	0.44
Ochiai-doc	0.31	0.20	0.20	-0.02	0.41	0.31
Ochiai-word	**0.43**	**0.31**	**0.29**	**0.08**	**0.62**	**0.44**
Jaccard/Dice-doc	0.31	0.20	0.18	-0.01	0.36	0.21
Jaccard/Dice-word	**0.43**	**0.32**	**0.21**	**0.09**	**0.59**	**0.35**

and the word count information with Wikipedia but not with the web. Our experiments demonstrate the utility of these information in improving the word association performance. In the current work, we have compared span-constrained word-count based versions with non-span-constrained document-count based versions, since for the Web we do not have access to the span information. In future, we plan to experiment with span-constrained document-count based versions for Wikipedia.

References

[1] P. Pecina and P. Schlesinger, "Combining association measures for collocation extraction," in *Association for Computational Linguistics*, 2006.

[2] K. W. Church and P. Hanks, "Word association norms, mutual information and lexicography," in *Association for Computational Linguistics*, pp. 76–83, 1989.

[3] A. Kilgarriff, "Googleology is bad science," *Computational Linguistics*, vol. 33, no. 1, pp. 147–151, 2007.

[4] J. R. Firth, "A synopsis of linguistics theory," *Studies in Linguistic Analysis*, pp. 1930–1955, 1957.

[5] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and wordnet-based approaches," in *NAAssociation for Computational Linguistics-HLT*, 2009.

[6] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *WWW*, pp. 757–766, 2007.

[7] H.-H. Chen, M.-S. Lin, and Y.-C. Wei, "Novel association measures using web search with double checking," in *Association for Computational Linguistics*, 2006.

[8] T. Wandmacher, E. Ovchinnikova, and T. Alexandrov, "Does latent semantic analysis reflect human associations?," in *European Summer School in Logic, Language and Information (ESSLLI'08)*, 2008.

[9] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Association for Computational Linguistics*, 2008.

[10] T. Hughes and D. Ramage, "Lexical semantic relatedness with random graph walks," in *Conference on Empirical Methods on Natural Language Processing*, 2007.

[11] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *International Joint Conference on Artificial Intelligence*, 2007.

[12] E. Yeh, D. Ramage, C. Manning, E. Agirre, and A. Soroa, "Wikiwalk: Random walks on wikipedia for semantic relatedness," in *Association for Computational Linguistics workshop "TextGraphs-4: Graph-based Methods for Natural Language Processing"*, 2009.

[13] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *Conference on Artificial Intelligence*, pp. 1419–1424, 2006.

[14] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: the concept revisited," *ACM Transactions on Information Systems*, vol. 20, no. 1, pp. 116–131, 2002.

[15] T. Chklovski and P. Pantel, "Verbocean: Mining the web for fine-grained semantic verb relations," in *Conference on Empirical Methods on Natural Language Processing*, pp. 33–40, 2004.

[16] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Conference on Artificial Intelligence*, 2006.

[17] W. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *20th International Conference on Computational Linguistics*, 2004.

[18] D. L. Chaudhari, O. P. Damani, and S. Laxman, "Lexical co-occurrence, statistical significance, and word association," in *Conference on Empirical Methods on Natural Language Processing*, 2011.

[19] L. R. Dice, "Measures of the amount of ecological association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[20] P. Jaccard, "The distribution of the flora of the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, 1912.

[21] A. Ochiai, "Zoogeografical studies on the soleoid fishes found in japan and its neighbouring regions-ii," *Bulletin of the Japanese Society of Scientific Fisheries*, vol. 22, 1957.

[22] J. Washtell and K. Markert, "A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations," in *Conference on Empirical Methods on Natural Language Processing*, pp. 628–637, 2009.

[23] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[24] ESSLLI, "*Free association* task at lexical semantics workshop esslli 2008." `http://wordspace.collocations.de/doku.php/workshop:esslli:task`, 2008.

[25] R. Goldfarb and H. Halpern, "Word association responses in normal adult subjects," *Journal of Psycholinguistic Research*, vol. 13, no. 1, pp. 37–55, 1984.

[26] G. Kent and A. Rosanoff, "A study of association in insanity," *American Journal of Insanity*, pp. 317–390, 1910.

[27] W. Russell and J. Jenkins, "The complete minnesota norms for responses to 100 words from the kent-rosanoff word association test," tech. rep., Office of Naval Research and University of Minnesota, 1954.

[28] K. K. White and L. Abrams, "Free associations and dominance ratings of homophones for young and older adults," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 408–420, 2004.

[29] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguists*, vol. 32, no. 1, pp. 13–47, 2006.

[30] StopWordList, "`http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`," *The Information Retrieval Group, University of Glasgow*, 2010. Accessed: November 15, 2010.

The Pitch Contour of Declarative Sentences in Urdu Language

Farhat Jeen[1], Sarmad Hussain[2]
Department of English, The Islamia University of Bahawalpur, Bahawalnagar Campus, Pakistan[1]. Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science, University of Engineering and Technology, Lahore, Pakistan[2].
farhat2iub@gmail.com[1]. sarmad.hussain@kics.edu.pk[2]

Abstract

This paper investigates the intonation contours in Urdu declarative sentences of Bahawalnagar district in Pakistan. For this purpose, three respondents are selected from the district. All the respondents use Urdu as their mother tongue while their parents are Punjabi speakers. The data is collected in the form of isolated sentences. Three repetitions each for intransitive, transitive and ditransitive are recorded for analysis. The results indicate that there is some variation in the intonation contours of declarative sentences but L H L- L% is the most frequently used pattern and does not change even as the sentence structure is modified due to the change in transitivity of the verbs.

1. Introduction

Intonation is an essential feature of any language, caused by changes in the pitch contour, and used by speakers to indicate a variety of linguistic phenomena. Speakers employ pitch to indicate whether they are stating something or asking a question. They also use pitch to indicate their emotional state, i.e. if they are angry, happy, surprised, etc.

These factors of an utterance are altered by simply changing its pitch contour. There has been very limited work on determining the intonation patterns for Urdu language. This work is an effort to understand Urdu intonation. This study aims to explore the range of pitch patterns available for the production of declarative sentences in Urdu and is part of a larger work which focuses on implications of first language (L1) intonation on English language learning (as L2), latter not discussed in the current paper.

2.Literature Review

Intonation may be defined as “the structured variation in pitch which is not determined by lexical distinctions.” [7, p. 253] Intonation is one of the widely discussed aspects of suprasegmental phonology. Intonation is generated by the rate of vibration of the vocal cords [1] and perceived as the pitch contours [2, 3, 4]. Intonation is described as the selection of pitch ranges available to the speakers [5] and includes:

> (a) categories of pitch 'peaks' and 'valleys' as well as their combinations at each sentence stress position (i.e., the last content word of the sentence), (b) types of pitch category concatenation, and (c) pitch of sentence fractions occurring before the first sentence stress. (p. 118)

These pitch ranges may be high or low and by exploiting these options, speakers convey linguistic and other information. In fact, many researchers claim that it is not easily possible for people to speak continuously in a monotone [1, 3, 6]. Intonation occupies a central position in speech production and perception.

The pitch contour of an utterance is usually denoted as a tone [3]. The pitch contour in an utterance may be rising, falling, falling-rising, rising-falling, level, high key and take-off [1, 2, 3, 4, 5]. Each of these contours performs different functions and is selected by the speakers on the basis of those functions.

Those languages of the world which have a fixed sentence order use pitch movements to convey emotions [5]. Generally pitch carries emotional information and the change in pitch does not change the meaning of an utterance. But there are certain languages of the world in which change in pitch contours results in change of meaning [4]. These languages are called tonal languages e.g. Punjabi [8] and Chinese. English is not a tonal language therefore pitch in English language is a superimposed feature which adds to the richness of meaning. For example, a simple declarative sentence may function as interrogative or exclamatory only by changing the tone of the utterance.

The pitch movement in English language has been studied with the help of Pierrehumbert’s auto-segmental metrical model of intonation. In [10], this model has been described comprehensively. She describes it as a collection of high (H) and low (L) target pitch targets. These targets are used in pre-nuclear and nuclear syllables, phrase accents (tone after the nuclear syllable) and at tone boundaries (at the end of a phrase). There are a maximum of six tones available at pitch accent level: H*, L*, L+H*, L*+H, H !H*, and !H*. Here the star indicates tonal targets falling on stressed syllables. Similarly there are two possible phrase accents available i.e. L- and H-. According to this model, English uses two tone boundaries symbolised as H% and L%.

Pierrehumbert's model was later modified into Tone and Break Indices (ToBI) model [6]. This model has been divided into two entities: phrasal tones and pitch accents. The latter are further divided into phrase accents and boundary tones.

It further describes English intonation with the help of boundary index ranging from one to four. Closely joined words in a tone unit are given the index of one and the boundary index at the end of a sentence is always four [6]. This model has also been used to study Japanese intonation patterns [9]. This paper also attempts to study the use of intonation contours in Urdu declarative sentences by using ToBI.

3. Methodology

3.1. Population

Three respondents have been selected for this study from district Bahawalnagar. All the respondents are Urdu L1 speakers whose parents speak Punjabi as their mother tongue. All the respondents and their parents have spent their initial years of language learning in Bahawalnagar, so we can reasonably assume that they all belong to the same linguistic group.

3.2. Data collection

The data has been collected from respondents in the laboratory atmosphere. As the stress patterns and the syllable structure of the utterances are controlled in order to get valid results, it has been impossible to use naturalistic speech for this study.

This work is limited to the study of intonation patterns in declarative sentences in Urdu language. The data set is designed in the form of sentences. Each sentence comprises words carrying only mono or disyllable words. However, the final verbs in the ditransitive sentences are of three syllables. The data set is designed on the basis of sentence types. The types studied in this research are SV, SOV, and SOVO^2.

For each sentence type, three sentences have been devised and each speaker has pronounced three repetitions of each sentence. The data set is given in the table below.

Table 3.1. Data set for the analysis of declarative sentences

1	S V	ناز نے کھایا	Naz ate.
		ناز نے مارا	Naz hit.
		ناز نے گایا	Naz sang.
2	S O V	ناز نے تیر مارا	Naz hit with an arrow.
		نازنے کھانا کھایا	Naz ate food.
		ناز نے گانا گایا	Naz sang a song.
3	SOVO^2	نازنے زین کو کھانا کھلوایا	Naz made Zain to eat food.
		ناز نے زین کو تیر سے مروایا	Naz got Zain killed with an arrow.
		ناز نے زین کو گانا گوایا	Naz made Zain to sing a song.

The stress is controlled and all the utterances have been produced with stress on the first word (ناز, Naz).

In order to compare the results, an alternative way of structuring the Urdu sentences i.e. SVO has also been studied. The sentences with alternative word order have been pronounced by the respondents and their recordings analysed in order to examine if the verb/object replacement affects the intonation contour of a declarative statement in Urdu language. Here again the stress is placed on the first word of the sentence.

The data has been recorded in wav. format. A total of 486 recordings (3 participants * 3 repetitions) have been analysed using PRAAT software of speech analysis. The sound files have been manipulated to generate pitch contours. Each sentence has been separately analysed and the final contour generated by eliminating the redundant pitch points. The final speech pictures have been saved in an excel file. Then the contours have been determined and listed. The lists have been used to produce pie charts and percentages of usage of various pitch contours.

4. Results

4.1. SV

The results indicate that the most dominantly used intonation pattern in Urdu intransitive declarative sentences is L H L- L%. 67% utterances in the data have used this pattern. However, two more contours, i.e. L L- L% and L H H- L%, have been traced in the recordings. Nevertheless, their frequency of usage is lesser than the L H L- L% contour. The summary of the use of these contours along with the percentage of their respective usage is given in table 4.1.

Table 4.1. Summary of pitch contours used in intransitive declarative sentences in Urdu

Pitch Contour	Percentage
L H L- L%	67
L L- L%	26
L H H- L%	7

4.2. SOV

The results indicate that in transitive sentences, the frequently used pattern is L H L- L% which has 52% occurrences in the data. The remaining 48% of the data is divided into various contours such as L H H- L%, L L- L%, L L L- L%, H L- L%, L H H- H%, H L L- L%, L H L H L- L%. The summary of all the contours along with their percentage of usage is given in table 4.2.

Table 4.2. Summary of pitch contours used in transitive declarative sentences in Urdu

Pitch Contour	Percentage
L H H- L%	4
L H L- L%	52
L L- L%	11
L L L- L%	4
H L- L%	7
L H H- H%	3
H L L- L%	15
L H L H L- L%	4

4.3. SOVO^2

Table 4.3 indicates that in ditransitive sentences too, L H L- L% contour covers 74% of the data. L H L L- L% contour comprises 15% of the data. With the addition of a low pitch accent, this contour may be merely an extension of the previously mentioned pattern. The other two contours, i.e. H L H- L% and L H L H- L% also share certain features with the previously mentioned contours as all of them share the low boundary tone (L%). The summary of the contours is presented in table 4.3.

Table 4.3. Summary of pitch contours used in ditransitive declarative sentences

Pitch Contour	Percentage
H L H- L%	4
L H L L- L%	7
L H L- L%	74
L H L H- L%	15

4.4. SVO

There is no definite sentence structure in Urdu language and SVO is perfectly acceptable and intelligible sentence structure for Urdu, albeit it is a feature of spoken language. This study sets out to verify if the alternative sentence structure affects the pitch contour of Urdu declarative sentences.

The results indicate that the change in sentence structure does not affect its pitch pattern. Figure 4.4 indicates that 66% data comprises L H L- L% contour. However, a wide variety of pitch patterns have been used in this context. H L- L% contour covers 11% of the data and L H L L- L% covers 7%. The remaining contours, L H L H- L%, L H L L- H%, L L- L%, L H H- L%, encompass only 4% of the data each. The summary of all the contours used in this context is given in table 4.4.

Table 4.4. Summary of pitch contours used in SVO declarative sentences

Pitch Contour	Percentage
L H L- L%	88
L H L H- H%	4
L H L L- L%	7
L H L L- H%	4
H L- L%	11
L L- L%	4
L H H- L%	4

5. Discussion

The above mentioned results indicate that the predominant pitch pattern used for Urdu declarative sentences is L H L- L% with the overall usage of 65%. Moreover, the difference of transitivity does not influence the use of pitch contour in declarative sentences in Urdu. The intransitive, transitive as well as the ditransitive sentences use the same L H L- L% contour.

The study of the intonation patterns of declarative sentences of alternative sentence structure also corroborates the above mentioned findings. Although a variety of contours have been used in these utterances, the dominant pattern is L H L- L%.

Yet there are other contours which use or drop one or two additional phrase accent contours but may be interpreted as extensions of the basic L H L- L% pattern. These extension contours are L H L L- L%, L L- L%, H L- L%, H L L- L%, L H L H L- L%. A shared contour among all these patterns is the use of low tone boundaries. So we can claim that there may be differences in the use of low and high phrase accents but the predominant contour for the tone boundaries is low. The results indicate that only 3% of the utterances made use of high boundary tones.

6. Conclusion

This work has aimed to study and determine the pitch contour used by the Urdu sentence to pronounce declarative utterances. It has proved that the major pitch contour to produce a declarative utterance is L H L- L%. It has also verified that transitivity and reverse sentence order do not affect the use of pitch contour in Urdu.

This study has pedagogical implications as it may help Urdu language teachers to teach the pitch contour accompanying declarative sentences in Urdu. It may also help them teach the variety of pitch contours available to pronounce a declarative utterance.

Furthermore, this work may be useful for the development of Urdu speech synthesis systems. The findings of this study may be used to improve the quality of those systems by identifying the principal contour used for declarative sentences in Urdu and by incorporating the variations in pitch contour that are possible in the production of Urdu declarative utterances.

However, there are many unexplored aspects of pitch contours in Urdu. This paper is limited to the study of declarative sentences only but other sentence structures such as interrogative exclamatory, imperative etc. have not been studied yet. Similarly, the change in pitch contour under the influence of different emotions also needs to be studied. The influence of complex sentence structure is yet another field that needs to be studied.

References

[1] J. Sethi and P. V. Dhamija, *A Course in Phonetics and Spoken English*, PHI Learning Pvt. Ltd., 2004.

[2] A. Cruttenden, *Gimson's pronunciation of English*, Hodder Education, 2008.

[3] P. Roach, *English phonetics and phonology*, Cambridge University Press, 1983.

[4] P. Skandera and P. Burleigh, *A Manual of English Phonetics and Phonology: Twelve Lessons with an Integrated Course in Phonetic Transcription*, Gunter Narr Verlag, 2005.

[5] P. Birjandi and M. Salmani-Nodoushan, *An introduction to phonetics*, Zabankadeh, Tehran, 2005.

[6] P. Ladefoged, *Vowels and consonants: An introduction to the sounds of languages* (Vol. 1): Wiley-Blackwell, 2001.

[7] C. Gussenhoven, Intonation. In P. V. De Lacy (Ed.), *The Cambridge handbook of phonology*, Cambridge University Press, 2007.

[8] T. K. Bhattia, World Englishes in global advertising. In B. B. Kachru, Y. Kachru & C. Nelson (Eds.), *The handbook of world Englishes* (Vol. 48), Wiley-Blackwell, 2006.

[9] J. J. Venditti, The J_ToBI model of Japanese intonation. *Prosodic typology: The phonology of intonation and phrasing*, pp. 172-200, 2005.

[10] J. Pierrehumbert, “Phonological and phonetic representation”, *Journal of Phonetics*, pp. 375-394, 1990.

Comparing Talks, Realities and Concerns over the Climate Change: Comparing Texts with Numerical and Categorical Data

Yasir Mehmood, Timo Honkela
AaltoUniversity, School of Science
yasir.mehmood01@estudiant.upf.edu, timo.honkela@aalto.fi

Abstract

The conference on the climate change, UNFCCC 2010, took place from 29 November to 10 December 2010 in Cancun, Mexico. This paper presents an analysis of the opening speeches by various countries at the conference, combined with the statistics of the countries regarding their socioeconomic indicators and memberships of different climate treaties. A central objective is to compare different sources of the information that reflect the underlying complex system where there are obvious and less obvious relationships between the rhetoric and some aspects of reality. At the level of argumentation, we are interested in the occurrence of topics related to the climate change, i.e., whether some topics are mentioned or avoided in the speeches. The recognition of the topics is based on a semi-automatic term selection process that provides the input for the subsequent steps of the analysis. The data preparation process includes optical character recognition, machine translation and approximate string matching. We assume that the collection of terms serves as a relevant set of features that reflect the content of the speeches. These text-based features are then compared with the country statistics. The basic hypothesis is that there is a detectable but complex relationship between the content of the speeches and known facts. The most important contributions in this paper are the formulation of the basic questions and the overall hypothesis, an analysis of the relationships between the countries as well as between the topics and indicators, and the qualitative analysis of the results.

1. Introduction

The concerns over the climate change are growing particularly with the high emissions of greenhouse gases (GHG) that are raising the temperature of the planet. The United Nations Framework Convention on Climate Change (UNFCCC) is an international convention for the climate change. It came into force in 1994 with a mission of forming consensus among 30 developed countries to reduce GHG emissions into the atmosphere. However, since UNFCCC is a convention, its members do not have an obligation to guarantee a reduction in the GHG emissions. The legally binding framework of UNFCCC is Kyoto Protocol, which aims to reduce GHG emission by 5% from 1990s (see UNFCCC website http://unfccc.int/). The supreme body of Kyoto Protocol, responsible for the implementation of its aims, is called Meeting of the Parties to the Kyoto Protocol (CMP). Similarly the supreme body of the UNFCCC is called Conference of Parties (COP). In the opening session of both the COP and CMP, in UNFCCC 2010, all the member countries and organizations delivered their speeches to reflect their views on the climate change. In our analysis, the text of the speeches is the primary source of data. A secondary data consisting of statistics of the countries and their membership in climate treaties has been used to provide context to the primary data. This is further detailed later in this section.

1.1. Main Objectives

The analysis that we have performed is of two types. The aim of the first type of analysis is to visualize the data points on a two-dimensional grid. The data points are prepared from two data sources namely the primary and secondary data. The primary data is the speech text represented by a set of terms that include both unigrams and bigrams. The procedure for selecting the elements (or terms) is explained in section 3. The secondary data includes contextual information that is divided into two data sets namely *Country-wide Statistics* and *Membership of Climate treaties*. The *Country-wide Statistics* consist of quantitative information such as the GDP of a country, per capita income, mortality rate and various related facts. The *Membership of Climate treaties* includes information regarding a country's membership to treaties including Kyoto Protocol, Biodiversity, Desertification, etc. In our first analysis, the primary and secondary data are combined by a scheme explained in Section 3.3, and the combined effect on the analysis is visualized by the Self-Organizing Map (SOM) [1]. The SOM is a widely used technique for analyzing and visualizing high-dimensional data. The SOM algorithm, positions each data vector, in the high-dimension, to a two-dimensional area without sacrificing the orientation of the vector in the original space. This results in a two-dimensional representation of complex data points where, the data points that are close to each other in the original space tend to remain close to each other in the two-dimensional space as well. This aids the visual perception of the data and hence one can identify the groups of similar data points. In the context of our analysis, the data points are the members of the climate change conference. Thus, visualizing a map of the (combined) data will unveil the similarity between/among different countries based on the contents of speeches, when contextualized with the respective country statistics. In addition to visualizing high-dimensional data, the SOM enables visualizing the distribution of each dimension - and therefore it can help in understanding the contribution of each data feature (textual or statistical) over the map. This is important, since the similarity in the speech documents is greatly influenced by the contribution of text [2] or statistics. Therefore, in the second part of our analysis, we select various features from the (combined) data and observe their distribution across the SOM map that was created in the first part of analysis. The SOM is one possible choice for conducting this kind of analysis but there are currently many other methods that could also be considered [3, 4]. However, the choice of methodology is not central aspect in our research, it is rather the formulation of the overall question and analysis architecture to increase understanding of the complex phenomenon.

1.2. Related Work

The topic of this paper is multifaceted and therefore providing a conclusive review of related work is challenging. Methodologically important aspect is the combination of text and data mining. There are a large number of scientific articles concerning text mining and data mining separately in domains like business and finance [5, 6] or biomedicine [7, 8]. The combination of text and data mining has been used much more infrequently. Existing work concentrates in the area of business and finance [9, 10]. The content of the negotiations have been widely reported in media and also analyzed in scientific articles in the areas of law [11] and environmental research [12].

This paper is organized so that Section 2 explains the data sources and its acquisition process. Section 3 details the data preparation process. Section 4 explains the results of experiments, and finally, Section 5 concludes the paper. In this article, the terms *speech*, *document* and *talk* refer to the same concept; similarly, the words *member* and *country* have been used interchangeably.

2. Data Acquisition

This section details the acquisition of primary and secondary data.

2.1. Primary Data

The primary data refers to the talks presented at UNFCCC 2010. There were 163 talks given by the heads of states and governments. We have managed to acquire 143 of them from the UNFCCC website since some of the talks are not available or easily translatable. The talks can be downloaded as PDF files and we have used Google's built-in OCR support for extracting the text from the PDF files. This process does not guarantee full accuracy of data acquisition since several *words* are distorted. On the other hand, we have avoided manual work to correct all the text documents considering it *a)* an interesting case of text mining, and *b)* time consuming human effort which can be addressed by smart methods. The details of dealing with this problem are explained in section 3. There are some talks that were presented in the national languages. We have used Google's translation facility in order to convert those in English. However, these have not been considered while creating the feature set. The textual data, acquired as described previously, is the primary data source for our experimentation because it has been given more weight in the analysis. However, the secondary source of data is explained in the next subsection.

2.2. Secondary Data

The secondary data, in our experimentation, refers to the contextual data. It is of two types: *Country-wide Statistics* and *Membership of Climate Treaties*, described earlier in Section 1. The contextual data has been acquired from CIA Factbook (https://www.cia.gov/library/publications/the-world-factbook/). The data preparation and representation is explained in Section 3.1. The contextual data has been given less weight in the analysis in order to amplify the impact of speech on the results of analysis.

3. Data Preparation and Representation

This section details the steps of data preparation as well as its appropriate representation for the analysis. Section 3.1 explains the preparation of primary data, Section 3.2 explains the preparation of secondary data and Section 3.3 outlines the final representation of data, which is necessary for the analysis.

3.1. Term Extraction, Validation and Weighting

In the primary data, the English text of the speeches is available for 106 countries out of 143. This is sufficient to create a collection of word features for the entire data set, without including the non-English speeches. The primary reason for excluding non-English text is that the Google translation facility requires human effort to understand the translation better [13] and therefore does not guarantee the exact translation. Moreover, if the individual terms, in the word features, are not chosen carefully, it may increase the feature space, giving rise to *sparsity* in the document vectors [14]. This has a great potential of affecting the quality of results by adding numerous features that are not meaningful to the analysis, and thereby adding to the computational complexity [15]. Feature extraction techniques such as principal component analysis (PCA) [16], Random Projection [17] and many others [18], can help overcoming this problem; however, these techniques represent the original feature space into a new space. This is not suitable for our analysis because we want to preserve the original (and meaningful) features in order to visualize the distribution of those features over the analysis map (as highlighted in the Section 1). On the other hand, a suitable feature selection technique can help alleviating this problem. Liu et al. [14] outlines various methods for reducing the original high-dimensional space by selecting useful features. In our work, we have employed *Entropy* to select the most informational features that comprise the set of terms (see also [19]). These features are essentially unigrams and bigrams having high entropy values. Considering each term vector a random variable *T*, the entropy *H* is calculated as follows:

$$H(T) = \sum_{t \in T} p(t)\log_2 p(t) \tag{1}$$

The informational terms (unigram and bigrams) are selected by setting a threshold on the *H(T)*. Both the unigrams and bigrams obtained by this procedure are a little more than 1100 each. We have selected 400 unigrams with a little manual work. However, for the bigrams we have marginalized a lot of them based on their occurrences in a reference dataset (Europarl: http://www.statmt.org/europarl). This gives us nearly 200 bigrams and we select 35 most informational bigrams by a little manual selection.

Selecting the *set of terms* require an initial term-frequency matrix, comprising all the features (or term vectors). However, after selecting the features, based on entropy values, we perform a second pass on the whole document space in order to create a low-dimensional term-frequency matrix. In this pass, the term frequency is not calculated purely based on strict comparisons but with a slight tolerance for errors. The rational behind this is that the text extracted by Google OCR, misspells various words and frequent misspellings include missing letters or presence of some special characters inside the word. We have used a tool for approximate string matching called *agrep* (http://www.tgries.de/agrep/) for this purpose. The tolerance level was set to one edit error in the unigrams and two in the bigrams. The precision drops slightly by this method but the coverage of words or recall [2] is high. Finally, the term' frequencies are weighted by multiplying them with the inverse of documents in which they appear. This is commonly known as tf-idf representation. Finally, the term-frequency matrix obtained by the aforementioned procedure contains each document as a 435 dimensional vector *D*

$$D = [uni_1, uni_2, ..., uni_{400}, bi_1, bi_2, ..., bi_{35}] \tag{2}$$

where, the first 400 dimensions represent unigrams and next 35 represent bigrams.

3.2. Information Extraction and secondary data

The preparation of secondary data mainly requires parsing the HTML pages of CIAFactbook (https://www.cia.gov/library/publications/the-world-factbook/) for both the *Country-wide Statistics* and *Membership of Climate Treaties*. This data has only been gathered for the countries that presented their speeches in UNFCCC. The data preprocessing includes mean-centering as well as scaling the values between 0 and 1. Table 1 shows a snippet of *Country-wide Statistics* that include 29 variables in total. The dataset for the *Membership of Climate Treaties* includes 26 different treaties (variables) for all 143 countries. Table 2 shows some of them. The numerical and categorical

Table I
A SELECTION OF COUNTRY-WIDE STATISTICS.

Countries	GDP Growth Rate	GDP Per Capita	Birth Rate	...	Pop. Below Poverty Line (%)
Guatemala	2.20	5200	26.96	...	56.20
Kenya	4	1600	33.54	...	50.00
Norway	1.50	59100	10.84	...	NaN
Germany	3.30	35900	8.30	...	11.00
Singapore	14.70	57200	8.50	...	NaN
Mexico	5.0	13800	19.13	...	18.20
...	...	...	...	...	...

Table II
A SELECTION OF THE MEMBERSHIP OF CLIMATE TREATIES.

Countries	Air Polution	Biodiversity	Climate Change	...	Whaling
Guatemala	0	1	1	...	1
Kenya	0	1	1	...	1
Norway	1	1	1	...	1
Germany	1	1	1	...	1
Singapore	0	1	1	...	0
Mexico	0	1	1	...	1
...	...	...	...	...	...

data shown in tables 1 and 2 respectively represent a 29 + 26 dimensional vector for each country.

$$D=[stat_1, stat_2, ..., stat_{29}, cat_1, cat_2, ..., cat_{26}] \quad (3)$$

3.3. Combining Text, Numerical and Categorical data

In order to carry out the analysis, the data from primary and secondary data sources has been combined. Since the secondary data, consisting of numerical and categorical values, provides context to the textual data, it is weighted less. Thus, the textual data will have more influence on the results of the analysis and this enables us to visualize the groups of countries on the SOM map primarily based on the content of the speeches presented in UNFCCC. Nonetheless, the real information regarding the countries, in the form of statistics, is present in the data providing less influential context to the data. The strategy to combine data from both the sources, to form a unified data *V* is shown in the following equation.

$$V=[D_1,...,D_{435}]+e[N_1,...,N_{29},N_{30},...,N_{55}] \quad (4)$$

The dimensionality of *V* is 490 and the value of e is set to a small number in order to reduce the impact of numerical and categorical data on the analysis. The + sign in equation 4 does not signify an addition operation but combining both the datasets (textual + numerical and categorical).

4. Experimentation and Results

The experimentation is primarily of two types. In the first part of experimentation a SOM map of the entire dataset *V*, as shown in the equation 4, is created in order to see the position of data points (precisely countries/organizations in the conference) over the map. This map is shown in Fig. 1. In this figure, the shades of gray denote distances in the original high-dimensional data space. The darker the color the higher the distance is in the original space. The most interesting part of the map in Fig. 1, from the analysis point of view, is the top left and top center region. In this area most of the European and rich countries are in the proximity of each other. Moreover, various developing countries are scattered in small groups in the bottom of the map and a considerably visible group of some African countries can be found in the middle of the map.

The next part of experimentation deals with analyzing the distribution of various variables or components on the map. The first set of variables includes unigrams. These are 400 in total; the

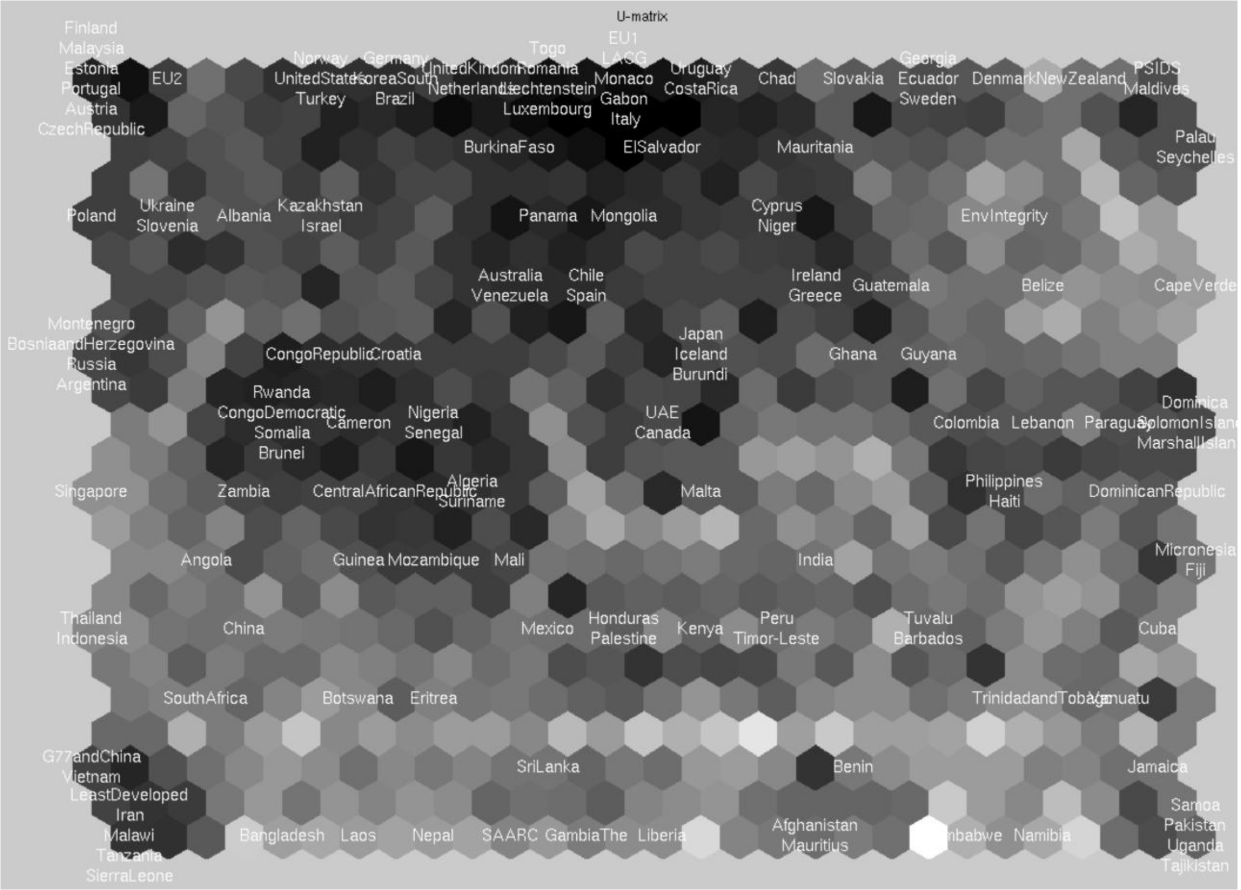


Fig. 1: SOM map of countries

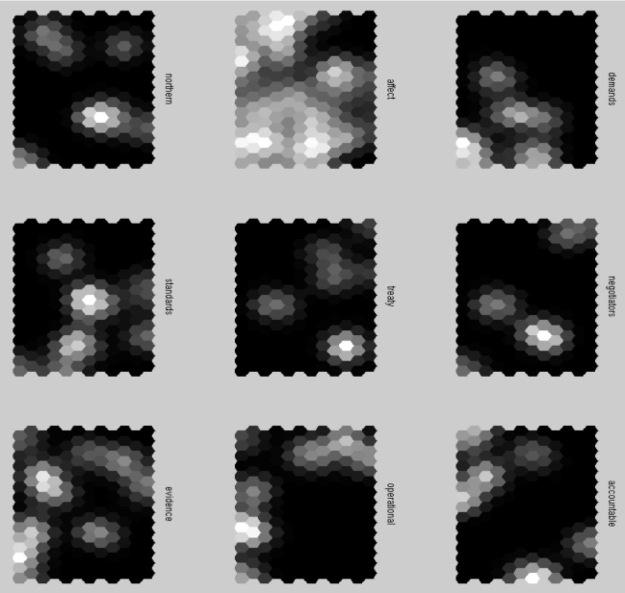


Fig. 2: Component map of unigrams

distribution of a few of them is shown in Fig. 2. The light shades of gray on the component maps show large values of the corresponding variable and dark shades show low values. Thus, we can see that in Fig 2 the unigram *affect* is unevenly scattered throughout the map except the top right. However, the top right portion is dark, explaining that countries in that region have used the word *affect* lesser than other countries. Moreover, we can also find interesting correlations between/among the unigrams. For instance, the density of *operational* is higher in those regions where the density is higher for *affect*. Interestingly this is the region where many underdeveloped countries are located as shown in Fig. 1. The next figure shows a component map of bigrams similar to the Fig. 2.

In Fig. 3 we can clearly see that the distribution of bigrams such as *climate change*, *Kyoto protocol*, *developing countries*, *legally binding* etc. is relatively higher in the regions of underdeveloped countries. This explains a correlation among these bigrams in the content of the speeches of underdeveloped countries.

The next two figures (Fig. 4 and Fig. 5) show the component maps of country-wide statistics and membership to the climate treaties respectively. Fig. 4 shows that *Life Expectancy at Birth* and *Literacy* is higher in the top regions of the map. This is clearly inline with the results shown in Fig. 1 since most of the developed world (including European countries and the US) is located in the top regions. Interestingly, these distributions do not deviate from the general understandings about the developed and developing world. In Fig. 5 we can see that significant portion of component maps of *Climate Change* and *Climate Change-Kyoto Protocol* are lighter in colors. This means that most of the countries have signed both of these treaties; however, some countries in the region of various African countries (see Fig. 1) have not signed *Climate Change*.

5. Conclusion

In this paper, we present a methodology to perform document analysis while putting contextual information in the background. The contextual information comprises of real statistics and their affect in the analysis is marginalized so that the analysis is dominated by the textual data. This has further helped us comparing the two sets of information and it reveals several interesting and obvious findings. We have found that the countries that are geographically close to each other and/or have similar socio-economic conditions are located in the proximity of each other on the analysis map. This is reinforced by analyzing the distribution of several variables, on the analysis map, representing concrete and real information regarding the countries. Our findings suggest that the countries that belong to similar groups in terms of their socio-economic conditions tend to speak in a similar manner when it comes to addressing the issue of climate change.

Our analyses open several areas of investigations for both social scientists as well as computer scientists. For example, it is worth asking if the divide between the rich, developing and under developed countries is also reflected in their actual concerns over the climate change. Or, as regards the global warming, do the underlying groups of countries (as investigated by this research) also take similar steps in order to address the challenges of climate change? Another simple step further in the context of this research is to take help from human experts to annotate the speeches (the data can be made available upon request) and then measuring the accuracy of the results presented in this paper. Finally, a concept level analysis of textual data, in a way that data features are essentially concepts (like ontologies in Semantic Web), would be an interesting investigation to consolidate the results of this papers in particular and document analysis in general. A by-product of such a research is a significant reduction in the size of lexicon and thereby the feature space [13, 20]. In general, we wish that the kinds of analyses presented in this paper could contribute in supporting sustainable developments in the world.

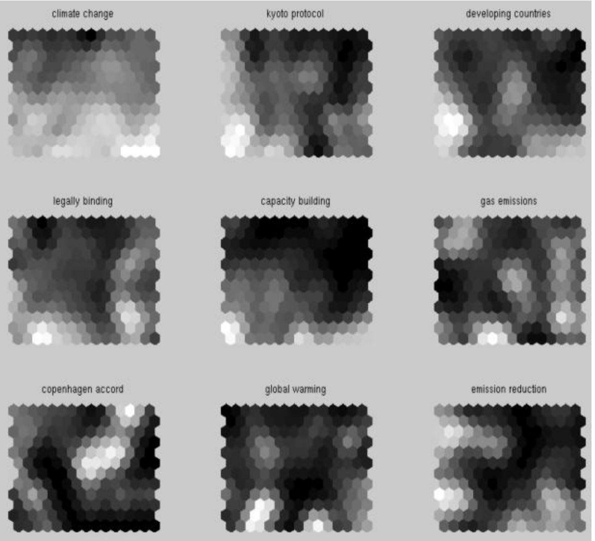


Fig. 3: Component map of bigrams

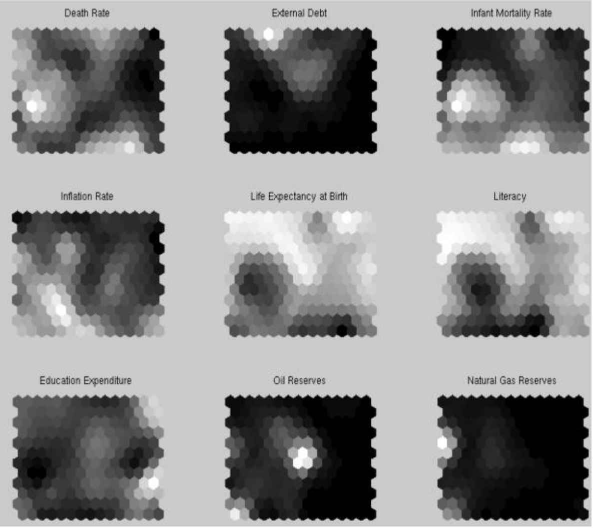


Fig. 4: Component map of numerical stats

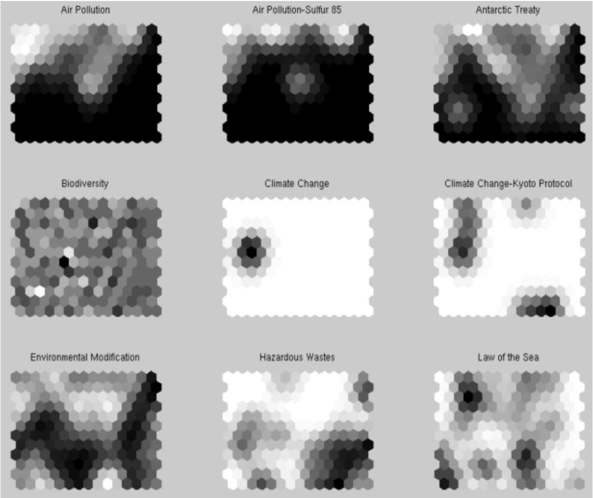


Fig. 5: component map of categorical data regarding climate treaties

References

[1] T. Kohonen. *Self-organizing maps.* Springer Series in Information Sciences, 2001.

[2] P. Senellart and V.D. Blondel. Automatic discovery of similarwords. *Survey of Text Mining II,* pages 25–44, 2008.

[3] M.A.A. Cox and T.F. Cox. Multidimensional scaling. *Hand- book of data visualization,* pages 315–347, 2008.

[4] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *Neural Networks, IEEE Transactions on,* 8(1):148– 154, 1997.

[5] G.J. Deboeck and T. Kohonen. *Visual explorations in finance with self-organizing maps,* volume 2. Springer, 1998.

[6] Debbie Zhang, Simeon J. Simoff, and John K. Debenham. Exchange rate modelling for e-negotiators using text mining techniques. In *E-Service Intelligence,* pages 191–211. 2007.

[7] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics,* 8(5):358–375, 2007.

[8] S Kaski, J Nikkilä, M Oja, J Venna, P Törönen, and E. Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics,* 4(1):48, 2003.

[9] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management,* 12(1):29–41, 2004.

[10] S. Takahashi, M. Takahashi, H. Takahashi, and K. Tsuda. Analysis of stock price return using textual data and numerical data through text mining. In *Knowledge-Based Intelligent Information and Engineering Systems,* pages 310–316. Springer, 2006.

[11] D. Freestone. From copenhagen to cancun: Train wreck or paradigm shift? *Environmental Law Review,* 12(2):87–93, 2010.

[12] S.C. Walpole, S. Singh, and N. Watts. International climate negotiations: Health to the rescue? *The International Journal of Occupational and Environmental Medicine,* 2, 2011.

[13] P. Koehn. *Statistical machine translation,* volume 9. Cambridge University Press, 2010.

[14] T. Liu, S. Liu, Z. Chen, and W. Ma. An evaluation on feature selection for text clustering. In *MACHINE LEARNING- INTERNATIONAL WORKSHOP THEN CONFERENCE-,* volume 20, page 488, 2003.

[15] C.C. Aggarwal and P.S. Yu. Finding generalized projected clusters in high dimensional spaces. *ACM SIGMOD Record,* 29(2):70–81, 2000.

[16] I. Jolliffe. Principal component analysis. 2002.

[17] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on,* volume 1, pages 413–418. IEEE, 1998.

[18] I.K. Fodor. A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory,* 2002.

[19] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters,* pages 83–86, Manchester, UK, August 2008.

[20] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on Data Mining,* pages 541–544. IEEE, 2003.

Urdu Syllable: Templates and Constraints

Mazhar Iqbal Ranjha
Minhaj University, Lahore
mazranjha@gmail.com

Abstract

Urdu has 4 heavy syllables or bi-moraic (CVC, CVV, VC, VV) and 5 super-heavy syllables (CVCC, CVVC, CVVCC, VCC, VVC), "C" denotes consonant, "V" light vowel, "VV" long vowel and "CC" denotes consonant cluster. In Urdu, "CC" is not allowed in the onset position. It only occurs in the coda position. This paper aims to consult one of the authentic Urdu dictionaries and search all templates with "CC" consonant clusters and constraints. It also aims to investigate, on one hand, all possible consonant combinations and on other hand try to investigate the syllabic principles followed in Urdu.

1. Introduction

Urdu is national language of Pakistan. "Urdu is second or third highest spoken language in the world. It is approximately spoken by 591,000,000 people in more than ten countries as first or second language" [7]. It has rich phonology which has not been fully explored yet. Every language has its rules and parameters. Some languages take complex onsets and other complex codas. Some have complicated syllable structure other have easy. Some follow one rule and some other. Some languages allow different combinations of consonant clusters in coda or onset positions. As Persian, Arabic, Hindi, Sanskrit, Portuguese and other local languages contribute a lot in the vocabulary of Urdu so these languages continue to cast their influence in the direction of evolution of Urdu. The syllable templates of Urdu are similar to the templates of these languages.

Working on stress in Urdu, Hussain [2] has touched its other aspects as well. Describing syllable weight, he has listed 12 syllable templates in Urdu. He has divided them into simple and complex onset templates and urges further need of study to confirm the complex ones. Besides him, Ghazali [4] and Nazar [5] have worked on it as well. To some extent they agree on the number of templates but differ in their opinion in phonotactic constraints and consonants clusters in syllables. Akram [6] has also investigated the syllabification and phonotactic constraints observed in Urdu and come with some new ideas. This paper aimed to search such clusters and analyze their behavior. For this purpose, an Urdu dictionary *Feroz ul lughaat* [1] was studied. All the words with consonant clusters were collected. Other phonotactic constraints were also investigated.

2. Literature Review

2.1. Syllable

"Syllable is an essential concept for understanding phonological structure" [11, p. 250]. It is an important unit of language but controversial to be defined. "Different attempts have been made to define the syllable in terms of muscular contraction and in terms of peaks of sonority but no completely satisfactory definition has been found" [12, p. 214]. "It is relatively easy for people to count the syllable of a word – much easier than counting the segments" [9, p. 250]. Ladefoged [3] says that although everybody can identify it, nobody can define it. He further states that every utterance must contain at least one syllable [3, p. 230]. Hayes [9] calls it the stressed bearing unit. Perhaps everybody finds syllable comparatively easy to define that is why no serious attention has been paid on defining it. "Every speaker has an intuitive notion of how many syllables each word has. It is less easy for speakers to reflect consciously on the internal structure of a syllable" [13, p. 105].

2.2. Syllable structure

Though the native speakers of any language find it easy to tell how many syllables are present in particular utterance yet it is difficult to give is proper definition that can clarify its phonetic and phonological character. The best way to define a syllable is to talk of its segments. A speech consists of two segments i.e. consonants and vowels. The universal syllable template accepted by most phonologists is denoted by Latin symbol σ (sigma).

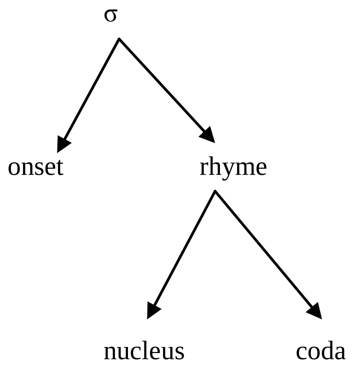


Figure 1

Onset is a consonant or group of consonants that precedes rhyme. Hayes [9] defines onset as consonant or sequence of consonants at the beginning of a syllable. Rhyme of a syllable consists of the vowel and any consonant/s that come(s) after it [3, p. 230]. The nucleus of syllable is the vowel or diphthong found at the syllable's core and functioning as sonority peak. It is an obligatory part for a syllable to have nucleus [9, p. 251]. It is a vocalic part [3, p. 230]. The final part of a syllable, consists of consonant/s, is called coda. The coda and onset are optional parts of a syllable. Onset, nucleus and coda are also called constituents of a syllable.

2.3. Syllabification

An analytical procedure of grouping or dividing a syllable into its components is called syllabification. Bartlett, Kondrak, and Cherry [8] define syllabification as the process of dividing a word into its constituent syllables. Referring Muller et al. and Bouma, Marchand & Damper, they write that technically speaking, syllables are phonological entities that can only be composed of strings of phonemes. Referring Blevins, they say that most of linguists view syllables as an important unit of prosody because many phonological rules and constraints apply within syllables or at syllable boundaries. Citing Goldsmith, Akram [6] names it a process that associates a linear string of segments with a syllable structure. He further writes that from a descriptive point of view, word should be factorable into sequences called syllables, which should have a specifiable internal structure that is roughly constant across the language [14, p. 107].

2.4. General principles of syllabification

"The basis on which syllabification is derived must be (partly) language specific: every language has its own principles of syllabification" [9, p. 251]. Hayes [9, p. 252] describes the following three general principles of syllabification for any language.

- Finding the syllable nucleus
- σ Syllabic affiliation of consonants
- An outline scheme for syllabification

2.5. Theories of syllabification

"There is debate as to the exact structure of a syllable" [8, p. 309]. There is some agreement between the linguistis about a nucleus preceded by onset and followed by coda being constituents of a syllable. A syllable is language specific. Every language has different typological parameters. Number of theories concerning syllabification have been presented by different linguists in different eras.

2.5.1. The Legality Principle. According to this principle "a syllable is not allowed to begin with a consonant cluster that is not found at the beginning of some word, or end with a cluster that is not found at the end of some word" (Goslin and Frauenfelder quoted in [8, p. 309]). Giving an example of an English word *admit,* he says that according to this principle, it is written as [əd-mit]. It cannot be written as [ə-dmit] because no word in English starts with [dm] cluster. In the same way an Urdu word [əlhəmd] cannot be syllabified as [ə-lhəmd] because no cluster is allowed in Urdu in the onset position.

This principle has its limitations and cannot be accepted as universal principle for all languages. Giving an example of a word like *askew* [əskju], [8] comments that "this principle cannot rule out any of [ə-skju], [əs-kju], or [əsk-ju], as all three employ legal onsets and codas".

2.5.2. Maximal Onset Principle (MOP). According to this rule, maximum consonants are preferred in the onset position. "A consonant which may in principle occupy either rhyme or onset will occupy onset position" (Trasj, 1996, p. 217). In Maximal Onset Principle "the consonants are preferred in the onset and thus allowing no coda consonants except for the word final position" (Golsmoth, 1990, p. 128 as in [6]). In a word where there are more than two consonants, leaving the one in the coda of preceding syllable, the rest will go to onset of following syllable For example *escritoire* [eskritwa:] can be syllabified as [e-skrit-wa:], [esk-rit-wa:] or [es-krit-wa:]. According to MOP, we assign the position of onset to maximum consonants so the division [es-krit-wa:] is correct. And the syllabification of [əskju] will be [əs-kju], Urdu language is very sensitive to onset. It allows only one consonant in the onset position. There are very few words in Urdu with have clusters of more than two consonants. For example an Urdu compound word [?ərzmɚď] has three consonants[rzm], According to this principle, it will be syllabified as [?r.zmɚď], which is not allowed in Urdu. So it can be said that all languages do not follow this principle either.

2.5.3. Maximal Coda Principle (MCP). For syllabification, Maximal Coda principle(MCP) is also used. Opposite to MOP, this principle prefers maximum consonants in the coda allowing no onset consonant except for the word initial position. For example the word [əskju] will be syllabified as [əsk-ju] and [eskritwa:] as [esk-rit-wa:]. As mentioned earlier, Urdu is very sensitive to onsets and it does not follow MCP. In Urdu syllable, if there is only one consonant, then it is preferred in onset. In case of two consonants, the first one will go the onset and the other to coda. If there is cluster of three consonants, very rare case, then it will follow the MCP i.e. two consonants will go to coda position leaving one in the onset position. For example a word [?ərzmɚď] will be syllabified as [?ərz-mɚď], This principle also has flaws to be accepted universally.

2.5.4. Sonority Sequence Principle (SSP). This principle states that sounds should occur on their sonority basis. The sonority of sound is determined primarily by the size of their resonance. “The sonority of a sound is its inherent loudness, holding factors like pitch and duration consonant (Crystal quoted in [8, p. 309]). Ladefoged [3, p. 227] defines as “the sonority of a sound is its loudness relative to that of other sounds with the same length, stress and pitch”. Citing Selkirk, [8] describes that “SSP states that sonority should increase from the first phoneme of the onset to the syllable’s nucleus, and then fall off to the coda”. One can observe it by producing vowel and consonant sound alternatively. The vocal tract is more open while producing vowel sound than that of consonant. According to this rule, sonority slope of sounds rises from onset to nucleus and then falls to coda. For example, an English word *vintage* [vintiʤ] cannot be syllabified as [vi-ntiʤ] because [n] is more sonorant than [t]. [t] being stop is least sonorant and according to SSP, sonority should increase from first phoneme to the nucleus i.e. [nt] is not possible cluster. In Urdu word [?ərz-mɚď], the coda of first syllable [rz] follows SSP as liquid [r] is more sonorant than fricative [z]. The sonority hierarchy [14, p. 111] is listed in the figure2.

- Vowels
  - Low vowels
  - Mid vowels
  - High vowels
- Glides
- Liquids
- Nasals
- Obstruent
  - Fricatives
  - Affricatives
  - Stops

Figure 2: sonority hierarchy by Goldsmith

According to Goldsmith, this is “necessary condition for basic syllabification and is universally accepted with few exceptions”.

2.6. Urdu Language

“Urdu is national language of Pakistan and spoken by more than 100 million people across a more than score countries” [15, p. 01]. “It is popularly regarded as offspring of Persian. It borrows words from different languages to expand its vocabulary. Major languages participating in the camp of Urdu are: Persian, Arabic, Portuguese and English” (Saksena, 1990 quoted in [10]). There are also considerable words of Sanskrit and Hind in Urdui. “Urdu belongs to the family of New Indo-Aryan (NIA) language, which is a sub-branch of the Indo-European language”[2, p. 39]. “Indo-European language family is the most widely studied language as more than half of the world’s population speaks one or more of these languages either as a mother tongue or as a business language” [7]. “Urdu and Hindi both belong to NIA Language. They are different literary styles based on the same linguistically defined sub dialect” [16, p. 27]. “In spite of having the same origins and having a very similar linguistic structure, Urdu phonetics and phonology have diverged form Hindi phonetics and phonology. The divergence is perhaps caused by the strong Perso-Arabic influence on Urdu and the strong Sanskrit influence on Hindi” [2, p. 40].

2.7. Urdu Templates

As Persian, Arabic, Hindi, Sanskrit, Portuguese and other local languages contribute a lot in the vocabulary of Urdu so these languages continue to cast their influence in the direction of evolution of Urdu. The syllable templates of Urdu are similar to the templates of these languages. Not much research has been done on Urdu phonology and little material is available about the structure of Urdu syllables. The work found on Urdu templates is only done by [17] and [2]. As mentioned earlier, the identification of a syllable of a language is intuitive to native speakers but very hard to define. Different theories about the syllabification have been given above but none of these has proved adequate. The syllable is considered as an abstract unit of prosodic organization through which a language expresses its phonology.

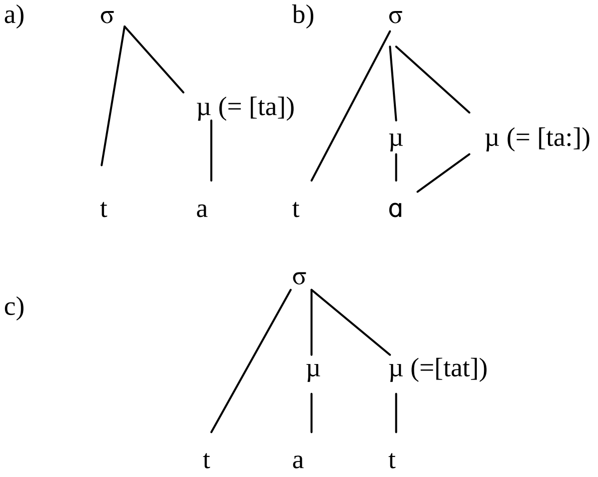
Recently, among the United States linguists, there has been found a trend of defining "mora as the element bearing phonological weight". [18] describes that the linguists like Hyman, McCarthy & Prince, Ito, Hayes, Archangeli are of the opinion that "mora plays a major role in syllabic structure". In the theory of Hayes, moras have replaced syllables altogether. The basic concept expressing syllabicity is the Weight Unit (WU). Each segment has a WU.

2.7.1. Mora. Han in [19] defines mora as a "unit equal to a short syllable". Ladefoged defines it as "a unit of timing". Long vowels are often considered to be bimoric, whereas short ones are monomoraic. Husssain describes it as "a time unite equivalent to a single short vowel". However, a mora is not a "species of sound but rather an elementary prosodic unit …… like the syllable ….intervening between the [syllable] and the phonetic string" (McCarthy and Prince, and Hayes as in [11, p. 392]).

2.7.2. Moraic Structure. Syllables are divided into moras, which function as WU. Moraic structure of languages can vary. For example, in some languages like Latin, both CVV and CVC syllables are counted as heavy (figure 2) and CV as light. Other languages, CVC does not count as heavy (figure 3).

2.7.3. Moraic Languages. Moraic languages are those in which "the mora plays a part in the phonology or the metrical system". The distinction between mora-timing and syllabic-timing languages is not clear. From terminology "mora-timing" does not mean "moraic". In a mora-timing each mora takes approximately the equal time to be pronounced. It means that a heavy (2-mora) syllable will take twice as long as a light mora. So a moraic language is not essential to be moraic-timing language. These two ideas stand quite apart.

2.7.4. Syllable template. The study reveals that "the syllable templates of any language provide a better understanding of the phonological properties of that language". The templatic syllabification permits a simpler and more successful analysis of a language. "Syllable templates represent a fixed static set of constraints that dictate the syllable structure in the language concerned". "Templatic syllabification may need some revision" [11, p. 276] but it still stands valid for elementary syllable inventory like Arabic.



(adapted from www.rnsoke.home.xs4all.nl/CV/publicaties/proefschrift/diss2.pdf)

Figure 3

2.7.5. Syllable Structure of Urdu Templates. "A syllable template is formulated in terms of sequence of consonants and vowels" [5]. For example, "a syllable template of Arabic language is CV(V)C, where C denotes a consonant, V denotes vowels and ( ) stands for optional templatic element" [11, p.270]. Using the moraic concept, it can be said that a short vowel in Urdu is mono-moraic and long vowel a bi-moraic. In Urdu both vowels and coda consonants are moraic. Consonant clusters in the coda of a syllable are bimoraic. Open syllable with short vowels are mono-moraic. Closed syllable with short vowels and open with long vowels are bi-moraic. "Closed syllables with long vowels or with short vowels and coda cluster are tri-moraic" [2, p. 44-45]. "The moraic differences are represented as a difference of the weight of the syllables. Mono-moraic syllables are called *light*, bi-moraic are called *heavy* and tri-moraic syllables are called *super heavy"* [2, p. 45]. Urdu language counts each short vowel and consonant as mono-moraic, long and nasal vowels bi-moraic and consonant cluster is also bi-moraic. In Urdu two consonants are allowed in the coda position. (See 2.5.3 and 2.5.4).

2.8. Templates and Constraints

Though Urdu is one of the largest spoken and understood language of the world yet little work has been done on its phonetic and phonology. Referring [17, p. 17-19], Akram [6] relates 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 consonants r, rʰ or ŋ nor ends in ŋ. Short vowels ə, cannot occur consecutively within a word nor can any one of them follow the middle consonant of three consonants syllable. The biggest Urdu word is tri-syllabic hence complex words containing more than three syllables are compressed and sounds are assimilated to three syllables”.

Talking about phonotactic constraints, [2] has related that open syllables with short vowels do not occur in the word final position. He further says that there can be complex codas and complex onsets in Urdu syllables; however, “there are limitations on formation of these complex onsets and codas” [2, p. 41]. For the formation of syllable, he describes two conditions. “First of all sonority sequencing principle should be satisfied. Secondly, these complex codas can contain at most two consonants. If there are two consonants in the onset, the second consonant in the onset is limited to the glides /w/ or /y/ or may be /h/”. Relating the position of two consonants in the codas, he says that “first consonant is limited to a voiceless fricative (/f/, /s/, /ʃ/ or /x/) or nasals (/m/ or /n/)”. He is also of the opinion that “alveolar flap cannot occur in the onset position” [2, p. 42].

Ghazali [4] and Nazar [5], in their reports, have described eleven (CV, CVC, CVCC, CVV, CVVC, CVVCC, V, VC, VCC, VV, VVC) syllable templates, they found in their researches.

Ghazali [4] has divided these templates into five categories. He enlists 2 light syllables or mono-moraic (CV, V), 4 heavy syllables or bi-moraic (CVC, CVV, VC, VV) and 5 super-heavy syllables (CVCC, CVVC, CVVCC, VCC, VVC). Both of the above persons agree on CVV to be the most and VC, CVVCC and V least frequent syllable respectively.

Relating the phonotactic constraints, Ghazali [4] claims to have found one restriction that in CVC template, “consonants in the onset and coda are allowed to be same only if they belong to the set /t/, /ʈ/, /l/, /ʃ/, /s/, /b/, /m/, /p/, /ʧ/. Talking about Hussain’s [2, p. 42] opinion about phonotactic constraints, he writes that first coda consonant could also be /l/, /z/, /r/, /ɤ/, /b/, ʈ// or /k/ other than fricatives or nasals”. He concludes his discussion about templates by saying that there are only six basic templates in Urdu and rest five have been derived from these basic ones.

Akram also agrees with Ghazali and Nazar on the number of templates found in Urdu. Discussing the syllable structure of Urdu, he relates that “Urdu is onset loving language as if there is only one consonant in between two vowels then it prefers it in the onset rather than in the coda”. For example [azar] will not be pronounced as [az.ar] but [a.zar]. If there are “two consonants together at the end of the word than they both will go in the onset and the coda respectively i.e. first will go in the coda of the first syllable and the second in the onset of the second one”. For example [abdoz] will not be as [a.bdoz] or [abd.oz] but [ab.doz]. The cases of three cons active consonants occurring in the middle of the syllable are very rare. In case of their presence, the first two will go to the coda of the first syllable and third in the onset of the next syllable. For example [gondni] will be [gond.ni] not as [gon.dni].

2.9. Summary

Syllable, the smallest unit of speech sound, consists of three constituents; onset, nucleus and coda. There are different syllabification theories, some of which have been discussed in this paper. Every language has its constraints and parameters. Some languages take complex onsets while other complex codas. Urdu is very sensitive to onset position and takes only one consonant. In the coda position it takes maximum two consonants. There are some sounds which are orthographically present but missing from the spoken language. Sometimes speaker use epenthesis to make pronunciation easy. The fundamental templates found in Urdu are CV, CVC, CVV, CVCC, CVVC and CVVCC where C denotes consonant and V denotes vowel. Urdu is moraic language where short vowel is mono-moraic, long and nasal vowels are bi-moraic. Consonant in Urdu is also mono-moraic. With the deletion of some sound or usage of epenthesis, the template undergoes a change. For example, orthographically an Urdu word *abr* [?əbr] has CVCC template structure but when a speaker deletes glottal stop, it becomes [əbr] with VCC template. In the same way [sərʰ-ɑ-ne] has CVC-VV-CVV template structure but now in present Urdu it is produced as [sər-hɑ-ne] with CVC-CVV-CVV template structure. With the deletion of [h] sound from [sʊbh] (mornig), it becomes [sʊbɑ]. Hence template CVCC changes to CVCVV. In same way *sang* [səng] (accompany) has CVCC and when [n] nasalizes the preceding vowel, it becomes [sɚŋ] with CVVC. Urdu word [ʔəql] (wisdom) has CVCC template and when [ə] is inserted [ʔəqəl], templates becomes CVCVC.

This paper aims to study the syllables templates, consonant clusters, constraints in Urdu.

3. Methodology

The best medium to study the phonology of any language is either its dictionary or native speakers of the target language. As my paper aimed to study all consonant clusters available in Urdu, so dictionary seemed the best medium to study them. For this purpose, I have selected "*Urdu Feroz ul lughaat*" containing more than one lac words, proverbs and idiomatic sentences.

All the root words with "CC" cluster were copied to examine the behavior of the clusters. All the words were transcribed and target templates were identified. Words belonging to English origin were not considered. "[n] has two orthographic representations for nasal. One of which is called "*noon ghunna*" like [hɑ̃s] (smile) and other is "*noon*" like [həns] (swine)". In current study, only *noon* sound was considered as *noonghunna* only nasalizes the preceding vowel.

In the first step all words with consonant clusters were searched and enlisted. In the second step, frequency of all clusters was found (see Appendix A). Sounds on basis of manners were also grouped.

4. Results

In *Fero-ul-lughaat*, 699 words having 702 consonant clusters have been found (see Appendix A). None of these was found following the MOP and 242 out 702 clusters were observed violating SSP.

Total templates with CC clusters = 702
The number of templates violating SSP = 242
Other cluster combinations are: Fricative-stops =168, Tap-stops = 077, Both stops = 046, Tap-fricatives = 043, Fricative-taps = 040, Both fricatives and Fricatives-nasals = 039, Stop-fricatives = 038, Stop-taps = 037, Fricatives-liquids = 027, Liquid-stops = 022, Stop-liquids and Stop-nasals = 016, Nasal-fricatives = 015, Tap-nasals = 014, Nasal-stops = 012, Liquid-fricatives and Tap-affricatives = 010, Affricative-stops and Affricative-fricatives = 005, Both nasals = 004, Nasal-liquids, Affricatives-taps, Nasal-taps, Liquids-nasals and Affricative-nasals = 003, Liquids-nasals and Fricative-affricatives = 002, Affricative-liquids and Fricative-retroflex = 001, Both approximants, Both Affricatives, Both Liquids, Both Taps, Nasal affricative, Liquid affricatives, Stop affricatives, Tap liquids and Liquid taps = 000.

As mentioned above these results were bases upon Urdu dictionary *Feroz ul lughaat*.

5. Discussion

Urdu is a language which has the ability to absorb new words in it easily. As Persian, Arabic, Hindi, Sanskrit, Portuguese and other local languages contribute a lot in the vocabulary of Urdu so these languages continue to cast their influence in the direction of evolution of Urdu. The syllable templates of Urdu are similar to the templates of these languages. Each language has its own paradigms and parameters and puts restrictions on its templates and syllabification.

The data observed reveal that Urdu does not follow Maximal Onset Principle (MOP) as it is very sensitive to its onsets and allows only one consonant in this position. Urdu has mixed behaviour towards Sonority Sequence Principle (SSP). It neither follows it completely nor goes against it completely. However, it follows Maximal Coda Principle (MCP) completely.

MCP prefers maximum consonants in the coda position leaving only one in the onset position. There are rare clusters with three consonants in Urdu. Mostly there are two consonants clusters. In case of three consonants, the two will go to the coda of the preceding syllable leaving one to form the onset of following syllable. For example a word **[**gʊ**zaʃţn**i] has a cluster of three consonants [ʃţn]. According to this principle, the first two [ʃţ] will take coda position of the first syllable and [n] will occupy the place of coda of the following syllable like **[**gʊ.**zaʃţ.n**i]

The number of consonant clusters found in the above said dictionary was 185. The cluster [sţ] had highest frequency 39 and [rď], [xţ] with 30 and 22 respectively. (For detail see Appendix B)

If we talk about the clusters from articulator point of view then some of the sounds are more frequent than the others. Sometime both consonants in the cluster are voiced and other time voiced-unvoiced combination is occurring mostly. Some other interesting facts have also been found. Fricative-stop clusters have been found in the highest number i.e., 130 out of 168.

If we observe these clusters in a little detail, then we find that in case of both stops, Urdu does not prefer both voiced consonants. Unvoiced-unvoiced combinations are more preferred e.g., out of 168 Fricative-stops, 130 are both unvoiced, both voiced, unvoiced-voiced, voiced-unvoiced are 13, 19 and 06 respectively. Similarly 77 Tap-stops with 47 both voiced and 30 voiced-unvoiced; 46 both stops with 03 both voiced, 11 both unvoiced, 17 unvoiced-voiced and 15 voiced-unvoiced have been observed.

Urdu does not take voiced fricatives as second consonant, if the first one is tap [r]. Tap [r] also prefers unvoiced before it. In case of both fricatives, both unvoiced are preferred. Nasals [m], [n] also prefer unvoiced preceding fricatives. Both unvoiced stop-fricatives are preferred in clusters. Tap [r] likes unvoiced stops in preceding position. Liquid [l] also prefers unvoiced preceding fricatives and unvoiced stops in both positions. In case of Stop-nasal cluster combination, unvoiced stops are more frequent. Nasals also prefer unvoiced fricatives and unvoiced stops in the following position. Liquid [l] mostly takes unvoiced fricative consonants. Tap [t] prefers voiced affricatives in the following position and most of the affricative-stops clusters are voiced. Unvoiced affricatives do not take unvoiced stops. Nasals are always preceded by voiced affricatives. Taps always take voiced affricatives before them. There are only two examples of fricative-affricative combinations. These show that fricative-affricatives are either both voiced or unvoiced fricative-voiced affricative combination.

Only one example of Affricative-liquid found shows that liquids is followed by voiced affricative. Only one fricative retroflex combination was observed. (See Appendix C)

As for as constraints are concerned, primary data also reveal that [ʈ], [ʧ], [ɖ], [v], [ɽ], [ʒ] sounds do not occur in the first position in "CC" clusters where as [ɖ], [ɽ] and [j] do not occur in the second position of these clusters. In Urdu, approximant-approximant, affricative-affricative, liquid-liquid, tap-tap, liquid-affricative, nasal-affricative, stop-affricative and tap-liquid combinations are non-existent. Or we can say that Urdu constrains the following cluster combinations.

None of these clusters has been found in the dictionary. Both approximants, Both Affricatives, Both Liquids, Both Taps, Liquid affricatives, Nasal affricatives, Stop affricatives, Tap liquids, Liquid taps

6. Conclusion

After studying the foresaid dictionary, it was observed that overall 185 consonant clusters within 702 templates were found. The detail study reveals that in Urdu that no consonant cluster is allowed in the onset position. Maximum two consonants are allowed in the coda position (There is one exception [ʔəmrv] pronounced as [ʔəmr] where [v] is silent). For syllabification, in Urdu only MCP is followed. Mostly there are clusters of unvoiced sounds. Cluster combination of unvoiced fricative-stops is more frequent. If both are stops, voiced stops are least preferred. If there is fricative-tap cluster, then ratio of unvoiced fricatives is higher than voiced ones. Taps do not take voiced fricatives after them.

Unvoiced affricative-stop, affricative-fricative, unvoiced affricative-voiced fricative combinations are not allowed in Urdu. Nasals and taps are always preceded by voiced affricatives. Urdu constrains Approximant-approximant, Affricative-affricative, Liquid-liquid, Tap-tap, Liquid-affricative, Nasal-affricative, Stop-affricative, Tap-liquid, Liquid-tap the following cluster combinations.

References

[1] *Feroz-ul-lughat Urdu*, Feroz Sons Private Limited, 2005.

[2] S. Hussain, *Phonetic Correlates of lexical stress in Urdu*", Northwestern University, IL, USA, 1997.

[3] Ladefoged, *A Course in Phonetics* , Thomson Wadsworth, USA, 2000.

[4] M. Ghazali, "Urdu Syllable Templates", *Annual Report of Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2002.
Available: www.crulp.org/research/reports/streports02.htm

[5] M. Nazar, "Syllable Templates in Urdu Language", *Annual Report of Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2002. Available: www.crulp.org/research/reports/streports02.htm

[6] B. Akram, "Analysis of Urdu Syllabification Using Maximal Onset Principle and Sonority Sequence", *Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2002. Available: www.crulp.org/research/reports/streports02.htm

[7] N. Wyne, "Languages and their families". *Annual Report of Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan. Available: www.crulp.org/research/reports/streports02.htm

[8] S. Bartlett, G. Kondrak, and C. Cherry, *On the Syllabification of Phonemes* 2009. Available: www.aclweb.org/anthology-new/N/N09/N09-1035.pdf

[9] Hayes,. *Syllables*, 2009.
Available:
www.udel.edu/~heinz/classes/2011/607/materials/…/Hayes2009-13.pdf

[10] A. Saleem, H. Kabir, K. Riaz, M. Rafique, N. Khalid and S. Shahid, "Urdu consonantal and vocalic sounds" *Annual Report of Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2002. (www.crulp.org/research/reports/streports02.htm)

[11] M. Kenstowicz, *Phonology in Generative Grammar*, Blackwell Publication, UK, 1994.

[12] R. Trask, *A Student's Dictionary of Language and Linguistics* , Arnold, London, 1997.

[13] A. McMahon, *An Introduction to English Phonology*, Edinburgh University Press, Edinburgh, 2002.

[14] J. Goldsmith, *Autosegmental and Metrical Phonology*, Basil Blackwell Ltd, UK, 1990.

[15] S. Hussain, "Letter to sound Conversion for Urdu text to speech system", *Annual Report of Center for Research in Urdu Language Processing (CRULP)*, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2004.

[16] C. Masiaa, *The Indo-Aryan languages*, Cambridge University Press, Great Britian, 1991.

[17] S. Bokhari, *Phonology of Urdu Language*, Royal Book Company,Karachi, Pakistan, 1985.

[18] *Moraic versus constituent syllables*, Available: http://rnoske.home.xs4all.nl/CV/publicaties/proefschrift/diss2.pdf

[19] *Moraic Phonology*, Available:
http://www.ling.fju.edu.tw/phono/farrah/Moraic%20Phonology.htm

Appendix

The appendices are available at the website of the conference. i.e. http://www.cle.org.pk/clt12/

Developing a Part of Speech Tagset for Sindhi

Mutee U Rahman
Department of Electrical Engineering and Computer Science, Isra University, Hyderabad Sindh 71000, Pakistan
muteeurahman@gmail.com

Abstract

Part-of-Speech (POS) tagging is process of assigning unique grammatical tags to every word in a sentence. POS tagset is primary requirement of POS tagging process. This research paper discusses various grammatical classes of Sindhi with reference to POS tagset design and tagging. Various issues like tagset design considerations, tagset size and granularity, part of speech types, subtypes and their attributes for tagging are discussed in detail. General guidelines for designing Sindhi POS tagset of any possible size and granularity are given. Obligatory and proposed tagsets for Sindhi are presented which provide basis for further research in part of speech tagging, tagged corpus, chunking, syntax analysis, information retrieval, part of speech usage analysis and other natural language processing applications.

1. Introduction

Part-of-Speech tagging is key research area in natural language processing and computational linguistics. POS tagging is prerequisite of chunking and parsing of natural language text. It is an essential requirement to develop computational grammar of any language. Information retrieval systems make extensive use of POS tags for text indexing. Text to speech systems use these tags for pronunciation of words. POS taggers are used to tag huge amount of text to develop POS tagged corpus which is key resource for text analytics and intelligent text processing. Basic requirement of POS tagging is a properly defined part of speech tagset.

Despite of few research initiatives of NLP resource development for Sindhi language [1], [2], [3], [4] the development of Sindhi POS tagset is still an open research area. Neither the published work regarding tagset design nor the design guidelines are available.

Following sections discuss tagsets, existing work in Sindhi POS tagging, Sindhi word classes and their division with reference to POS tagset design, possible attributes of word classes, obligatory tagset, and recommended tagset of Sindhi in detail.

2. Tagsets

Tagset is a list of lexical entries with their grammatical categories or tags (POS tags). Language specific tagset is primary requirement of any tagging algorithm. Tagset design issues include representation of linguistic information needed, size, and granularity of the tagset. Generally, larger tagsets are more useful but result in lower accuracy and smaller tagsets are less useful but result in more accuracy.

Some example tagsets for English include London Lund Corpus [5] with 197 tags, Lancaster UCEREL [6] with 165 tags, LOB corpus [7] with 135 tags, Penn POS tagset [8] with 48 tags.

Apart from English, various tagsets are also available for other languages including; Urdu [9], [10], Hindi [11], Pashto [12], Sindhi [1] and Punjabi [13].

3. Existing Work in Sindhi POS Tagging

Sindhi is one of the less resourced languages in NLP studies. Only few published research papers are available on POS tagging of Sindhi. Which discuss rule based and wordnet based POS tagging in Sindhi. A rule based POS tagger [1] is first ever published work on POS tagging of Sindhi which presents a rule based tagging algorithm with a tagset of 67 tags. Few disambiguation rules for tags and some tokenization issues with tokenization scheme are discussed.

Sindhi POS tagging using wordnet [2] is another published work available. The work is almost identical to the above discussed research work. Instead of rules wordnet is used for tag disambiguation.

A comparison of rule based and wordnet based tagging algorithms is given in [3]. The paper concludes that wordnet based approach gives more accurate results compared to rule based approach.

Problems with the tagset proposed in above papers include un-necessary granularity (for example nouns are given twenty different tags), ambiguity in tags (generic as well as specific tags are present) and use of Preposition tag (in Sindhi prepositions don't exist instead postpositions are used). Also, the tagset does not follow any standard guidelines for POS tagset design.

POS tagging research work in Sindhi is still subject to research and major areas of concentration are POS tagset design and comprehensive approach of POS tagging that can serve as the basis of further research on parsing, machine translation and other NLP applications.

4. Sindhi Word Classes

Sindhi word classes are divided into eight different parts of speech which include: noun (اسم), pronoun (ضمير), postposition (حرف جر), adjective (صفت), adverb (ظرف), verb (فعل), conjunction (حرف جملو) and interjection (حرف ندا). Most of the classes are further divided into subclasses. While defining the POS tagset word classes and their subclass considerations directly affect the size of the tagset. Various issues related to the selection of subclasses for defining tagset need to be considered before tagset finalization.

Generally, word classes are divided into two broad categories: closed word classes and open word classes. Following sections discuss open and closed Sindhi word classes in detail.

4. 1 Closed Word Classes in Sindhi

Closed word classes in a language are those classes which have relatively fixed membership. These classes usually contain small number of words compared to open word classes. Postposition (حرف جر) in Sindhi (preposition in English) is an example of closed word class. New postpositions are rarely added. Table-1 shows closed word classes of Sindhi and their examples.

4.1.1. Pronouns (ضمير). Like other languages pronouns in Sindhi are forms that act as a kind for referring a noun and are considered closed word class.

Sindhi pronouns are divided into seven different types which include: demonstrative pronoun, wh-pronoun, reflexive pronoun, relative pronoun, co-relative pronoun, and indefinite pronoun. Personal pronouns are further divided into three categories namely: first person, second person and third person pronouns.

Table 1: Closed word classes in Sindhi.

S.No	Word Class	Examples
1.	Pronoun (ضمير)	تون، توهان، اسين، مان، آئون، اُهي، اِهي
2.	Postposition (حرف جر)	کي، تي، ۾، جو، جا، جي، مان، تان، تائين
3.	Conjunction (حرف جملو)	۽، يا، پر، ته به، پر، ڇاڪاڻ ته
4.	Intransitive (فعل لازمي), Transitive (فعل متعدي) Auxiliaries (فعل معاون)	آهي، ٿو، ها، پيو، هلان، ٻوڙيو
5.	Numerals, Cardinals, Ordinals, Fractals, Multipliers	هڪ، پنج، ڇهون، پنجوٽو، اڌ، منو، پاءُ
6.	Negative, Affirmative	نه، ڪونه، ڪونهي، ها، پلي
7.	Articles	ٻه، پڻ، ئي، ته

Demonstrative pronouns are also divided into two categories. Table-2 shows different types of pronouns in Sindhi along-with their examples.

Third person pronouns have ambiguity with demonstrative pronouns; for example: third person pronouns ho:a هو (that/he) and uhe: اُهي (they) can also be demonstrative pronouns. In the sentence ' اُهي ڇوڪرا اچن پيا' (those boys are coming) 'اُهي' is remote demonstrative pronoun and in sentence 'اُهي اچن پيا' (they are coming) 'اُهي' is third person pronoun. This ambiguity needs to be resolved during POS tagging process.

Table 2: Sindhi Pronouns and their subclasses

S.No.	Type	Subtype(s)	Example
1.	Personal Pronoun	1st Person	اسين، آئون، مان،
		2nd Person	توهان، اوهين، تون،
		3rd Person	اهي، انهن
2.	Demonstrative Pronoun	Proximate	هيءُ، اِهو
		Remote	هو، اُهو
3.	Wh-Pronoun	-	ڪهڙو، ڇا
4.	Reflexive Pronoun	-	پاڻ، خود
5.	Relative Pronoun	-	جهڙو
6.	Co-relative Pronoun	-	تهڙو
7.	Indefinite Pronoun	-	ڪونه ڪو

4.1.2. Postpositions (حرف جر). Another closed word class in Sindhi is postposition. Postpositions usually come after nouns, pronouns, adjectives and adverbs within their own syntactic position. They show relationship between two nouns, noun and pronoun or nouns and adjectives. There are two types of postpositions in Sindhi: simple postpositions and compound postpositions. Another type of postpositions may also be considered as hidden postpositions which may be tagged as postposition or postpositional/ ablative case of nouns or adverbs. Table-3 shows three types of postpositions and their examples.

4.1.3. Conjunction (حرف جملو). Conjunctions in Sindhi are divided into copulative, concessive, adversative, conditional, interrogative, casual and final categories [14] but syntactically all these categories fall into two main types coordinate and subordinate conjunctions [15]. Table-4 shows these two types with examples.

Table 3: Types of Sindhi Postpositions

S.No.	Postposition Type	Example
1.	Simple	۾، تي، کان، تان
3.	Compound	جي اڳيان، جي وچ ۾، کانسواءِ
4.	Hidden	ڳوٺان، گهران

Table 4: Conjunction types in Sindhi

S.No.	Conjunction Type	Example
1.	Coordinate	۽، يا، پر
2.	Subordinate	جيڪڏهن، جيتوڻيڪ، تنهن هوندي به

4.1.4. Transitive verb (فعل لازمي), Intransitive verb (فعل متعدي) and Auxiliaries (فعل معاون). Sindhi verbs are divided into four major categories: intransitive, transitive, auxiliary and compound verbs. Intransitive verbs (فعل لازمي) are verbs without object in a sentence. For example: in sentence "آئون ڊوڙان ٿو" the word "ڊوڙان" is intransitive verb.

Transitive verbs (فعل متعدي) are verbs with subject and object in a sentence. For example: in sentence "احمد خط لکي ٿو" the word "لکي" is a transitive verb as it has subject "احمد" and object "خط". Transitive and intransitive verbs are further divided into active and passive types.

Auxiliary or helping verbs in Sindhi are used to complete the sentence in different tenses. Auxiliaries are used with main verbs, adverbs and nouns. Examples of auxiliary verb include: آهي، پيو، ٿو etc.

All three types of verbs discussed above are closed class type verbs in Sindhi.

Table 5: Participles in Sindhi

S.No.	Participle Type	Example
1.	Present Participle (اسم حاليه)	ٻوڙندو، ڏيندي، ٻچندا
2.	Past Participle (اسم مفعول)	پڙهيل، ٻوڙيو، لڌل، ڳاتو
3.	Future Participle(اسم استقبال)	وڃڻو، وڙهڻو، وٺيون
4.	Verbal Noun (اسم فاعل)	اپائيندڙ، اپائڻهار،ڪاشيگر
5.	Conjunctive Participle(اسم معطوفي)	ڪائي، پڙهائي، وٺيو

4.1.5. Participles (ڪردنت يا مشتق). Participles are derived from verb roots. As root forms of verbs are closed classes in Sindhi so is the case with participles. Five participle types of Sindhi and their examples are shown in Table-5. Past participle and future participle sometimes act as adjectives in sentences and can create ambiguity during POS tagging.

4.1.6. Adjectives (صفت). Few types of adjectives in Sindhi which include cardinals, ordinals, multipliers, fractals and pronominal adjectives belong to closed class and usually new words are not added in these types. Table-6 shows examples of such types of adjectives.

Table 6: Closed class adjectives in Sindhi

S.No.	Adjective Type	Example(s)
1.	Cardinal	هڪ، ٻه، چار
2.	Ordinal	پهريون، پنجون
3.	Multiplier	ٻيڻو، پنجوڻو
4.	Fractal	اڌ، منو
5.	Pronominal Adjective	اسين ماڻهو

4.1.7. Adverbs (ظرف). Subset of adverbs also belongs to closed class which includes: negative, affirmative, temporal, manner, quantity, and pronoun adverbs. Table-7 shows different types of closed class adverbs in Sindhi.

Table 7: Closed class adverb types in Sindhi

S.No.	Adverb Type	Example(s)
1.	Temporal Adverb	هينئر، ڪڏهن
2.	Manner Adverb	آهستي، ڏاڍو
3.	Negation Adverb	نه، ڪونه
4.	Quantity Adverb	ڪيترا، گهڻا
5.	Affirmative Adverb	ها، پلي
6.	Pronoun Adverb	هيئن، هوئن

4.1.8. Articles (حرف). Articles also belong to closed word class. Examples of articles include: به، پڻ، ئي، ته.

4.2 Open Word Classes in Sindhi

Major open classes in most of the languages are: nouns, verbs, adjectives, adverbs and interjections. Sindhi also has these open classes. Following sections discuss open Sindhi classes in detail.

4.2.1. Nouns (اسم). Three categories of Sindhi Nouns are: proper noun (اسم خاص), common noun (اسم عام) and abstract noun (اسم ذات). Like all languages of the world, noun in Sindhi is open word class. Table-8 shows examples of three types of noun along-with newly included words in these classes.

Table 8: Types of Sindhi Nouns

S.No.	Noun Type	Example	Newly Included Word(s)
1.	Proper Noun	احمد، ڪراچي، پاڪستان	انٽرنيٽ، مائڪروسافٽ، نوڪيا
3.	Common Noun	ملڪ، شهر، ڪتاب	اي ميل، اٿارٽي، ڪمپيوٽر
4.	Abstract Noun	اچاڻ، ٿڏاڻ، شاهوڪار	ڪنيڪٽيوٽي

4.2.2. Adjectives (صفت). Adjectives define properties of nouns. Most of the native Sindhi adjectives belong to closed class as cardinals, ordinals, multipliers, fractals and pronominal adjectives discussed in section 4.1.6. But there are examples of adjectives which are adopted from other languages. For example: in the sentence "احمد هڪ اينيميٽيڊ ويب سائٽ ٺاهي" the word "اينيميٽيڊ" is an adjective which is adopted and is transliterated form of English word "animated". The word خوبصورت and شاندار are also examples of adopted words in Sindhi adjectives.

4.2.3. Adverbs (ظرف). Temporal, manner, negation, quantity, affirmative, and pronoun adverbs usually belong to closed word classes as discussed in section 4.1.7. Space, noun and adjective adverbs belong to open class types. As nouns and subset of adjectives themselves are open class types so is the case with noun and adjective adverbs. For example: in the sentence "آن لائن هليو اچ" the word "آن لائن" can be considered as space adverb or noun adverb.

4.2.4. Compound Verbs (مرڪب فعل). The only open class verb types in Sindhi are compound verbs. Compound verbs in Sindhi are formed by combining nouns, adjectives/adverbs and participles with verbs. Table-9 shows examples of compound verbs and newly formed compound verbs.

Table 9: Sindhi Compound Verbs

S.No.	Compound Verb	Description	Newly adopted Compound Verb(s)
1.	راند ڪرڻ، شادي ڪرڻ، عرض ڪرڻ	Formed by combining nouns with verbs	چيٽ ڪرڻ، اي ميل ڪرڻ
3.	خوش ٿيڻ، ناراض ڪرڻ	Formed by combining adjectives/adverbs with verbs	آن لائن ٿيڻ، آف لائن ڪرڻ
4.	ڪري پوڻ، بڪيو چڏڻ، ٻڌي سگهڻ	Formed by Combining verbs and participles	ڪنيڪٽ ٿيڻ

4.2.5. Interjection (حرف ندا). Interjections convey emotions in sentences. There are a small number of interjections used in Sindhi. Few examples are: واه واه، هاءِهاءِ!، افسوس!، گهوڙا!، ڪاش!، اڙي! and شاباس. However, new emotions can be included at any time therefore, interjections belong to open class type.

5. Developing a Sindhi POS Tagset

According to EAGLES (Expert Advisory Groups on Language Engineering Standards) [16] guidelines for morpho-syntactic tagging of languages three different levels of constraints may be considered for POS tagging are: obligatory, recommended and optional. Following sections discuss and present obligatory and recommended tagsets for Sindhi according to EAGLES guidelines.

5.1. Obligatory POS Tagset for Sindhi

EAGLES guidelines recommend only one attribute as obligatory tag for every grammatical category. Therefore, the obligatory POS tagset for Sindhi according to EAGLES guidelines will be as given in Table-10.

The NU (numeral) tag is used to tag numeric values in text, PU (punctuation) is for tagging punctuation marks, as punctuation marks are treated as words in POS tagging. The tag 'R' (residual) is used for foreign words or mathematical formulae.

Table 10: Obligatory Tagset for Sindhi

N	[noun]	V	[verb]	AR	[article]
PN	[pronoun]	AV	[adverb]	NU	[numeral]
PP	[postposition]	C	[conjunction]	PU	[punctuation]
AJ	[adjective]	I	[interjection]	R	[residual]

5.2. Recommended POS Tagset for Sindhi

The recommended POS tags for widely recognized grammatical categories of Sindhi according to EAGLES guidelines are discussed below. Appendix-A shows recommended attributes and their values for various word classes. At some places due to language specific features of Sindhi these recommendations may differ from EAGLES guidelines.

Noun (N) can have three different tags: common noun (NC), proper noun (NP), and abstract noun (NA). These basic tags can be further extended to recommended and optional tags as per EAGLES guidelines. For example: an intermediate tag for common noun, masculine, plural, nominative can be N1121, in which every position represents the attribute;

number at that position represents one of the attribute values given in Appendix-A. The intermediate tag can have an equivalent final tag NCmsn. This method can be extended for every attribute and value of noun class. However, this will result in 36 different tags for nouns which is theoretically more complete and useful but practically infeasible. Due to increase in number of tags automatic tagging of corpus will be exponentially complex as this increase directly affects the accuracy of tagging algorithms [17]. Therefore three basic tags for nouns are considered. However, according to attribute values of nouns given in Appendix-A there can be any number of possible tags. Table-11 shows proposed tags for nouns. The intermediate tags shown in the table are generated according to EAGLES guidelines from subtype attribute values of Appendix-A. For instance: in N2000; N refers to noun, 2 is type of noun (proper noun), and three 0's show that none of the gender, number and case attributes is considered for tagging.

Table 11: Proposed tags for Sindhi nouns

S.No.	[TAG] Word Class	Intermediate TAG	Example
1.	[NC] Common Noun اسم عام	N1000	هي گهر <NC> اسانجو آهي.
2.	[NP] Proper Noun اسم خاص	N2000	ڪراچي<NP> تمام وڏو شهر آهي.
3.	[NA] Abstract Noun اسم صفاتي	N3000	هي سيٺ <NC> شاهوڪار<NA> ماڻهو آهي.

Seven different types of pronouns and their basic tags are: personal pronoun (PP), demonstrative pronoun (PD), wh-pronoun (PWh), reflexive pronoun (PRF), relative pronoun (PRL), co- relative pronoun (PCR), and indefinite pronoun (PI). Personal pronoun (PP) is further divided into first person (PP1), second person (PP2) and third person (PP3) types. In Appendix-A attribute DemType at (vi) is used only for demonstrative pronouns, which are further divided into proximate (PDP) and remote (PDR) types and are proposed to be considered for basic tagset. Total ten different tags are proposed for pronouns. In Sindhi demonstrative pronouns are also used as third person personal pronouns and this ambiguity needs to be sorted out in tagging algorithm. Table-12 shows different tags for pronouns with examples.

Verb (V) attributes are divided into ten different categories. Auxiliary and main verbs are considered in tagset design. According to transitivity type main verbs are divided into: transitive (VT), intransitive (VI), casual transitive (VC), casual transitive double (VCTD) and casual transitive double twisted (VCTDT). All transitive and intransitive verbs are further divided into: active and passive types. These active and passive forms are handled via voice attribute in recommended attributes given in Appendix-A. Verb participles are divided into five different types and are tagged separately in the tagset. Tags of these five participles are: present participle (VPPrs), past participle (VPPst), future participle (VPFutr), verbal noun (VPVNn) and conjunctive participle (VPConj). Table-13 shows different verb type tags.

Table 12: Proposed tags for Sindh pronouns

S.No.	[TAG] Word Class Example
1.	[PP1] 1st Person Pronoun ضمير متڪلم آئون <PP1> هلان ٿو.
2.	[PP2] 2nd Person PN ضمير حاضر اوهين <PP2> اچو ها ته سٺو ٿئي ها.
3.	[PP3] 3rd Person PN ضمير غائب اهي <PP3> چوڪرا ڏاڍا تڪڙا آهن.
4.	[PDP] Demonstrative Pronoun Proximate ضمير اشارو ويجهو هيءُ <PDP> چوڪرو آهي.
5.	[PDR] Demonstrative Pronoun Remote ضمير اشارو ڏور هو <PDR> چوڪرو آهي.
6.	[PWh] Wh-Pronoun ضمير استهفام تنهنجو <PP2> هتي ڪهڙو<PWh> ڪم.
7.	[PRF] Reflexive Pronoun ضمير مشترڪ مون <PP1>پاڻ<PRF>هي<PDP>ڪم ڪيو آهي.
8.	[PRL] Relative Pronoun ضمير موصول جهڙي <REP>ڪرڻي تهڙي <CRP> پرڻي.
9.	[PCR] Co-relative Pronoun ضمير جواب موصول جهڙي <PRL>ڪرڻي تهڙي <PCR> پرڻي.
10.	[PI] Indefinite Pronoun ضمير مبهم ڪونه ڪو <PI> ته ايندو.

Adjective (AJ) tags can be divided into: characteristic (AJ1), cardinal (AJC), ordinal (AJO), pronominal (AJP), aggregate (AJA), quantifier (AJQ), fractal (AJF) and multiplier (AJM). Adjective types and their tags are shown in Table-14.

Adverb (AV) tags include: temporal adverb (AVT), space adverb (AVS), manner adverb (AVM), negation (AVNEG/NEG), quantity adverb (AVQ), affirmative adverb (AVA), noun adverb (AVN), adjective adverb (AVAJ) and pronoun adverb (AVP). Table-15 shows adverb types and their tags.

Two types of postpositions (prepositions are not used in Sindhi) are: simple and compound and have (PP) and (PPC) tags respectively. Table-16 shows postposition types and their tags.

Conjunctions are syntactically divided into: coordinate (CC) and subordinate (CS) types.

Interjection has tag (I) and is not further divided into any type. Conjunction and Interjection types with tags are shown in Table-17.

Table 13: POS tags for Sindhi verbs

S.No.	[TAG] Verb Class Example
1.	[VIA] Active Intransitive Verb فعل لازمي معروف آئون ٻوڙان <VIA> ٿو.
2.	[VIP] Passive Intransitive Verb فعل لازمي مجهول گاڏي ۾ چڙهجي <VIP> ٿو.
3.	[VTA] Active Transitive Verb فعل متعدي معروف مان خط لکان <ATR> ٿو.
4.	[VTP] Passive Transitive Verbفعل متعدي مجهول خط لکجي <PTR> ٿو.
5.	[VCTA] Active Casual Transitive Verb فعل متعدي بالواسطه معروف مان خط لکايان <ACT> ٿو.
6.	[VCTP] Passive Casual Transitive Verb فعل متعدي خط لکائجي<PCT> ٿو. بالواسطه مجهول
7.	[VCTDA] Active Casual Transitive Double Verb فعل مان خط لکارايان<DCT> ٿو متعدي بالواسطه ٻٽومعروف
8.	[VCTDP] Passive Casual Transitive Double Verb فعل متعدي بالواسطه ٻٽو مجهول خط <VCTDA> لکارائجي ٿو.
9.	[VCTDTA] Active Casual Transitive Double Twisted Verb فعل متعدي بالواسطه ٻٽودهرو مان خط لکارارايان< VCTDTA > ٿو
10.	[VCTDTP] Passive Casual Transitive Double Twisted Verb فعل متعدي بالواسطه ٻٽودهرومجهول خط لکارارائجي < VCTDTP > ٿو.
11.	[VAux] Auxiliary Verb فعل معاون مون خط لکيو آهي <VAux>
12.	[VPPrs] Present Participle اسم حاليه هوٻوڙندو < VPPrs > گهر ويو.
13.	[VPPst] Past Participle اسم مفعول اهو ڪتاب ڦاٽل < VPPst > آهي
14.	[VPFutr] Future Participle اسم استقبال مون کي خط لکڻو <VPFutr> آهي.
15.	[VPVNn] Verbal Noun اسم فاعل گاڏي هلائيندڙ<VPVNn> کي روڪيو.
16.	[VPConj] Conjunctive Participle اسم معطوفي استاد ٻارن کي پڙهائي <VPConj> گهر ويندو.

Table 14: Adjective types and tags

S.No.	[TAG] Word Class
1.	[AJ1] Characteristic Adjective صفت عارف سلڇڻو <AJ1> چوڪر آهي.
2.	[AJC] Cardinal عددي صفت اڪبر مون کان پنج <AJC>سئو اڏارا ورتا.
3.	[AJO] Ordinal قطاري صفت شمس ڊوڙ ۾ پنجون <AJO> نمبر آيو.
4.	[AJP] Pronominal Adjective ضميري صفت اسين <AJP> ماڻهو لاڙ جا.
5.	[AJA] Aggregative Adjective صفت مطلق اوهان مان هڪ به <AJA>انعام جي لائق ڪونهي.
6.	[AJQ] Quantifier صفت مقداري گهڻا <AJQ> ماڻهو گوڙ ڪندا.
7.	[AJF] Fractal عدد جا جزا پاءُ <AJF> کير وٺي آ.
8.	[AJM] Multiplier عدد جو پيرو تيل جي قيمت پنجوٽي <AJM> ٿي وئي.

Table 15: Adverb types and tags

S.No.	[TAG] Word Class Example
1.	[AVT] Temporal Adverb ظرف زمان احمد هينئر< AVT > آيو آهي.
2.	[AVS] Space Adverb ظرف مڪان هيٺ < AVS > نه ويهو.
3.	[AVM] Manner Adverb ظرف تميز هي همراه ڪم ڪار جو چڱو < AVM > آهي.
4.	[AVNeg] Negation Adverb ظرف نفي سليم ڪونه < AVNeg > ايندو.
5.	[AVQ] Quantity Adverb ظرف مقدار ماڻهو گهڻا < AVQ > هوندا ته گوڙ ٿيندو.
6.	[AVA] Affirmative Adverbظرف اثباتي تون ڀلي< AVA > هليو اچ.
7.	[AVN] Noun Adverb اسميه ظرف اڳيان< AVN > هليو اچ.
8.	[AVAJ] Adjective Adverb صفتي ظرف ڏاڍيان< AVAJ > نه ڳالهاءِ.
9.	[AVP] Pronoun Adverb ضميري ظرف پوءِ جيئن < AVP > توهين چئو.

Table 16: Postposition types and tags

S.No.	[TAG] Word Class
1.	[PP]Post Position حرف جر وڃي کڏ ۾ <PP> پئو.
2.	[PPC]Compound Post Position مرڪب حرف جر احمد اسان کانسواءِ <PPC> هليو ويو.

Table 17: Conjunction, Interjection types and tags

S.No.	[TAG] Word Class
1.	[CC] Coordinate Conjunction احمد ۽ <CC> سليم گڏجي آيا.
2.	[CS] Subordinate Conjunction ڏاڍو چيو مانس تنهن هوندي به <CS> هو ڪونه مڙيو.
3.	[I] Interjection حرف ندا ڪاش! <INJ> تون اچين ها.

Pronominal suffixes (PSx) are commonly used in Sindhi like few other south Asian languages. Due to complex nature, important morpho-syntactic structure and major role in semantics pronominal suffixes are considered as a different word class for POS-tagging purpose.

Three types of pronominal suffixes are nominal (PSxN), postpositional (PSxP) and verbal (PSxV). Other attributes are not considered in tagset being proposed; however, one can consider other attributes of Appendix-A and generate tags accordingly. Table-18 shows different types of pronominal suffixes and their tags.

Articles are assigned article (A) tag. Only few articles are there in Sindhi as discussed in section 1.1.8.

Residual (R) is considered for foreign words, formulas, acronyms etc. Residuals can be assigned many tags according to their types but only one tag (R) is considered for all residuals.

Numerals are numbers which occur in text. These numbers are tagged as numeral (NU) tags.

Other tags considered for proposed tagset include DATE, Title (divided into Pretitle (PRT) and posttitle (POT)) and sentence marker (SM). Table-19 shows various tags and examples including article, residual, numeral, title, sentence maker and date.

Table 18: Pronominal suffix types and their tags

S.No.	[TAG] Word Class Example(s)
1.	[PSxN] Pronominal Suffix with Nouns (Nominal Suffix) اسميه ضمير متصل پنم، پٽس، چاچهين
2.	[PSxV] Verbal Pronominal Suffix فعليه ضمير متصل اٿم، اٿئون، لڪندم
3.	[PSxP] Postpositional Pronominal Suffix جري ضمير متصل ڪين، ساٿس، وٽئون

Table 19: Miscellaneous Tags

S.No.	[TAG] Word Class	Example
1.	[AR] Article	به، ، ته، ئي، پڻ
	[R] Residual	IBM, ن.ڪ , Device
	[NU] Numerals	12, 445،٢٣
	[PRT]Pre title	محترم، جناب، سائين
	[POT]Post title	صاحب، خان
	SM	اٿي ته هلون.<SM>
	DATE	12/12/2011, 2012 فيبروري 13

As discussed earlier, POS tagset design considerations depend on the purpose for which tagset is being designed. The intension of proposed tagset usage is to tag a corpus with necessary tags, so that it can be used for syntactic analysis and parts of speech usage analysis in Sindhi text. Therefore, necessary level of granularity is considered. For example: only two tags for verbs are considered: main verb (V) and auxiliary verb (VAux); all verb types other than auxiliary are tagged as main verbs. Verb participles are also tagged separately as discussed in section 1.1.5. Compound verbs are not treated separately but their individual parts are separately tagged.

The proposed tagset is shown in Table-20. The tagset contains total 52 tags. The table shows proposed tags with intermediate tags; these intermediate tags are generated by using attributes and values given in Appendix-A as per EAGLES guidelines.

6. Conclusion

Discussion on various issues about the open and closed Sindhi word classes, tagset design issues in general, and Sindhi tagset design in particular, proposed tagset and attribute values of Sindhi word classes will provide basis for further research in various fields of Sindhi NLP. The proposed tagset discussed and presented provides basis for POS tagging and tagged Sindhi corpus construction. While designing the tagset necessary granularity level is considered to cope with the basic word level syntactic information and is therefore useful for parts of speech usage analysis in Sindhi corpus. The syntactic analysis of tagged corpus with these tags is also possible. The tagged corpus can also be used as training corpus for automatic grammar learning applications. By using the EAGLES guidelines and the word class attribute values of Appendix-A automatic tagset generation can also be implemented depending on the NLP application requirements.

Acknowledgement

Special thanks go to Dr. Ghulam Ali Allana and Dr. Hameedullah Kazi for their guidance, suggestions and valuable comments.

References

[1] J. A. Mahar, , G.Q. Memon. "Rule Based Part of Speech Tagging of Sindhi Language," in proc. *International Conference on Signal Acquisition and Processing, ICSAP 2010, Bangalore, India*, February 9-10, 2010, pp.101-106.

[2] J. A. Mahar, , G.Q. Memon. "Sindhi Part of Speech Tagging System using WordNet", *International Journal of Computer Theory and Engineering*. 2010. 2(4): 538-545.

[3] J.A. Mahar, H. Shaikh, A.R. Solangi. "Comparative Analysis of Rule Based, Syntactic and Semantic Sindhi Parts of Speech Tagging Systems". *International Journal of Academic Research*. 2011. Vol. 3. No. 5.

[4] M. Rahman. "Towards Sindhi Corpus Construction In *Linguistics and Literature Review* 63 Vol. 1 No 1, 2011 pp. 74-85. UMT Lahore Pakistan.

[5] S. Jan, ed. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press. 1990.

[6] L. Löfberg. , D. Archer., S. Piao., P., Rayson., T. McEnery., K., Varantola., J-P. Juntunen., "Porting an English semantic tagger to the Finnish language". In *Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University. 2003. pp. 457- 464.

[7] G. Leech., R. Garside and E. Atwell. *The automatic grammatical tagging of the LOB corpus*. Newsletter of the International Computer Archive of Modern English. 1983. 7, 13-33.

[8] M. P. Marcus, M. A. Marcinkiewicz, B. Santorini. “Building a large annotated corpus of English: the penn Treebank”, *Computational Linguistics*. 1993. v.19 n.2.

[9] A. Hardie. “Developing a tag-set for automated part-of-speech tagging in Urdu”. In *Archer, D, Rayson, P, Wilson, A, and McEnery, T (eds.) Proceedings of the Corpus Linguistics 2003 conference*. UCREL Technical Papers Volume 16. Department of Linguistics, Lancaster University, UK 2003.

[10] CRULP, *Urdu Part of Speech Tagset*. Center for Research in Urdu Language Processing. National University of Computer and Emerging Sciences. 2007. Lahore Pakistan.

[11] S. Manish. and P. Bhattacharyya. “Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge”. In *Proceedings of the International Conference on NLP (ICON08)*, Pune, India. 2008.

[12] I. Rabbi., M. A. Khan, and R. Ali. “Developing a tagset for Pashto part of speech tagging”. In *Proceedings of the International Conference on Electrical Engineering*.2008. pp. 1-6.

[13] Aglsoft. “Punjabi Part of Speech Tagger”. 2009. Retrieved (February 2012).
Available: http://http://punjabi.aglsoft.com/?show=tagger

[14] E. Trumpp. *Grammar of the Sindhi Language*. London-Leipzig. 1872.

[15] G. A. Allana “Sindhi boli jo tashrehi grammar (Descriptive Grammar of Sindhi Language)”. Sindhi Language Authority, Hyderabad, Pakistan. 2010.

[16] G. Leech., A. Wilson. *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora.* Istituto di Linguistica Computazionale, Pisa, Italy. 1996.

[17] Meyer, Chalres. F. *English corpus linguistics: An introduction*. Cambridge University Press. 2002. pp. 91-91.

Table 20: Proposed Sindhi POS Tagset

S.No.	TAG	Intermediate Tag	Word Class
1	NC	N1000	Common Noun
2	NP	N2000	Proper Noun
3	NA	N3000	Abstract Noun
4	PP1	P100010	1st Person Pronoun
5	PP2	P100020	2nd Person PN
6	PP3	P100030	3rd Person PN
7	PDP	P200001	Demonstrative PN Proximate
8	PDR	P200002	Demonstrative Pronoun Remote
9	PWh	P300000	Wh-Pronoun
10	PRF	P400000	Reflexive Pronoun
11	PRL	P500000	Relative Pronoun
12	PCR	P600000	Co-relative Pronoun
13	PI	P700000	Indefinite Pronoun
14	V	V1000000000	Verb
15	Vaux	V2000000000	Passive Intransitive Verb
16	VPPrs	V1000000101	Present Participle
17	VPPst	V1000000102	Past Participle
18	VPFutr	V1000000103	Future Participle
19	VPVNn	V10000002000	Verbal Noun
20	VPConj	V10000003000	Conjunctive Participle
21	AJD	AJ1000	(Characteristic) Adjective
22	AJC	AJ2000	Cardinal
23	AJO	AJ3000	Ordinal
24	AJP	AJ4000	Pronominal Adjectives
25	AJA	AJ5000	Aggregative Adjectives
26	AJQ	AJ6000	Quantifier
27	AJF	AJ7000	Fractal
28	AJM	AJ8000	Multiplier
29	AVT	AV1	Temporal Adverb
30	AVS	AV2	Space Adverb
31	AVM	AV3	Manner Adverb
32	AVNeg	AV4	Negation Adverb
33	AVQ	AV5	Quantity Adverb
34	AVA	AV6	Affirmative Adverb
35	AVN	AV7	Noun Adverb
36	AVAJ	AV8	Adjective Adverb
37	AVP	AV9	Pronoun Adverb
38	PP	PP1	Post Position
39	PPC	PP2	Compound Post Position
40	CC	CC1	Coordinate Conjunction
41	CS	CC2	Subordinate Conjunction
42	I	I	Interjection
43	PSxN	PS10000	Pronominal Suffix with Nouns (Nominal Suffix)
44	PSxV	PS20000	Verbal Pronominal Suffix
45	PSxP	PS30000	Postpositional Pronominal Suffix
46	AR	AR	Article
47	R	R0	Residual
48	NU	NU	Numerals
49	PRT	PRT	Pre title
50	POT	POT	Post title
51	SM	SM	Sentence Maker
52	DATE	DATE	DATE

Appendix-A

Sindhi Part-of-Speech Tags and Attributes

Noun (N) Attributes	**Value(s)**
(i) Type:	1. Common 2. Proper 3. Abstract
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural
(iv) Case:	1. Nominative 2. Oblique 3. Vocative

Verb (V) Attributes	**Value(s)**
(i) Type:	1. Main 2. Auxiliary
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural
(iv) Person:	1. First 2. Second 3. Third
(v) Transitivity	1. Transitive 2. Intransitive 3. Casual Transitive 4. Casual Transitive Double 5. Casual Transitive Twisted
(vi) Finiteness:	1.Finite 2. Non Finite
(vii) Participle Type:	1. Tense Participle 2. Verbal Noun 3. Conjunctive
(viii) Voice:	1. Active 2. Passive
(ix) Mood / Word Form:	1. Subjunctive 2. Imperative 3. Presumptive 4. Counter Factual
(x) Tense:	1.Present 2. Past 3. Future

Adjective (AJ) Attributes	**Value(s)**
(i) Type:	1. Descriptive 2. Cardinal 3. Ordinal 4. Pronominal 5. Aggregate 6. Quantifier 7. Fractal 8. Multiplier
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural
(iv) Case:	1. Nominative 2. Oblique 3. Vocative

Pronoun (PN) Attributes	**Value(s)**
(i) Type:	1. Personal 2. Demonstrative 3. Wh 4. Reflexive 5. Relative 6. Co-relative 7. Indefinite
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural
(iv) Case:	1. Nominative 2. Oblique
(vi) Person:	1. First 2. Second 3. Third
(v) DemType:	1. Proximate 2. Remote

Adverb (AV) Attributes	**Value(s)**
(i) Type:	1. Temporal 2. Space 3. Manner 4. Negation 5. Quantity 6. Affirmative 7. Noun 8. Adjective 9. Pronoun

Postposition (PP) Attributes	**Values(s)**	
(i) Type:	1. Simple	2. Compound

Conjunction (C) Attributes	**Value(s)**	
(i) Type:	1. Coordinate	2. Subordinate

Interjection (I)	

Pronominal Suffix Attributes	**Value(s)**
(i) Type:	1. Nominal 2. Postpositional 3. Verbal
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural
(iv) Case:	1. Nominative 2. Oblique 3. Agentive
(v) Tense:	1. Present 2. Past 3. Future

Article (AR)	

Punctuation Attributes	**Value(s)**
(i) Type:	1. Period 2. Comma 3. Semicolon 3. Colon 4. Dash (Long) Hyphen) 5. Hyphen 6. Ellipsis 7. Question Mark 8. Exclamation 9.Opening Inverted Comma 10. Closing Inverted 12. Opening Bracket 13. Closing Bracket

Residual (R) Attributes	**Value(s)**
(i) Type:	1. Foreign Word 2. Formula 3. Symbol 4. Acronym 5. Abbreviation 6.Unclassified
(ii) Gender:	1. Masculine 2. Feminine
(iii) Number:	1. Singular 2. Plural

Numeral (NU)	

Title Attributes	**Value(s)**
(i) Type:	1. Pre Title 2. Post Title

Date	

Sentence	

CLE Urdu Digest Corpus

Saba Urooj*, Sarmad Hussain*, Farah Adeeba*,
Farhat Jabeen**, Rahila Parveen*
* *Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science,*
*University of Engineering and Technology, Lahore, ** Islamia University Bahawalpur*
firstname.lastname@kics.edu.pk, farhat2iub@gmail.com

Abstract

The paper presents design scheme and details of the first large publically available corpus of Urdu language. This includes the collection and cleaning techniques for the first 100k derivative of the larger corpus and the issues related to corpus design such as size, genres along with their ratio. The same design and techniques are being scaled to develop larger derivatives of the corpus with 500k, 1000k and 5000k words. The corpus, due to its public license, will significantly contribute towards linguistic and computational aspects of Urdu analysis.

1. Introduction

In this paper, we present CLE Urdu Digest Corpus, which is a balanced, corpus of Urdu to promote the further research on Urdu linguistics and its computational modeling. Although there has been work published on Urdu Lexicon development based on much larger corpora i.e. 1.8 million words [1], however it is not publicly available due to licensing constraints. CLE Urdu Digest Corpus will be made publicly available through license agreement from Urdu Digest[1], a leading general interest magazine, with a history of 52 years of publication, with articles and stories covering a range of subjects including education, health, politics, international affairs, sports, business, humor and literature. CLE Urdu Digest corpus is collected from Urdu Digest published ranging from 2003-2011.

2. Literature Review

Corpus development criteria include corpus size, domains, target audience, genres and proportion of these genres. Bozkurt et al. [2] have suggested that corpus selection and collection decisions can be made by focusing the planned coverage of domains and sub categories. Additionally, Biber [3] has presented recommendations concerning representativeness, with general sampling frames including writing (published), writing (unpublished), speech and scripted speech.

One of the most widely used corpora of the English language; the Brown corpus comprises of one million words of written American English [4]. It is one of the earliest developed corpora, released in 1961, and has proved to be a guide for developing many other corpora such as Freiburg-Brown (Frown), Lancaster/Oslo Bergen (LOB) and FLOB (Freiburg-LOB). The corpus was divided into two components: informative and imaginative written American English. The informative component is further subdivided into the following categories: press, religion, skill trades/hobbies, popular lore, Belles La Hoes/biography/essays, government documents and learned and scientific writing. Furthermore, the imaginative component has been divided into fiction, romance/love-story and humor [5]. This is a balanced corpus as it covers a wide range of genres and text types.

Lancaster/Oslo Bergen (LOB) corpus is another English corpus which belongs to the Brown corpus family. It is also a balanced corpus. Just like Brown corpus, it consists of one million words from British English. The text domains used in this corpus are also modeled after Brown corpus. Both LOB and Brown corpora are important because they capture the trends in British and American written language respectively in 1961. However, these corpora have been further used as a guideline for developing Freiburg-Brown (Frown) and Freiburg-LOB (FLOB) corpora of English. Both of these corpora were released in 1991 and their purpose is to capture the differences in British and American English that had evolved between 1961 and 1991 [4].

The British National Corpus (BNC) consists of 2 million words. It has been divided into major domains: spoken and written texts. Each of these domains has

[1] http://www.urdudigest.pk

been divided into many sub-domains. The speech component is broken up in context dependent texts including fields such as leisure, business, educational, public/institutional and the demography related speech. Like the Brown corpus, the written text in the BNC has been split into two sub-domains: informative and imaginative. The informative component comprises of texts from pure sciences, applied sciences, belief and thought, commerce and finance, social science, world affairs, and leisure, covering 75% of the written component. The remaining 25% is covered by the imaginative fiction [5].

American National Corpus (ANC) has also been developed. This corpus is made up of 11 million words of written and spoken American English and was released in 2003 [6]. It is modeled after the BNC [4] and covers domains including email, essay, fiction, journal, letters, newspaper, non-fiction, spoken, court transcript, technical and travel.

Survey of English Usage corpus covers both written and spoken components of English language [7]. The written component is further divided into printed and non-printed text. Within printed text, instructional, informative and imaginative domains constitute the major categories. It is also one of the early corpora of British English released in 1960 by the University College London. The spoken component of this corpus was later used in the London-Lund corpus [7]. The London-Lund corpus is different from all the previously discussed corpora because it covers only the spoken component of British English.

Apart from these corpora based on the regional varieties of English, there have been attempts to develop an international corpus of English. These efforts culminated in the shape of the International Corpus of English, which has been divided into components of different regional varieties such as English in Great Britain (GB), America, Pakistan, India etc. The ICE-GB has been divided into spoken and written domains. Among the spoken constituent, there are dialogues and monologues whereas in the written part printed and non-printed texts have been included.

There are other corpora designed on the basis of population characteristics. One of them is the International Corpora of Learner English. It contains written text samples from 14 countries where English is used as a foreign or second language. Similarly, there is another learner corpora named The International Corpus Network of Asian Learners of English (ICNALE) which also comprises of samples of non-native writing in English. Based on its size, ICNALE is claimed to be one of the largest corpora of the English language [8]. Corpora like ICLE and ICNALE provide ample opportunities for research in the field of learner English and help in understanding the nuances of learner inter-language.

The attempts have also been made in developing corpus based lexicon for other languages as well. Alansary et al. [9] have presented a technical design for international corpus of Arabic language (ICA) that will cover Arabic language as is used all over the Arab world. They intend to collect the corpus from newspapers of different Arab countries. The corpus is collected from magazines, novels, net articles and academic sources. The paper also describes the importance of corpus in language studies. The ICA also contains a diverse range of written genres and sub-genres in some cases. This classification of genres includes strategic sciences, social sciences, sports, religion, literature, humanities, natural sciences, applied sciences, art and biography.

Weerasinghe et al. [10] have developed a corpus-based Sinhala lexicon of 10 million words drawn from diverse genres. The text is obtained from different online sources. The genres covered in the corpus are creative writing; technical writing and news reportage in which technical writing covered the highest percentage and creative writing covered the lowest percentage.

Baker et al. [13] developed publically available corpus of 96 million words under the EMILLE project. The corpus consists of three components: monolingual, parallel and annotated corpora. The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Gujarati, Punjabi and Urdu. The corpus has been translated from English.

As discussed earlier, corpus based Urdu lexicon of 19.3 million words has also been developed [1]. Text was collected from two news websites i.e. Jang and BBC. Data is collected from different domains for the purpose of ensuring diversity. These domains include sports, news, finance, culture/entertainment, consumer information and personal communication with their further categorization into sub-domains. In deciding on the corpus design, certain conventions have been followed; first of all each domain is represented by at least one million tokens, secondly no data is collected before the year 1990 as the time of appearance of a corpus does influence the extracted word lists and thirdly data from chat rooms has not been included.

3. The Process of Corpus Construction

A corpus seeks to represent language or some part of a language. So while deciding on corpus design, it is crucial to decide certain parameters, including the following.

- Text source
- Length of individual text samples
- Diversity among domains
- Time-frame for text selection

Further, in the construction of a corpus, it is essential to document the information about the author, the date of publication and information about the publisher (in our case it is Urdu Digest). The studies say that there should be some restriction in selecting the text from an individual article for the purpose of ensuring diversity of styles and authors. It has been argued that for written texts, one can include the first 2,000 words of an article, which contains the introduction and part of the body of the article, or one can take the middle of an article, which contains a significant amount of text developing the main point made in the article, or even its end [4]. The study also adds that not all samples need to be exactly 2,000 words i.e. a sample should not be broken off in mid-sentence but at a point (often over or just under the 2,000-word limit) where a natural break occurs. So it is more realistic to include text fragments in a corpus rather than complete texts. These fragments can be as short as 2,000 words, especially if there are frequently occurring grammatical constructions in the text.

Moreover, the range of genres to be included in a corpus is determined by whether it will be a multi-purpose corpus (a corpus intended to have wide uses) or a special-purpose corpus (a corpus intended for more specific uses, such as the analysis of a particular genre like scientific writing). In either case, the text needs to be from diverse sources to encompass variation across authors.

There are two types of corpora as far as time-frame is concerned. Synchronic corpora (i.e. corpora containing samples of text as it is presently spoken and written) contain texts created within a relatively narrow time-frame. In creating a synchronic corpus, the corpus compiler wants to provide an overview of contemporary language uninterrupted by language change. According to Mayer [5], time-frame of five to ten years is reasonable for the construction of a synchronous corpus. Diachronic corpora are used to study historical periods of a language.

The decision about time-frame for corpus design should be made before time i.e. before the collection of corpus. In the corpus based Urdu lexicon development [1], it has been ensured that text collected from two news websites i.e. Jang (www.jang.com.pk) and BBC (www.bbc.co.uk/urdu/) is not older than 2002 as the time of appearance of corpora has a large impact on the extracted word lists. The current data collected from Urdu digest is not older than 2003, so the corpus for the current work falls under the category of synchronic corpora. The reasons for this selection is that the corpus is designed to analyse and model and current use of Urdu language.

The corpus construction process has three phases, corpus acquisition, corpus organization and corpus cleaning

3.1. Corpus acquisition

As a first step, the data is gathered from Urdu Digest ranging between years 2003-2011. The data received in the format of Inpage[2] files. As Inpage uses its own encoding scheme, the data cannot be used for further processing. Due to this reason, the original files are converted into Unicode format. For this conversion, a third party utility is used. After the whole process of conversion, the converted files are analysed and matched with the original files to trace any unusual symbols generated or ignored during the conversion process. The following discrepancies are found between the original and converted texts.

3.1.1. Special symbols. Some special symbols fail to convert into Unicode, e.g. symbol of ◌ؐ. These are incorporated manually in the cleaning phase.

3.1.2. Garbage symbols. Certain symbols are added, e.g. ؟ ,Ñ,%,#. These symbols are removed by a cleaning utility.

3.1.3. Punctuation marks. Incorrect punctuation marks are detected during the cleaning process. The comma in the original files is written in the English form (i.e. ‘,’). It is replaced with the Urdu comma (‘،’). Moreover, the glossed words, proper nouns and direct speech are surrounded by an apostrophe from one side and by a comma on the other. Some examples are shown in Table 1. Such cases are also corrected.

Table 1: List of Glossed Words

Original	Modified
'بیماری،	'بیماری'
'اصول شفا،	'اصول شفا'
'سنٹرو تھراپی،	'سنٹرو تھراپی'
"آج میرے سر میں درد ہے۔ ،،	"آج میرے سر میں درد ہے۔"

[2] http://www.inpage.com/

3.2. Corpus organization

While designing a corpus, a number of considerations have to be taken into account including "the kinds of texts included, the number of texts, the selection of particular texts, the selection of text samples from within texts, and the length of text samples" [2]. CLE Urdu Digest Corpus is divided into two major categories, Informational (which covers 80% of the corpus) and Imaginative (which covers 20% of the corpus). The reason for taking a small percentage of imaginative text is that it contains figurative language, which is not good for computational modeling of the language. But the imaginative texts cannot be completely ignored as corpora need to represent language use. Therefore, a smaller percentage of imaginative part was kept, as is also the case in the BNC and the ICE. The Informational part includes texts from letters, interviews, press, religion, sports, culture, entertainment, health and science. The Imaginative part includes texts from short stories, novels, translation of foreign literature and book reviews. The data is distributed over 348 files whereby each file contains a minimum of 300 words, selected from the beginning or middle of the text. The corpus make-up is shown in Table 2.

3.3. Corpus cleaning

In the cleaning phase the errors of space, compound words, affixation and typological errors are removed. The details of these errors are given in table 2.

3.3.1. Typographical Errors. Typographical errors introduced during the conversion process are corrected manually. Examples include duplication of letters when *Tashdeed* diacritic (◌ّ) is found, deletion of word final *Noon Ghunna* letter (ں), etc. Where spellings are unclear, *Urdu Lughat*[3] (Urdu Dictionary) is used to confirm them. Some examples are given in the Table 4.

3.3.2. Compound words. Compound words in Urdu, can be written either with a space between them or without it. In the latter case, a Zero-Width-Non-Joiner (ZWNJ[4]) is needed to form the correct shape of the final letter of the words, in case it is a joining letter, e.g. the last row of Table 4. *Urdu Lughat* is used to resolve the ambiguity. When a compound word is found in this dictionary, it is written without a space, else with a space.

Table 2: Genres of CLE Urdu Digest Corpus

Category	Sub-category	Percentages
1. Informational (80%)		
a) Informal (20%)	Letters	10%
	Interviews	10%
b) Formal		
	Press	8%
	Religion	8%
	Sports	8%
	Culture (travel, history)	8%
	Entertainment	4%
	Health	8%
	Science (education, technology)	16%
2. Imaginative (20%)		
	Short Stories	8%
	Translation of foreign literature	4%
	Novels	4%
	Book reviews	4%

Table 3: Errors of letter insertion

Original	Modified
االله	الله
التتمش	التمش
می	میں

[3] Online version available at http://www.clepk.org/oud/

[4] The Zero-Width Non-Joiner (ZWNJ) is a Unicode character U+200C. ZWNJ is used to prevent joining.

Table 4: Examples of Compound Words

Compound with Space	Compound without Space (with ZWNJ if needed)
بے چین	بے چین
جدید و قدیم	جدیدوقدیم
طلبا و طالبات	طلباوطالبات

3.3.3. Reduplication. In case of reduplication, if the compound has been created with meaningful + meaningless word such as ٹھیک ٹھاک and سچ مچ it is written without a space (with ZWNJ if needed, as discussed) and if it is formed by repeating meaningful words, a space is inserted between them, e.g. آہستہ آہستہ and بار بار.

3.3.4. Loan words. Transliterated loan words are also written without a space (with ZWNJ where needed) as shown in Table 5. If multiple words are formed, they may also be written with a space between them, though that is not practiced at this time, as in the last row of Table 5.

Table 5: Examples Loan Words

Original	Modified
ٹیلیوژن	ٹیلی وژن
یونیورسٹی	یونی ورسٹی
پروٹون ایکسچینج میمبرین	پروٹون ایکسچینج میمبرین

3.3.5. Zer-Azafat/Hamza-Azafat. Urdu uses these diacritics for compounding of words (to show possessiveness or quality). Though this is a productive phenomenon in Urdu, many of these forms are also lexicalized. It is decided that lexicalized forms will be written without a space (with ZWNJ if needed) after consulting *Urdu Lughat*.

Table 6: Words with Zer-Azafat/Hamza-Azafat

Compounded Word
قواعدِانشا
سرِبالیں
ردِ عمل
اجرِآخرت

3.3.6. Abbreviations. Transliterated abbreviations of English are also found in the corpus. The abbreviations which should be treated as single word are written without space (with ZWNJ if needed). Otherwise, they are separated by a space. These examples are presented in Table 7 (a) and 7 (b).

Table 7 (a): Single Word Abbreviations

Original	Modified
ایس ڈی او	ایس‌ڈی‌او
اے ایم ڈی	اے‌ایم‌ڈی
ایم این اے	ایم‌این‌اے
آئی ایم ایف	آئی‌ایم‌ایف

Table 7 (b): Abbreviations with Multiple Words

Original	Modified
اے کے سومار	اے‌کے سومار
این این مارک	این‌این مارک
ڈی واے لی	ڈی‌واے لی
این بی سی فوکس	این‌بی‌سی فوکس

3.3.7. Affixation. Urdu corpus contains single words containing prefixes and/or suffixes separated with spaces from the root of the word. However, as they are inherently a single word, these spaces were deleted (and ZWNJ was inserted, where needed).

Table 8: Words with Affixes

Original	Modified
دست مال	دست‌مال
ہموزن	ہم‌وزن
مندرجہبالا	مندرجہ‌بالا

4. Results

The current paper presents the initial corpus developed for 102,209 words of Urdu. Domain-wise corpus size distribution is given in Table 9. A total of 83,450 words have been collected in the Informational domain, amounting to 81.6% of corpus. Additionally, 18,759 words are collected in the imaginative domain, forming 18.4% of the corpus.

A complete record of author, date and genre has been kept. It is ensured that the text sample is continuous without poetry and a variety of authors is selected for genres. The texts were saved in UTF-8 format.

5. Discussion and Future Work

After the initial corpus acquisition, the main challenge was to convert Inpage files into UTF-8 format. There are a number of converters available but process a single file at a time. For the conversion of multiple files at once, a batch process has been developed.

Moreover, when collecting individual text samples it was found that the corpus is not available as per the requirement of the decided percentage. For example, entertainment text samples are very rare in the available data of the Urdu Digest. This problem has been resolved by including more test from the category of culture containing the data of history and travel, as the text of travelogue mostly resembles with that of entertainment. Moreover, these two categories fall under the same sub-domain. Similarly, there is very limited text available in the category of news in Urdu Digest. This issue is resolved by re-distributing the text among the categories of news and editorials and including both of them in the category of Press.

Table 9: Domain-wise Corpus Distribution

Domains	No. Of Words	Distinct Words	%
Letters	10340	3048	10.1%
Interviews	10599	3010	10.3%
Press	9076	2884	8.8%
Religion	8753	2694	8.5%
Sports	8997	2672	8.8%
Culture	7789	2703	7.6%
Entertainment	4433	1805	4.3%
Health	8533	2551	8.3%
Science	14930	4397	14.0%
Short stories	6039	2091	5.9%
Novels	3791	1446	3.7%
Book reviews	4393	1775	4.2%
Translation of foreign literature	4536	1696	4.4%

For future work, CLE Urdu Digest Corpus will be extended to 500k, one Million and five million words, and more layers will be added to it e.g. POS-tagging in the first stage and sense-tagging.

6. Conclusion

Corpus development is divided into three phases including acquisition, organization and cleaning. Each phase has been described in detail. A total of 100k corpus with 348 text files has been created. It includes texts from multiple authors from the domains of letters, interviews, press, religion, sports, culture, entertainment, health, science, short stories, novels, book reviews and translation of foreign literature. This synchronous corpus has been collected from text produced after 2003.

7. Acknowledgements

We acknowledge Urdu Digest for generously contributing its text to develop this publically available corpus. We would like to thank the the German Academic Exchange Service (DAAD), for funding the project under the German-Pakistani Research Collaboration Scheme with grants from the Federal Ministry of Education and Research. Also, very special thanks to Mr. Asad Mustafa for his research assistance in corpus acquisition, and cleaning phases.

References

[1] M. Ijaz and S. Hussain, "Corpus based Urdu lexicon development", in proc. *Conference on Language Technology (CLT07)*, 2007, Retrieved (06, 25, 2012).
Available:
http://crulp.org/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf

[2] B. Bozkurt, O. Ozturk and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection", in Proc. *European Conference on Speech*, Geneva, 2003, Retrieved (06, 25, 2012).
Available:
http://tcts.fpms.ac.be/publications/papers/2003/eurospeech03_bbootd.pdf

[3] D. Biber, "Representativeness in corpus design", *Literary and Linguistic Computing, 8(4)*, 1993, Retrieved (06, 25, 2012) pp. 243-257.
Available:
http://staff.um.edu.mt/albert.gatt/teaching/dl/biber93.pdf

[4] F. Mayer, *Corpus Linguistics: Introduction, (1st edition)*, Cambridge University Press, 2002, Retrieved (06, 25, 2012).

[5] D.Y. Lee, "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle", *Language Learning & Technology*, 2001, Retrieved (06, 25, 2012).
Available: http://llt.msu.edu/vol5num3/pdf/lee.pdf

[6] N. Ide and K. Suderman, "The American national corpus first release", in proc. *LREC 2004*, 2004. Retrieved (06, 25, 2012).
Available: http://www.cs.vassar.edu/~ide/papers/anc-lrec04.pdf

[7] S. Greenbaum and J. Svartvik, "The London Corpus of Spoken English: Description and Research" In J. Svartvik (Ed.), *Lund Studies in English 82*, Lund University Press, 1990, Retrieved (06, 25, 2012).
Available:
http://khnt.hit.uib.no/icame/manuals/londlund/index.htm

[8] S. Ishikawa, "A New horizon in learner corpus studies: The aim of the ICNALE Project", In G. Weir, S. Ishikawa and K. Poonpon (Eds.*), Corpora and language technologies in teaching, learning and research*, (pp. 3-11), Glasgow, UK: University of Strathclyde Press, 2011, Retrieved (06, 25, 2012).
Available:
http://language.sakura.ne.jp/s/ilaa/ishikawa_20110921.pdf

[9] S. Alansary, M. Nagi and N. Adly, "Building an International Corpus of Arabic (ICA): Progress of Compilation Stage", *7th International Conference on Language Engineering*, Egypt, 2007, Retrieved (06, 25, 2012).
Available:
http://www.bibalex.org/isis/UploadedFiles/Publications/Building%20an%20Intl%20corpus%20of%20arabic.pdf

[10] R. Weerasinghe, D. Herath and V. Welgama, "Corpus-based Sinhala Lexicon", in proc. *7th Workshop on Asian Language Resources (ALR7)*, 2009, Retrieved (06, 25, 2012).
Available: http://aclweb.org/anthology-new/W/W09/W09-3403.pdf

[11] Urdu Dictionary Board. *Urdu Lughat*, Urdu Dictionary Board, Karachi, Pakistan.

[12] Baker, J.P., Hardie, A.,McEnery, A.M., Cunningham,H., and Gaizauskas, R. "Emille a 67-million word corpus of Indicl: data collection, markup, and harmonization". In proc. *3rd Language Resources and Evaluation Conference (LREC'2002)*, 2002, pages 819-825.

Developing Urdu WordNet Using the Merge Approach

Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain,
Asad Mustafa
*Center for Language Engineering, Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore
firstname.lastname@kics.edu.pk*

Abstract

The current paper describes the process of developing an Urdu WordNet. The process includes selecting words, identifying their senses and documenting their use. The current work also ties the Urdu senses with corresponding senses in English. Challenges in developing the WordNet and the solutions being implemented are discussed. Finally, this paper presents the work planned in the future.

1. Introduction

Fellbaum [1] defines WordNet as an extensive lexical database in which words are divided by part of speech and organized into a hierarchy of nodes. Each node represents a concept and words denoting the same concept are grouped into a synset with a unique ID, for example, ENG20-02853224-n: {car, auto, automobile, machine, motorcar}). Concepts are defined by a short gloss (e.g., 4-wheeled motor vehicle; usually propelled by an internal combustion engine) and are also linked to other relevant synsets in the database (e.g. hypernym: {motor vehicle, automotive vehicle}, hyponym: {cab, hack, taxi, taxicab}).

WordNet is used for many computational linguistic tasks such as Word Sense Disambiguation, Information Retrieval and Extraction and Machine Translation, etc. Over time, WordNet has become a valuable resource, which has initiated the development of WordNets for many other languages as well.

Urdu is a language of the Indo-Aryan family, widely spoken in Pakistan and India. It is written using Arabic script from right to left, in Nastalique writing style. Process for the development of Urdu WordNet has been discussed in this paper. The purpose of the development of Urdu WordNet is to provide a lexical resource for Urdu language that can be used in natural language processing. The WordNet is being developed specifically to align with linguistic, cultural, religious and other contexts in Pakistan.

The roadmap for the rest of paper is as follows: Section 2 presents the literature on Urdu WordNet. Methodology for development of Urdu WordNet is described in Section 3 and the current status is discussed in Section 4. Section 5 discuses the relevant issues and solutions, and Section 6 concludes the paper.

2. Literature Review

WordNets in various languages have been developed both through manual [2, 13] and automated [3, 14] methods. The manual construction of each WordNet is more accurate, but is also more time-consuming and expensive. There are two common approaches for building a WordNet for a language [4]: (i) a top-down approach, using an existing WordNet in a source language to seed the linguistic data for the target language WordNet [4], and (ii) a bottom-up approach, where the linguists create the WordNet synsets without depending on an existing one [5].

In the top-down approach, the synsets from the source language are translated into the target language. However, for the synsets to be mappable, concepts in the source language must exist in the target language, which is not always possible. Additionally, generally a significant amount of language resource is required for building a WordNet. For example, a set of synsets strictly aligned with the source WordNet must exist before the new WordNet can be built. This is a significant drawback of building a WordNet from an existing one. For this approach to be successful there must be significant level of linguistic similarity between the two languages [5, 6].

Two methods have been discussed for developing a WordNet through the bottom-up process: the merge approach and the expand approach [7]. The merge approach builds the taxonomies of the language, synsets and relations, and then map to the Princeton WordNet (PWN) by using the English equivalent words from existing bilingual dictionaries [15]. Merge approach provides a description of lexico-semantic relations, closer to the spirit of the given language, in that it is less influenced by the design decisions in a WordNet for another language, often of a significantly different type. The merge approach, however, requires rich resources at the outset, for example, a monolingual dictionary with senses identified, detailed definitions, thematic codes for senses and some semantic structuring [15].

The expand approach is to map or translate local words directly to the PWN's synsets by using the existing bilingual dictionaries. Thai WordNet construction has used the expand approach due to budget and time reasons [7].

Previous work on Urdu WordNet [8, 9] is based on the top-down approach. Hindi WordNet (HWN) has been used due to its similarity with Urdu. However, this method faced the following challenges [8].

- There are number of Hindi words that are not used in Urdu due to the linguistic, religious, cultural and other differences, e.g. انتیرن (fail) is not normally used in Urdu.
- Many words which are commonly used in Urdu, e.g. those loaned from Arabic and Persian languages, are not present in Hindi WordNet synsets. For example ربا (interest) is used in Urdu but not available in HWN.
- In the explanation given for the synset and the example for its usage a lot of Hindi words are used, which are not part of the common cultural vocabulary of Urdu in Pakistan. For example in the sentence. ارجن پریکشا میں کامیاب رہا. ارجن and پریکشا are not commonly used.

  In addition, the compound words and complex predicates in verbs are not addressed.

3. Methodology

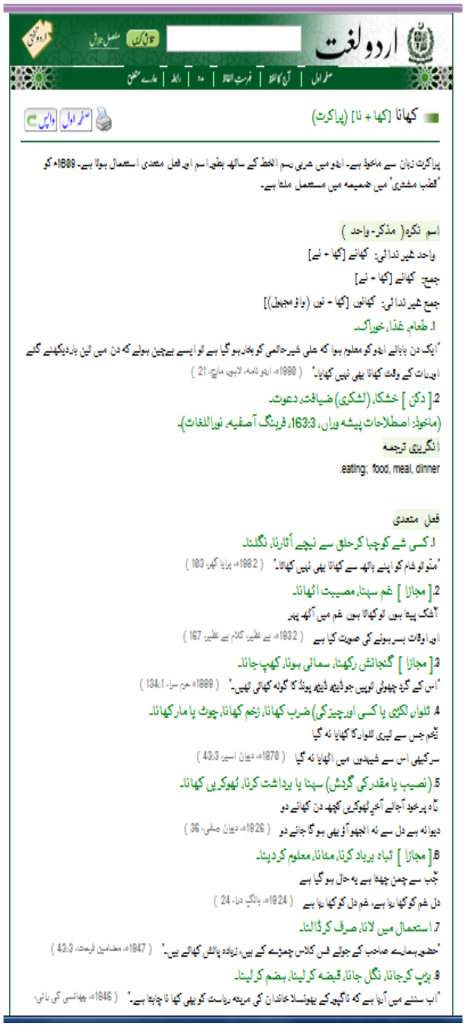
To build Urdu language WordNet merge approach has been used. 5000 high frequency nouns, verbs, adjectives and adverbs are selected from Urdu corpus [10] to develop the WordNet. The following process is used for the development of Urdu WordNet.

1. A word from the list of 5000 words is looked up into Urdu Lughat [11]
2. Its POS tag is determined by Urdu Lughat. For example the word کھاناwhich has two POS tags in Urdu Lughat i.e. کھانا (meal) a noun and کھانا(eat) a verb.
3. The number of senses for each POS of the particular word is determined from Urdu Lughat. The less common, literary and poetic senses are ignored. So the number of senses for each word varies according to its use. For example, the third sense is in Table 1 below is less common and poetic, and thus ignored.

Table 1: Urdu Word Senses

Concept	Sense	English Translation
پکڑا ہوا، قیدی ، محبوس	گرفتار	Capture
مبتلا ، پھنسا ہوا، گھرا ہوا	گرفتار	Entangled
عاشق، فریفتہ	گرفتار	Smitten

4. The English translation of the word according to its POS tag is looked up in Urdu to English Dictionary. If there are two or more POS tags of the word in Urdu Lughat then the English translation of the word is determined according to all its tags as the word کھانا (meal) is a noun as well as a verb کھانا (eat) . So both the categories will be created. Figure 1 shows different POS categories of the word کھانا.


اردو لغت
تلاش کریں
مفصل تلاش
اردو تحقیق
صفحہ اول | آج کا لفظ | فہرست الفاظ | مدد | رابطہ | ہمارے متعلق
صفحہ اول
واپس

کھانا [کھا + نا] (پراکرت)

پراکرت زبان سے ماخوذ ہے۔ اردو میں عربی رسم الخط کے ساتھ بطور اسم اور فعل متعدی استعمال ہوتا ہے۔ 1609ء کو "قطب مشتری" میں ضمیمہ میں مستعمل ملتا ہے۔

اسم نکرہ (مذکر - واحد)

واحد غیر ندائی: کھانے [کھا + نے]

جمع: کھانے [کھا + نے]

جمع غیر ندائی: کھانوں [کھا + نوں (واؤ مجہول)]

1. طعام، غذا، خوراک۔

"ایک دن بابائے اردو کو معلوم ہوا کہ علی شیر خانی کو بخار ہو گیا ہے تو ایسے بے چین ہوئے کہ دن میں تین بار دیکھنے گئے اور رات کے وقت کھانا بھی نہیں کھایا۔" (1990ء، اردو نامہ، لاہور، مارچ، 21)

2. [دکن] خشکا، (لشکری) ضیافت، دعوت۔

(ماخوذ: اصطلاحات پیشہ وراں، 163:3، فرہنگ آصفیہ، نوراللغات)۔

انگریزی ترجمہ

.eating; food, meal, dinner

فعل متعدی

1. کسی شے کو چبا کر حلق سے نیچے اُتارنا، نگلنا۔

"منو تو شام کو اپنے ہاتھ سے کھانا بھی نہیں کھاتا۔" (1982ء، پرایا گھر، 103)

2. [مجازا] غم سہنا، مصیبت اٹھانا۔

اشک پیتا ہوں تو کھاتا ہوں غم میں آٹھ پہر
اب اوقات بسر ہونے کی صورت کیا ہے (1932ء، بے نظیر، کلام بے نظیر، 167)

3. [مجازا] گنجائش رکھنا، سمائی ہونا، کھپ جانا۔

"اس کے گرد چھوٹی توپیں جو ڈیڑھ ڈیڑھ پونڈ کا گولہ کھاتی تھیں۔" (1899ء، حرم سرا، 134:1)

4. تلوار، لکڑی یا کسی اور چیز کی) ضرب کھانا، زخم کھانا، چوٹ یا مار کھانا۔

زخم جس سے تیری تلوار کا کھایا نہ گیا
سر کبھی اس سے شہیدوں میں اٹھایا نہ گیا (1870ء، دیوان اسیر، 43:3)

5. (نصیب یا مقدر کی گردش) سہنا یا برداشت کرنا، ٹھوکریں کھانا۔

راہ پر خود آجائے آخر ٹھوکریں کچھ دن کھانے دو
دیوانہ ہے دل سے نہ الجھو آؤ یہی ہو گا جانے دو (1926ء، دیوان صفی، 36)

6. [مجازا] تباہ برباد کرنا، مٹانا، معدوم کر دینا۔

جب سے چمن چھٹا ہے یہ حال ہو گیا ہے
دل غم کو کھا رہا ہے، غم دل کو کھا رہا ہے (1924ء، بانگ درا، 24)

7. استعمال میں لانا، صرف کر ڈالنا۔

"حضور ہمارے صاحب کے جوتے فس کلاس چمڑے کے ہیں، زیادہ پالش کھاتے ہیں۔" (1947ء، مضامین فرحت، 43:3)

8. ہڑپ کر جانا، نگل جانا، قبضہ کر لینا، ہضم کر لینا۔

"اب سننے میں آیا ہے کہ ناگپور کے بھونسلا خاندان کی مرہٹہ ریاست کو بھی کھانا چاہتا ہے۔" (1846ء، جھانسی کی رانی،)

Figure 1: POS Cat. of کھانا in Urdu Lughat

5. English translation of an Urdu word may be different for its multiple senses. So the English translation of each sense is looked up separately in Urdu to English Dictionary. The example is explained in the Table 2.

Table 2: English Translation of Urdu Word

English word	Concept of each sense	Urdu Word
Work	کسبِ معیشت کا وسیلہ یا ذریعہ	کام
Chores	روزمرہ یا مقررہ وقت کا کام	کام
Concern	سروکار یا واسطہ ہونا	کام
embroidery	کڑھائی، نقاشی وغیرہ کا کام	کام

6. The selected word is looked up in Princeton WordNet version 2.1 and each sense of Urdu is mapped on the sense of English according to its determined POS tag. The unique ID of English sense and its English word is recorded in separate columns. Table 3 shows the unique ID of English sense.

Table 3: Unique IDs of English Senses

English ID	English Word
578942	Work
708623	Chores
5600606	Concern
3248411	Embroidery

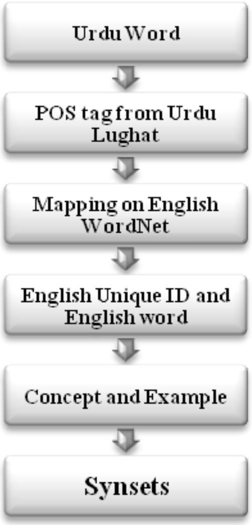

Urdu Word
POS tag from Urdu Lughat
Mapping on English WordNet
English Unique ID and English word
Concept and Example
Synsets

Figure 2: Urdu WordNet Process

7. The concept of each sense is explained with the help of Urdu Lughat in simple and precise language.
8. Further, an example is given to illustrate the concept, using a word from the synset. For formulating the example, as a first preference the example usage given Urdu Lughat is used. If this example is difficult to understand, a new example sentence is created. Where it is not easily possible, the corresponding example from PWN is translated as an alternative.
9. The synsets of the word are written from Qamos-e-Mutradifat (synonyms dictionary) [12]. Only those synonyms from Qamos-e-Mutradifat are selected that have the same concept. The concepts of these synonyms are confirmed from Urdu Lughat.

10. In the end, a linguist reviews the WordNet entries.

This process is summarized in the Figure 2.

4. Current Status

A sample Urdu WordNet entry is given in the table below.

Table 4: Urdu WordNet Entries

Synsets	دباؤ، بوجھ، بھار، بار، ثقل	دباؤ، سختی، جبر	دباؤ، خوف، ڈر، دہشت
Urdu ID[1]	1	2	3
Category	N	N	N
Concept	کسی چیز کا بوجھ یا وزن	سختی یا جبر کرنے کا عمل	خوف یا دہشت ہونا
Example	اس نے میز پر دباؤ ڈالا تو وہ ٹوٹ گئی	بچوں پر غیر ان ضروری دباؤ کو والدین سے متنفر کر دیتا ہے	وڈیرے کے دباؤ کی وجہ سے وہ کچھ نہ بولا.
English ID	11329024	416551	7418507
English Word	Pressure	Oppression	terror

At present, 2205 senses are completed. These include 1518 nouns, 560 adjectives, 80 verbs and 47 adverbs.

5. Discussion

This paper presents experience of building Urdu WordNet. Although it gives sufficient lexical information of Urdu words but still there are issues needed to be resolved. Some language specific challenges are observed during the development of Urdu WordNet process that are needed to be considered carefully. The diacritics need to be handled for Urdu. The words that change their meaning with the diacritics need to have a separate entry in Urdu WordNet. Table 5 shows the example. This is addressed in the Urdu WordNet.

Table 5: The Case of Diacritics

Urdu	Concept	English
گَنّا	بانس کے درخت کی وضع کا پودا جو	sugar cane
گِنا	گننا، شمار کرنا	count

There are Urdu words/concepts that do not exist in the English WordNet due to religious, cultural and other differences. Some examples are given in Table 6.

Table 6: The Case of Cultural Concepts

Words	Concept
صفر	name of the second Islamic month
مہندی	a cultural function which is celebrated before the marriage ceremony in which typical intricate patterns of Henna are applied to bride, celebrated mainly by the bride's family
ڈوپٹہ	a long scarf that is worn by females to cover their head

This difference creates problem when Urdu synset is mapped onto English ID.

Further, because of the difference in the structure of English and Urdu language it is difficult to map some of the words on the same POS tag. For example the word قیدی “prisoner” is a noun in English but Urdu Lughat lists it as an adjective. صارف “consumer” is a noun in English and an adjective in Urdu. Similarly the word پولنگ “polling” is a noun in Urdu and a verb in English. In order to incorporate this problem, there is need to improve Urdu Lughat.

Sometimes two different words are mapped on the same English ID, to avoid this problem and keep all the IDs unique that particular word is added into the synset of the previously added word.

In the future, 5000 senses will be completed. Currently nouns are more in number than other categories. The words added in the future will be

[1] This is an arbitrarily assigned number, which will be finalized upon release of Urdu WordNet.

selected from other categories as much as possible, to balance this distribution. Further the work will associate these synsets, to allow for more significant modeling of the semantic relationships.

6. Conclusion

In this paper, we present the process of developing a basic lexical resource for Urdu. This lexical resource is developed using the bottom-up approach. A few language and cultural issues faced in its development are discussed. This is a work in progress and future goals are also presented.

7. Acknowledgements

This work has been supported through a grant from DAAD Germany, and is being done in collaboration with University of Konstanz, Germany.

References

[1] C. Fellbaum, “WordNet: An Electronic Lexical Database.” MIT Press, Cambridge, Massachusetts, 1998.

[2] C. Fellbaum, M. Palmer, L. Delfs, S. Wolf, “Manual and Automatic Semantic Annotation with WordNet”, 2001, Retrieved (06, 27, 2012). Available at: https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2001/naacl/mwnw/pdf/invitedPaper.pdf

[3] M. Montezory and Heshaam Faili “Automatic Persian WordNet Construction” Coling 2010: Poster Volume, pages846–850, Available at: http://aclweb.org/anthology/C/C10/C10-2097.pdf

[4] M. Khan and F. Faruqe. “*BWN- A Software Platform for Developing Bengali WordNet*”, Center for research on Bangla language processing (CRBLP), BRAC University 2010, Retrieved (06, 27, 2012). Available at: http://crblp.bracu.ac.bd/papers/2008/BWN-architecture-CISSE08.pdf

[5] D. Fiser, “A Multilingual Approach to Building Slovene WordNet” In Proc. *Workshop on A Common Natural Language Processing Paradigm for Balkan Languages held within the Recent Advances in Natural Language Processing Conference RANLP'07*. Bulgaria, 2007.

[6] E. Barbu and V. B. Mititelu, “Automatic Building of WordNets" In Proc. *International Conference Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2005, Retrieved (06, 27, 2012). Available at: http://clic.cimec.unitn.it/eduard/publications/AutomaticBuildingWordnets.pdf

[7] S. Thoongsup, K. Robkop, C. Mokarat, T. Sinthurahat, T. Charoenporn, V. Sornlertlamvanich and H. Isahara, “Thai WordNet Construction” In. Proc. *7th Workshop on Asian Language Resources, ACL-IJCNLP*, 2009, pages 139–144, Retrieved (06, 27, 2012). Available at: http://aclweb.org/anthology/W/W09/W09-3420.pdf

[8] F. Adeeba and S. Hussain, “Experiences in Building the Urdu WordNet”, *IJCNLP*, 2011, Retrieved (06, 27, 2012). Available at: http://www.cle.org.pk/Publication/papers/2011/UrduWordNet.pdf

[9] T. Ahmed and A. Hautli, “Developing a Basic Lexical Resource for Urdu using Hindi WordNet”, in proc. *CLT10*, Islamabad, 2010, Retrieved (06, 27, 2012). Available at: http://ling.uni-konstanz.de/pages/home/pargram_urdu/main/files/Ahmed_Hautli_CLT10.pdf

[10] M. Ijaz and S. Hussain, “*Corpus Based Urdu Lexicon Development*”, Centre for Research in Urdu Language Processing, National University of Computer and Emerging Sciences, Lahore, 2007, Retrieved (06, 27, 2012). Available at: http://cle.org.pk/Publication/papers/2007/corpus_based_urdu_lexicon_development.pdf

[11] *Urdu Lughat*, Urdu Lughat Board Karachi, 2002, Retrieved (06, 27, 2012). Available at: http://www.clepk.org/oud/Default.aspx

[12] W. Sirhindi, *Qamoos-e-Mutradifaat*, Urdu Science Board, Lahore, 2006.

[13] P. Sathapornrungkij, C. Pluempitiwiriyawej, “*Construction of Thai WordNet Lexical Database from Machine Readable Dictionaries*”, Mahidol University, Bangkok, 2005, Retrieved (06, 27, 2012). Available at: http://www.mt-archive.info/MTS-2005-Sathapornrungkij.pdf

[14] M. Saveski, I, Trajkovski, “*Automatic Construction of Wordnets by Using Machine Translation and Language Modeling*”, Staffordshire University, UK, 2010, Retrieved (06, 27, 2012). Available at: http://www.time.mk/trajkovski/papers/is2010.pdf

[15] M. Piasecki, S. Szpakowicz, B. Broda *“A Wordnet from the Ground Up”*, Oficyna Wydawnicza Politechniki Wroclawskiej, Wroclaw, 2009, Retrieved (24, 8, 2012). Available at: http://www.site.uottawa.ca/~szpak/pub/A_Wordnet_from_the_Ground_Up.pdf

An Acoustic Study of Vowel Nasalization in Punjabi

Saira Zahid[1,] Sarmad Hussain[2]
Government College University Faisalabad[1], Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science,University of Engineering and Technology, Lahore[2.]
sara.linguistics@gmail.com[1], sarmad.hussain@kics.edu.pk[2]

Abstract

Nasalization is a very prominent but less understood feature of many languages spoken in Pakistan. This paper compares the contextual and contrastive nasalization phenomenon in Punjabi vowels. The degree and direction of nasalization is determined using acoustic measures. The results depict that contrastively nasal and contextually nasal vowels show almost the same degree of nasality except /ɪ/ vowel. For the latter, both the anticipatory and preservatory nasalization are observed in Punjabi.

1. Introduction

Punjabi language is a member of the Indo-Aryan family. Primarily, Punjabi is spoken in India and Pakistan but there are speakers of Punjabi in East Africa, United Kingdom and Canada as well. Nearly forty five million people use this language either as their first or second language. [1]

Many regional dialects of Punjabi are used y its speakers. The major dialects of Punjabi are Majhi, Malwi, Doabi and Powadi. Furthermore, Rathi, Ludhianwi, Patialwi, Bhattani are also some traditionally recognized Punjabi dialects whose independent status as dialect is in question [2]. These dialects of Punjabi differ from each other on the basis of distinct variation in their phonemic inventories [1].

Punjabi is written in three scripts which are Gurmukhi, Perso-Arabic and Devanagari scripts. Hindus in India write Punjabi in Devanagri script; Sikhs in India write in gurmukhi script while in Pakistan, Punjabi is written in shahmukhi (Perso-Arabic) script. [2]

The present paper aims to report the trend of vowel nasality in the speech of Majhi speakers of Punjabi living in Lahore.

2. Review of Literature

The oral sounds are produced with the complete closure of nasal tract, whereas the nasal sounds are produced with open velopharyngeal port. The phenomenon of vowel nasalization exists in almost all the languages of the world [3]. But the level of velopharengeal port's opening varies from language to language and from speaker to speaker.

All the languages of the world have oral vowels, but there are some languages which have nasal vowels as well. French, Taiwanese, Urdu, Punjabi, etc. are the examples of languages which have oral-nasal contrast in their vowel system. The nasal vowels are never observed to be greater in number than the oral vowels in any language [17]. Other than contrastive nasalization, there are also some languages which have contextual nasalization e.g. English, where the presence or the absence of nasality feature in the vowel does not change the meaning of the word [8].

During the production of vowels with neighboring nasal consonants, the languages with contrastive vowels restrict the level of velum lowering and make vowels less nasalized than the languages which lack this oral-nasal contrast. The velum lowering is restricted to maintain oral- nasal contrast and to avoid the contextual nasalization. Herbert [18] reports that only the languages which have oral- nasal contrast for vowels have this pattern of velum lowering restriction for oral vowel production in context of nasal sounds. Furthermore, Manuel [5] illustrates that the contrast of nasality in vowels and the degree of coarticulation are correlated inversely.

Cohn [4] reports a higher degree of contextual vowel nasalization before a nasal consonant in English, a language which does not contain nasality contrast in its vowels. It may be compared with French, which has nasal-oral contrast for vowels.

Ladefoged et al. [8] describe that the vowel nasalization phenomenon exists in all the dialects of

English language. In English, vowels tend to assimilate with the nasal consonants whenever they occur in nasal context. They illustrate the example of the English word “man”. In such circumstances where a vowel is followed or proceeded by the nasal sound, all the vowels become completely nasalized. So in English vowels are nasalized because of the phonetic context. Vowels in oral context never adopt nasality feature except in the disordered speech.

Languages having contextual nasalization or contrastive nasalization or even containing both types of nasalization differ from each other because of different nasality patterns. There is evidence that the languages which lack oral/nasal contrast for vowels show extensive degree of nasalization. English language is a good example of heavy nasalization of vowels in nasal context.

Furthermore, Delvaux et al. [14] describe that the languages which have oral/nasal contrast for their vowels may limit the degree of contextual vowel nasalization in both high and low vowels, in order to maintain the oral/nasal contrast between vowels. French allows an extensive degree of contextual nasalization for the high oral vowels as all the nasal vowels are mid-low and low in French. So the vowels which have oral and nasal contrast show lesser degree of nasal coarticulation than the vowels which have no nasal counterpart.

Moreover, Kawasaki [16] studies the degree of nasalization between Taiwanese contrastively and contextually nasalized vowels. He states a greater degree of nasalization in contrastive environment (nasal vowel) in comparison with non contrastive environment (contextually nasalized vowels).

On the other hand, Al-Bamerni discusses the extensive degree of velopharyngeal opening for the high back vowels in Gujarati and Hindi, the languages which have contrastive nasality in their vowel systems (as cited in [14]). This asymmetry between the degrees of nasalization among various languages suggests that the extent of nasal coarticulation is not dependent on the phonemic inventory of languages. Different languages have different patterns of nasalization for vowels regardless of the presence and absence of oral/nasal contrast for vowels.

The study of vowel nasalization is very complex because of the variation in the exact acoustic characteristics of nasalization among speakers. The acoustic characteristics of nasalization are difficult to examine due to the changes in the anatomical structure of the nasal cavity, vowel quality, and also because of the degree of oral and nasal tract’s coupling. [9]

The vowels are nasalized because of the nasal and oral tract’s configuration. The more the velum lowers; the heavier the degree of vowel nasalization. So, this variation in configuration between oral and nasal tract introduces change in spectrum at transition between the vowel and the nasal consonantal sounds [12]. These acoustic effects are transformed in spectra through introducing nasal poles and zeroes in the region of first formant (F1) and also the shift of vowel formants (especially F1).

Various acoustic effects of vowel nasalization are explored through multidimensional ways. Ladefoged et al. [8] report that the vowels which have extra nasality feature are distinguished with reduction in intensity of the first formant (F1) and increase in third formant (F3). This reduction in the intensity is because of the diversion of acoustic energy from the oral cavity to the nasal cavity. There is evidence from the perception based experiments that the reduction in F1 amplitude by 6-8 db is necessary to get a significant level of nasalization perception [10]. But later studies do not support this assumption providing the view that the degree of F1 amplitude’s lowering is somehow language and speaker specific. As Chen [7] reports the results of her study on nasalization, the degree of F1 amplitude varies among English speakers and the French speakers. So there is lack of any fixed measure of the lowering of F1 amplitude.

Furthermore, the flattening of spectral region is also studied as an indicator of nasality. Maeda [11] has studied spectral variations analyzing 11 French vowels. He reports that the diversion of energy from oral to nasal tract flattens the spectral region between 300 Hz and 2500 Hz. Similarly, Stevens [15] reports that the widened first formant (F1) and the overall reduced vowel amplitude is the indicator of the presence of nasality feature in a vowel.

Fant [13] also illustrates that the nasalized vowel has “a distortion superimposed on the vowel spectrum” which is significant by the nasal effect on harmonics in the region of low frequencies (below F1) (p. 156). Similarly, Beddor et al. [6] describe that the vowels with nasality feature have broader and flatter spectral prominence in the region of low frequency (below F1).

Chen [7] has introduced an acoustic approach for the measurement of nasality in her study of nasalized vowels of French and English. She finds the reduction of first formant as the primary cue of nasalization in vowels. She has distinguished nasalized vowels of French and English successfully, employing the two parameters which are A1-P0 and A1-P1. Here A1 is the amplitude of the first formant (F1), P0 is the amplitude of first nasal peak below the first formant (F1) and P1 is the measure of the amplitude of nasal peak between first formant (F1) and the second formant (F2) of the vowel. So, the results of her study confirm that the amplitude of F1 in nasalized vowel reduces relative to its amplitude in oral vowel, and the extra nasality peaks are also noticed. These measured acoustic parameters of nasalization in nasal and oral vowel are given below.

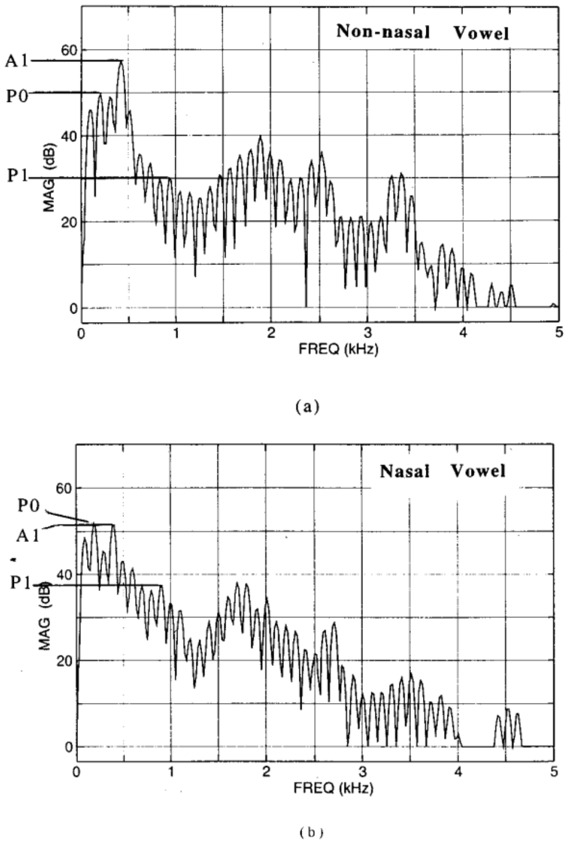


Figure 1: The measurements of A1, P1 and P0
Adapted from Chen [7]

Chen [7] has measured vowels at initial, medial and final positions. But she has not observed any differences among the measures taken at these three points in vowels. So she has averaged these measures across three points and has got results.

Punjabi is one of those languages which have oral-nasal contrast in their vowel system. There are ten oral vowels in Punjabi. It has three short /ɪ, ə, ʊ/ and seven long /ɪ, e, æ, a, ɔ, o, u/ vowels. All these oral vowels have their nasal counterparts as well [19]. This study is an attempt to explore nasalization phenomenon in Punjabi. This aims to determine the degree and direction of nasalization in both contrastive and non-contrastive environments.

3. Methodology

For this study, the participants with low fundamental frequency have been selected, so that the harmonics can be traced in spectrum accurately. Three male speakers of Majhi dialect from Lahore, with Punjabi L1 are selected for the present study.

The data consists of four syllable types; CVC, CṼC, CVN, NVC. The first two syllable types are chosen to measure the degree of nasality in contrastive environment while the other two are used to study the degree of nasality in non-contrastive anticipatory and preservatory environments respectively.

In the CVN and NVC contexts the N is /n, m, ŋ/. The vowels in CVC, CVN and NVC syllable types are /ɪ/, /æ/, /a/ and /ʌ/. While in CṼC context, the vowels are /ĩ/, /æ̃/, /ɑ̃/ and /ʌ̃/. All the words are embedded in a carrier phrase for recording:

میں -------- کیا

/mæn ______ kea/

“I said _______”

The detailed list of tokens is given in appendix A. Three repetitions of each word have been recorded. The acoustic measures A1-P1 and A1-P0 introduced by Chen [7] are used to study the degree of nasality. A1-P1 is measured for high vowels and A1-P0 for low and mid vowels. These measurements are taken at the initial, medial and final points of the vowels and are compared to study the degree of nasality in different contexts. The measurements are taken at different points of vowels so that the difference in vowel portions can be observed.

4. Results

A total of 144 utterances have been recorded and analyzed (3 speakers * 3 repetitions * 4 templates * 4 vowels). The measurements of A1-P1 and A1-P0 are made at three locations (initial, medial, final) within each vowel and are compared for oral vowels (O), nasal vowels (N) and contextually nasalized (CVN and NVC) vowels.

4.1. /æ/ vowel

The analysis of the vowel /æ/ clearly shows difference in the degree of nasalization among four syllable types. The measure A1-P0 has lower value for nasal vowels and the vowels in nasal context (VN, NV) than the oral vowels. There is no consistent trend of assigning nasality across contrastively nasal vowel and the contextually nasal vowel. Therefore, both these categories of nasality are significantly different from the oral vowel. The difference among these three can be seen clearly in Figure 2.

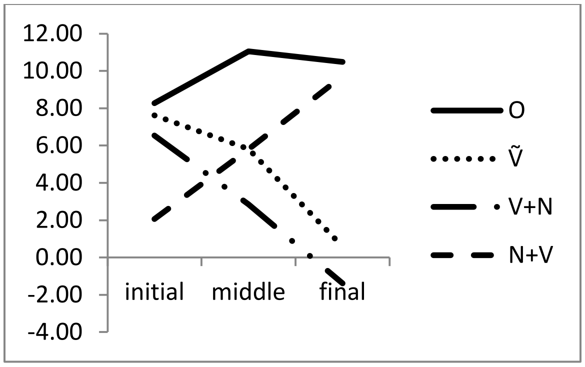


Figure 2: A1-P0 (db) values averaged across three speakers and three repetitions. V+N presents the values averaged across the contexts (v+n, v+m, v+ŋ) and N+V presents the values averaged across the contexts (n+v, m+v)

The average standard deviation (s.d.) for the vowel /æ/ across four syllable types and the measures at three points (initial, middle and final) within vowel is 3.55. The minimum standard deviation is 2.23 and the maximum is 5.18.

4.2. /ʌ/ vowel

The measures taken in the different locations of /ʌ/ vowel in different contexts show greater A1-P0 value in oral context than the other (CVN, NVC and Ṽ) contexts (see Figure.3). The vowel is nasalized greatly at its onset in N+V context and at its offset in V+N context. The nasal vowel /ʌ̃/ reflects greater degree of nasality at its offset.

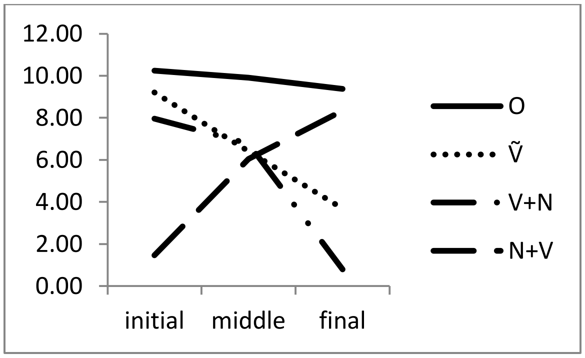


Figure 3: Average A1-P0 (db) across three speakers and three repetitions

The average standard deviation for the vowel /ʌ/ for A1-P0 measure is 3.38, ranging from 2.59 minimum and 5.36 maximum.

4.3. /ɑ/ vowel

The measured A1-P0 values for /ɑ/ in different syllable types provide a clear distinction of nasality in oral vowel and the nasal vowels. We can see the difference among oral and nasal vowels in the Figure 4. The values are lowest for the vowel in CVN and NVC contexts at offset and onset respectively, which depict the effect of neighboring nasal consonant on the preceding and following vowel. The measures at the mid points of contrastively and contextually nasalized vowel are less than the oral one.

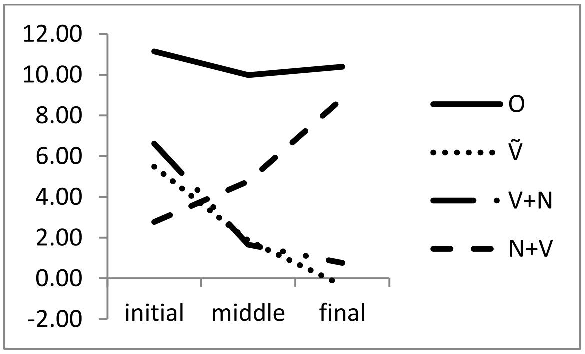


Figure 4: Average A1-P0 (db) across three speakers and three repetitions

The standard deviation for /a/ vowel is 2.75. The minimum s.d. is 2.12 and the maximum is 4.15.

4.4. /ɪ/ vowel

The A1-P1 values measured for /ɪ/ vowel show a greater difference between the oral and nasal vowel. The nasal vowel /ɪ̃/ depicts strong nasalization as compared to the contextually nasalized vowel /ɪ/. The measures of A1-P1 are given in Figure 5.

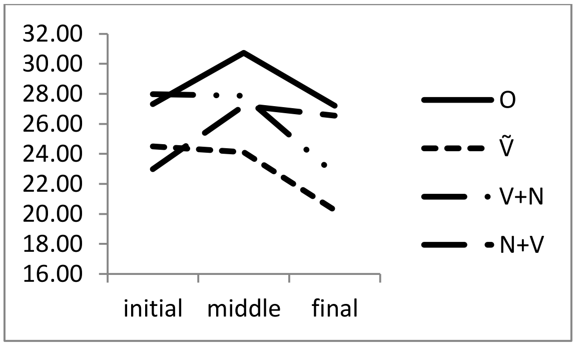


Figure 5: Average A1-P1 (db) across three speakers and three repetitions

The average standard deviation for the measures of the vowel /ɪ/ is 6.94 with 3.75 minimum and 10.14 maximum.

The direction of nasality is shown in the figure 6 clearly. The nasal vowels show greater degree of nasalization than the contextually nasalized vowels. The vowels in CVN context show greater degree of nasality in their offsets. On the other hand the vowels in NVC contexts have greater degree of nasality in their onsets. But the values measured at the middle of the vowels show a greater degree of nasalization in VN context than in NV context. So the Punjabi vowels tend to be nasalized in anticipatory direction

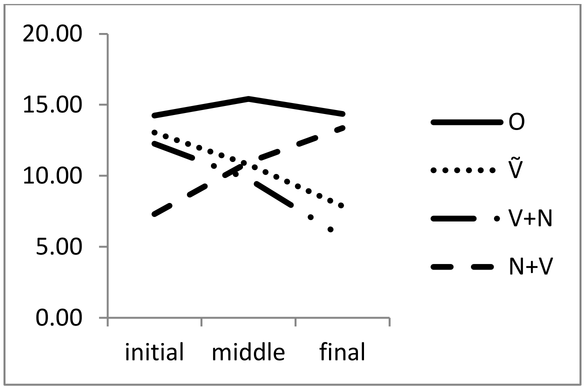


Figure.6. Average (A1-P1, A1-P0) in CVC, CṼC, CVN and NVC contexts across vowels

5. Discussion

The results obtained from the data show very consistent trend of nasalization for both the contrastively and contextually nasalized vowels.

There is significant difference between the A1-P0 and A1-P1 values for oral and nasal vowels. The three nasal vowels /ʌ/, /æ/ and /a/ show tendency to be less nasal at initial point of the vowel but it gradually shows higher degree of nasality towards the middle and final locations within the vowel. Therefore, the A1-P0 and A1-P1 measures show greater difference at middle and final portions of nasal vowels from the oral vowels. Only /ɪ/ shows the opposite trend. For this vowel, the measures of A1-P1 and A1-P0 are lesser at initial and middle locations than the final portion of nasal vowel. The overall averages of the four vowels show the nasal vowel to be nasalized heavily at final point and lesser at initial point (see fig.6).

There is no difference in the degree of nasalization for the contrastively nasal /æ̃/, /ɑ̃/ and /ʌ̃/ and contextually nasalized vowels /æ/, /ɑ/ and /ʌ/ in Punjabi. Only the nasal vowel /ĩ/ shows a greater degree of nasalization than the contextually nasalized vowel /ɪ/. So this study presents a different perspective from the previously cited work on the other languages which have oral/nasal contrast for vowels like French, Taiwanese etc. These languages confirm greater degree of nasalization in contrastively nasal vowels as compared to the contextually nasalized vowels. But Punjabi provides a different account. Punjabi contrastively nasal and contextually nasalized vowels have almost similar degree of nasalization, except /ɪ/ vowel. It contributes to the notion that languages differ on the basis of nasalization patterns.

The vowels in CVN context show greater degree of nasalization in their offsets (following nasal consonant). This tendency marks the influence of neighboring nasal consonants on the following and preceding vowels. The vowels following a nasal consonant, show minimal influence of nasality on their onsets.

On the other hand the vowels in NVC context adopt nasality from their preceding nasal consonant, which is more dominant in the onset of vowels than the other portions. The results of this study observe the Punjabi vowels' tendency to be nasalized in anticipatory direction.

6. Conclusion

This paper has acoustically studied the degree and direction of nasalization in Punjabi vowels. The findings indicate the patterns of nasalization in the contextually nasalized vowels and the contrastively nasal vowels. The contrastively nasal and contextually nasalized vowels are clearly different from the oral vowels. There is no clear difference in the degree of nasality between the vowels in contrastive environment and the vowels in non-contrastive environment, except /ɪ/.

The results also show a clear tendency of Punjabi vowels to adopt contextual nasality in anticipatory direction.

References

[1] T.K. Bhatia, *Punjabi: A Cognitive-Descriptive Grammar,* Routledge, 1993.

[2] T.K. Bhatia, "Punjabi", In K. Brown, & S. Ogilvie, *Concise Encyclopedia of languages of the World,* Elsevier Ltd., 2008, pp. 885-890.

[3] P.S. Beddor, "The perception of nasal vowels" In M. K. Huffman, *Phonetics and phonology: Nasals, nasalization, and the velum*, Academic press, 1993, pp. 171-196.

[4] A.C. Cohn, *"Phonetic and Phonological Rules of Nasalization",* University Microfilms International, 1993.

[5] S.Y. Manuel, “The Role of Contrast in Limiting Vowel-to-Vowel Coarticulation in Different Languages”, *Acoustical society of America* ,1990, pp. 1286-1298.

[6] P.S. Beddor and S. Hawkins, “The Influence of Spectral Prominence on Perceived Vowel Quality”, Huskins laboratory Status Report on speech Research,1991, pp. 187-214, Retrieved on 18.06.2012.
Available:
http://www.haskins.yale.edu/sr/SR105/SR105_14.pdf

[7] M.Y. Chen, “Acoustic Correlates of English and French Nasalized Vowels”, *Acoustical Society of America*, 1997, pp. 2360-2370.

[8] P. Ladefoged and S.F. Disner, *Vowels and Consonants,* John Wiley & Sons, 2012.

[9] T. Pruthi and C.Y. Espy-Wilson, “Acoustic Parameters for the Automatic Detection of Vowel Nasalization”, *Interspeech,* 2007.

[10] A.S. House and K.N. Stevens, “Analog Studies of the Nasalization of Vowels”, *Journal of Speech and Hearing Disorders,* 1956, pp. 218-232.

[11] S. Maeda, “Acoustic Cues of Vowel Nasalization: A Simulation Study”, *Acoustical Society of America,* 1982, pp. 102-102.

[12] A.S. House, “Analog Studies of Nasal Consonants”, *Journal of Speech and Hearing Disorders,* 1957, pp. 190-204.

[13] G. Fant, “*Speech Acoustics and Phonetics”*, Springer, 2004.

[14] V. Delvaux, D. Demolin, B. Harmegnies and A. Soquet, “The Aerodynamics of Nasalization in French”, *Journal of Phonetics,* Elsevier, 2008, pp. 578-606.

[15] K.N. Stevens, “*Acoustic Phonetics*”, MIT Press, 2000.

[16] T. Kawasaki, “Oral-Nasal Contrast for vowels and the degree of Nasalization in Taiwanese”, pp. 15-22, Retrieved on 06.27.2012.
Available:
http://www.hosei.ac.jp/bungaku/museum/html/kiyo/59/articles/Kawasaki59.pdf

[17] J.T. Wright, “The Behavior of Nasalized Vowels in the Perceptual Vowel Space”, *Experimental Phonology,* 1986, pp. 45-67.

[18] R.K. Herbert, “*Language Universals, Markedness Theory, and Natural Phonetic Processes*”, Walter de Gruyter, 1986, Retrieved on 06.20.2012.
Available:
http://18.7.29.232/bitstream/handle/1721.1/35283/71823018.pdf?sequence=1

[19] H.S. Gill and H.A Gleason, “*A Reference Grammar of Punjabi*”, Punjabi University, Department of Linguistics, 1969.

Appendix A: Data set used for the study

vowels	**V**	**Ṽ**	**v+n**	**v+m**	**v+ŋ**	**n+v**	**m+v**
æ	ɣæb	pæ̃da	b^{h}æn	qæm	b^{h}æŋə	nær	mæl
a	pak	bɑ̃g	pan	ʃam	Taŋ	nap	map
ʌ	kʌt	kʌ̃b	kʌn	kʌm	dʒʌŋ	nʌg	mʌt
ɪ	pɪt	pɪ̃d	dɪn	nɪm	dɪŋ	nɪb	mɪt̪

Survey of Urdu OCR: An Offline Approach

Naila Fareen[1], Mohammad Abid Khan[2], Attash Durrani[1]
Dept. of computer science, Allama Iqbal Open University, Islamabad, Pakistan[1]; Dept. of Computer science, University of Peshawar, Peshawar, Pakistan[2]
nailafareen@ hotmail.com, mabid@upesh.edu.pk, attash.durrani@inksoft.net

Abstract

Optical Character Recognition (OCR) is the process of converting printed, handwritten and typed printed text into its equivalent machine readable form. Scanning and comparison techniques are considered to recognize printed text or numerical data. Once the scanned document is converted into machine readable form, the text can then be used in different applications, just like normal machine readable text. It saves time by not typing already printed material for data entry. OCR software attempts to identify characters by comparing figures to those stored in the software library. The discipline of OCR is an offspring of Pattern Recognition, Artificial Intelligence, and Computer Vision. Arabic script (having characters that are connected cursively) makes the recognition of Urdu text more difficult as compared to a language such as English having isolated characters when forming a word. In this research paper, an analysis of 8 years research papers (2002 to 2009) on Urdu OCR has been conducted to show the endeavors for the development of offline Urdu OCR covering both history and future work.

1. Introduction

There are many scripts available in Arabic and Urdu and many people have worked on Nastaleeq type. According to the best of the author's knowledge, none of the work is done on pattern recognition of Typed Urdu Naskh font in which the treasure of knowledge exist left by our predecessors. T. Nawaz and his/her co-authors discussed about the Urdu OCR Naskh font by using pattern matching techniques. The authors worked on pattern recognition for Unicode based computer fonts [1]. So a lot of attention is needed towards Urdu Naskh typed font.

The recognition of characters of Arabic script based languages is not an easy task because of its cursive nature. Arabic characters are connected even when printed or typewritten. The characters of Arabic script and similar other characters are used by a greater percentage of the world's population to write languages such as Arabic, Farsi (Persian), and Urdu. According to V. Margner and H. El Abed, research work on Arabic Optical Text Recognition increased considerably since the 1980's. First systems for Arabic printed text were available at the market in the 1990's. The above authors also did the collection of real world data. They performed scanning and labeling of the collected data to construct a database [2].

Arabic OCR for Printed characters was a research topic in 1990's and a comparison on published papers was reported in 2000 There are three Arabic OCR systems (Sakhr, IRIS, ABBY) that are available in the market but they do not suit for Urdu script, because Urdu has some additional characters. Therefore the OCR used for Arabic or Farsi will not accomplish all the needs for Urdu OCR [2].

2. Urdu OCR History

Urdu is the national language of Pakistan with around 180 million speakers. Urdu script belongs to the family of scripts based on Arabic script. It is a cursive script, i.e. individual characters are usually combined to form ligatures. Although many fonts are available for Urdu, the predominant fonts are Nastaleeq and Naskh. In order to automatically convert an Urdu document image into electronic form, an Urdu OCR system is needed. However, there has been very little work done in the area of Urdu OCR [3]. Urdu is a derivational word from Turkish and it means “horde” (Lashkar). Urdu is an Indo-European language of the Indo Aryan family [4].

Urdu computing started in early 1980s, creating multiple encodings each in different places, as a standard encoding scheme was missing at that time. With the advent of Unicode in early 1990s, some online publications have switched to Unicode, but much of the publications still continue to follow the traditional ad hoc encodings [5].

3. An Analysis of Offline Urdu OCR

On the basis of actual situation of research Table 1 gives an overview of recently published offline Urdu recognition systems and their accuracy.

Table 1: overview of offline Urdu text recognition system

Author(s)	Methodologies\Algorithms	Data Used	Results
S.A. Hussain & S. H. Amin (2002)	Multi-tier holistic approach and Feed Forward Back Propagation Neural Network	200 Carefully selected ligatures	100%
U.Pal & Sarkar (2003)	Water Reservoir Principle	3050 Characters	97.8%
F. Shafait, et al. (2006)	Geometric layout analysis system	Text line detection of 25 scanned images	Books 90%, Magazines 80% and news papers 72%
Z. Ahmed, et al. (2007)	Neural Networks	Old and newly written scripts used to evaluate results	93.4%
I.Shamsher, et al. (2007)	Feed Forward Neural Network	Ariel font Type in the Urdu alphabet set, 72pt font size	98.3%
A.Gulzar & S.Rehman (2007)	Omega: as typesetting Engine	7000 ligatures	Only 65 Nuqta clash
N.Shahzad, et al. (2009)	Subset of Rubine features, weighted and linear classifiers presented, to perform results	38*4 characters	Native Urdu participants, achieved: 92.8%, Non- Native Urdu participants achieved: 73%
T. Nawaz, et al. (2009)	Algorithm Used: Chain code calculation, segmentation, classification, character matching, testing, Unicode file creation.	Different printed Urdu text image files of different font size of Urdu isolated characters	89%
M.W. Sagheer, et al. (2009)	SVM, RBF and MN	3770 images of words.	98.61%
S.T. Javed et al. 2010	HMM and HTK	3655 Ligatures of Noori Nastaleeq	92%
S.S. Bukhari et al. 2011	Multiresolution Morphology and Ridge Based Method	25 Arabic and 20 Urdu Document Images.	99% for non text segmentation, 96% for Arabic and 92% for Urdu text-line Detection.

S. A. Hussain and S. H. Amin present a new approach of offline character recognition. For this purpose, they select the Noori Nastaleeq script by using Multi-tier Holistic approach for the recognition of a ligature [3]. According to U. Pal and A. Sarkar, writing style in Urdu is from right to left whereas it is from left to right in other Indian scripts [6]. Under the sub-heading "PREVIOUS WORK" the above authors mentioned, the accuracy achieved by their system is 97.8%, but it will not handle the messiness and variation of handwritten characters. It can be noted that an Urdu basic character may have four components. There is a structural similarity between Urdu and Arabic script. These authors presented their paper on the recognition of printed Urdu script. They use the water reservoir principles for the character recognition. They use the Hough transform technique for skew angle estimation [6].

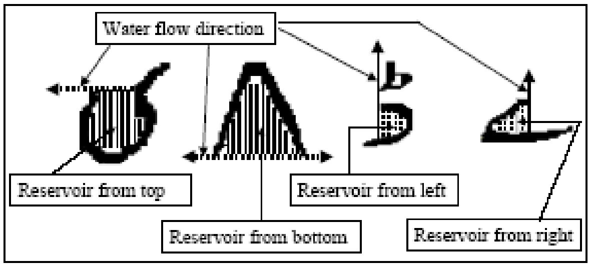


Fig.1. Different reservoirs and their water flow directions are shown in four characters. [6]

According to U. Pal and A. Sarkar, the proposed OCR system automatically detects individual text lines and then segments the character in each line. Their OCR system only recognizes the basic (isolated) characters. They actually want to explain about the complex and cursive nature of Urdu script. They tested the system on 3050 characters. They tested the prototype of this system on printed Urdu characters and currently achieved 97.8% character level accuracy on the average. They describe that Urdu alphabet consists of 39 basic characters [6].

According to F. Shafait and et al [7], layout analysis is the key component of an OCR system. They present a layout analysis system from Urdu document images for extracting text-lines in reading order. They are concerned with geometric layout analysis of Urdu documents. They present the block diagram which shows the flow of layout analysis as shown in the Fig 2.

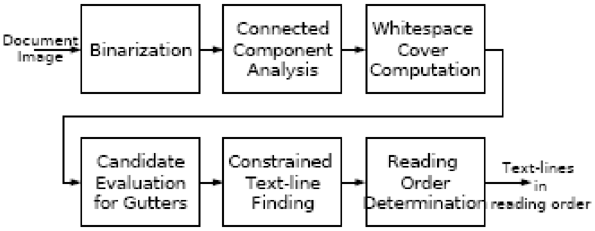

Document Image
Binarization
Connected Component Analysis
Whitespace Cover Computation
Candidate Evaluation for Gutters
Constrained Text-line Finding
Reading Order Determination
Text-lines in reading order

Fig 2: Block diagram of Layout analysis. [7]

According to these authors the results obtained on Urdu script are not as robust as those in the roman script. They collected 25 images of Urdu text to evaluate the performance of the described layout analysis system. They mentioned that due to the presence of dots and diacritics the results were not satisfactory as compared to Latin script. Mainly this paper analyzes, and to some extent detects, the layout of scanned documents. Their proposed algorithm achieved detection rate of more than 90% for line detection for the source taken from books and magazines, whereas in case of decreased inter-line spacing as in digest, decreased the detection results up to 80%. Newspapers source line detection was 72% [7].

4. Segmentation

Z. Ahmed and his/her co-authors describe that Urdu is written in different styles and shapes. They found that recognizing the connection of a word is not a big issue but understanding the shapes/forms of words become complex i.e. when a letter comes in the beginning, in the middle or at the end it changes its original shape. They recognized the printed Urdu script using neural networks and achieved 93.4% accuracy on the average. According to these authors, the main focus is on character segmentation. They mentioned that there is 58 character set defined by NLA (National Language Authority). They assumed that input script was diacritics free, where the diacritics in Urdu script distinguish the homonym easily [8].

M. I. Razzak and his co-authors try to combine the preprocessing steps for Online Character Recognition and present a novel technique for online preprocessing for removing the variation in both online and offline text. Whereas in printed offline approach, the variation rate is not high as found in handwritten documents. They mentioned that they performed different preprocessing steps on the input strokes, but they didn't name those preprocessing steps. They proposed segmentation that was based on the threshold value and position with respect to the previous base character. They did smoothing on the chain code of the stroke. They transformed the input stroke into image to perform offline preprocessing steps. They claim "By using joint processing for online and offline OCR the efficiency can be increased". But they didn't mention the accuracy and recognition rate in percentage, so we are not fully aware about the achievements of this research [9].

M. Akram and S. Hussain presented the word segmentation for Urdu OCR system, they tested their model on the corpus of 150 sentences, and these sentences were composed of 2156 words and 6075 ligatures. They also mentioned that 65 unknown words and 2092known words. The identification rate of this model was 96.10% with 65.63% unknown words [10].

S. T. Javed et al. extracted the global transformational features from non segmented ligature. They used Hidden Markov Model (HMM) for recognition and HMM Tool Kit (HTK) to implement

HMM. For achieving 92% accuracy they tested the system with 3655 ligature of Noori Nastaleeq font [11].

S. S. Bukhari and his co-authors presents a robust layout analysis system from scanned Arabic script document images written in different languages (Arabic, Urdu, Persian) and styles (Naskh, Nastaliq) for extracting text-lines in reading order. They evaluated their system on 25 Arabic and 20 Urdu document images. They achieved 99% non text segmentation accuracy by using multi resolution morphology based method and above 96% text line detection accuracy for Arabic dataset and 92% for Urdu dataset by using ridge based method [12].

5. Recognition

I. Shamsher and his co-researchers presented that there is a lot of work done on the literature of Islamic studies and Urdu, which need to be transferred into electronic form. The above mentioned authors use their own proposed Feed Forward Neural Network Algorithm of MLP (Multi Layer Preceptrons) for the implementation of Urdu OCR Shown in the Fig 3. Their methodology consists of three layers, i.e. input layer, hidden layer and output layer. They worked on OCR system for printed Urdu. The accuracy rate at character level is 98.3% on the average. The software is tested in 72 pt font size only. This software, however, only recognizes the individual characters, so its scope is limited [13].

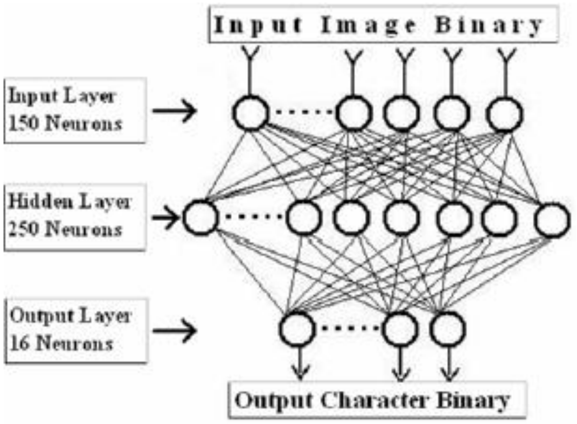
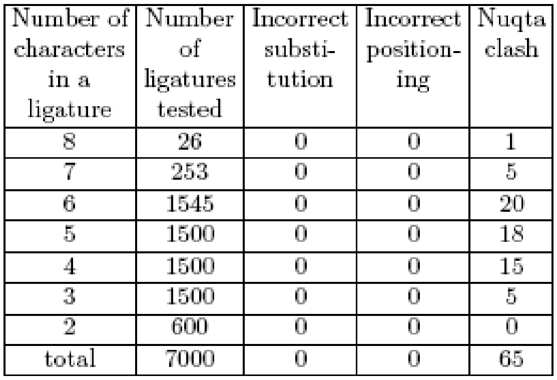


Fig 3: Implemented MLP Network. [13]

According to A. Gulzar & S. Rehman, Nastaleeq is a complex font. They discuss the complexity of Nastaleeq font and also provide the solution that uses Omega as the Typesetting Engine for rendering Nastaleeq. They discuss that there are more than 20,000 ligatures in Urdu and they use only 7000 ligatures randomly from the corpus of 20,000 valid ligatures [14]. Fig 4 shows the test results of A. Gulzar & S. Rehman.

Fig 4: Test results by using Omega. [14]



Number of characters in a ligature	Number of ligatures tested	Incorrect substi-tution	Incorrect position-ing	Nuqta clash
8	26	0	0	1
7	253	0	0	5
6	1545	0	0	20
5	1500	0	0	18
4	1500	0	0	15
3	1500	0	0	5
2	600	0	0	0
total	7000	0	0	65

N. Shahzad and his co-researchers in their paper "Urdu Qaeda: Recognition System for Isolated Urdu Characters" presented the online system for recognizing isolated, hand sketched Urdu characters. The system showed an accuracy of 92.8% for native Urdu writers. According to them, there is no significant research which has been directed towards the online recognition of the Urdu Language. They mentioned that "Urdu language consists of 38 basic characters". The authors further say that most of the characters in Urdu language are multi-stroke. Each character has a single primary stroke and zero or more secondary strokes. They give the example of " ش ". In this case, according to these authors, " س " is the primary stroke and three dots are secondary strokes. However in the discussion section they have shown the " ش " which has single primary stroke and two secondary strokes as shown ش . So their earlier statement that "sheen (ش) is a multi-stroke character which consist of four strokes is not matched with the given image of sheen (ش). According to the above mentioned authors, some characters were not correctly recognized due to similarity in writing. They presented an example of similar characters which cause recognition problem as in the following case.

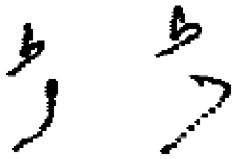
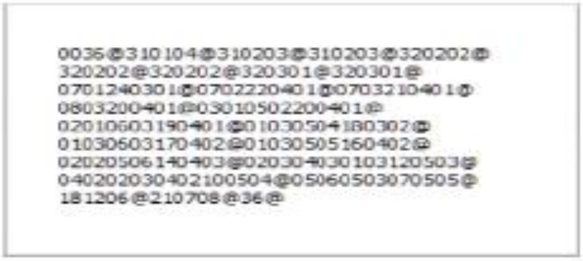


Fig 5: Similar Characters. [15]

These characters are similar in level of secondary stroke as the *toye-shosha* is similar, but the basic or primary characters are quite different. They only use the initial character of cursive script for recognition. They concluded that the system also recognizes a character that is not true in shape. The system should be able to reject the input as an invalid character so the user could learn and draw the character correctly [15].

J.Tariq and his co-workers develop a prototype of Urdu OCR which recognizes the isolated characters of Urdu language with the help of database, without using neural network and its accuracy rate is 97.43% [16].

T. Nawaz et al. discussed about the Urdu OCR Naskh font by using pattern matching techniques. These authors worked on pattern recognition for Unicode based computer fonts and proposed offline character recognition for isolated characters of Urdu language. They use the Chain code algorithm for character matching [1]. Calculating alternating on and off pixels the chain code is generated as shown in the Fig 6.



```
0036@310104@310203@310203@320202@
320202@320202@320301@320301@
0701240301@0702220401@0703210401@
0803200401@0301050220040 1@
02010603190401@010305041 80302@
01030603170402@010305051 60402@
02020506140403@0203040301031205 03@
04020203040210 0504@05060503070505@
181206@210708@36@
```

Fig 6: Calculated String for chain code. [1]

There are many words in Urdu, which are formed using only isolated characters. For example, see Fig 7.

وہ , اور , ا ر د و

Fig 7: Urdu Isolated Characters

But these are only few words, so the claim made in T. Nawaz's paper that "Urdu language forms words by combining isolated characters" is partially correct, as examples mentioned above. They also mentioned, "Urdu is a cursive language, having connected characters making words". The examples of such words are in Fig 8.

پاکستان , اسلام آباد , تحقیق , نسخ

Fig 8: Urdu Ligatures

They worked only for isolated characters, so the scope for their research is limited. The majority of Urdu words are connected characters. They claimed about accuracy up to 89% with the rate of 15 chars/sec. They also mentioned, "Urdu character set has 40 characters". But there is a controversy over the number of letters in Urdu alphabet. National Language Authority declared 58 letters of Urdu, in a meeting dated 26 January 2004 [17].

As per Dr. Rauf Parekh "The controversy over the total and correct numbers of letters in Urdu alphabet has been running for over 200 years now". The algorithm chain code calculation which is normally used in such OCR systems to recognize characters with the segmentation i.e. line segmentation, character segmentation and two levels segmentation [18] .

A. Durrani describes the shapes of four letters as given in Fig 1 that have become the reason of segmentation or space between words. Wherever any of these four letters appear, the ligature of work will break. There is a rule in Urdu that all letters become combined until the word is finished and the last letter will be in full shape [19].

ا ر د و

Fig 9: Letters of the word Urdu

Urdu characters in isolated shape have only one form but in case of connected, their shapes change ranging from 2 to 4 as given in Fig 10.

ثـ ـثـ ـث ث

Fig 10: Urdu character: Shapes

There are many fonts available in Arabic and Urdu but it is considered that "Naskh" and "Nastaleeq" are the main two fonts, Nastaleeq has some complexities. Many people have worked on Nastaleeq type. Calligraphy of Nastaleeq is still waiting for research projects. In Pakistan, none of the work is done on pattern recognition of Typed Urdu Naskh font.

M.W. Sagheer and his co-authors worked on large database for off-line handwriting recognition. They conducted the recognition of Urdu digits and achieved the accuracy of 98.61%. For the recognition process they use the methodologies of Support Vector Machine (SVM), Radial Base Function (RBF) and Moment Normalization (MN) [20].

6. Conclusion and Future Work

This paper describes research on Urdu OCR. It has discussed methods of OCR and classified them according to different criteria. It is the first Urdu offline character recognition survey to give testing procedures and recognition rates for as many systems as possible. However, current systems are applied to restricted domains and/or have only been tested on small datasets.

Future research and testing are needed to develop systems for widespread use. Considering all the aspects in the previous section, the next step is to provide better offline Urdu OCR for the typography and pattern recognition. So a lot of attention is needed towards Urdu Naskh to bring it in a working stage, especially for the type cast by type foundries the pages partitioned accordingly. Urdu typography and calligraphy are enormously different fields but most of the people mixed-up both fields, so the distinction between them may be considered. The work on Urdu typography of Typed Urdu Naskh font that was developed and used by foundries before the advent of computerized printing has not been touched by any researcher. So there is dire need to convert that work into electronic form so that everyone can get benefit from the work that has almost been diminished.

References

Articles in journals:

[1] T. Nawaz, S. A. Naqvi, H. Rehman and A. Faiz "Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique", International Journal of Image Processing, 3(3), pp. 99-104, 2009.

Articles in conference proceedings:

[2] V. Margner and H. El Abed, "Arabic Word and Text Recognition -- Current Developments ", *2nd International conference on Arabic Language Resources and tools,* Cairo, Egypt, pp.31-36, April 2009

[3] S. A. Hussain, and S. H. Amin, "A Multi-tier Holistic approach for Urdu Nastaliq Recognition", *IEEE INMIC*, Pakistan. 2002.

[4] W.Anwar , X. Wang and X.L. Wang. "A Survey of Automatic Urdu Language Processing" Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 13-16, August 2006.

[5] S. Hussain, "Resources for Urdu Language Processing", *the 6th Workshop on Asian Languae Resources*, 2008.

[6] U. Pal and A. Sarkar. "Recognition of Printed Urdu Script", *ICDAR , IEEE*, 2003.

[7] F. Shafait, Adnan-ul-Hassan, D. Keysers and T. M. Breuel "Layout Analysis of Urdu Document Image", *Multitopic Conference,, INMIC'06. IEEE*, 05-07. 2006.

[8] Z. Ahmed, J. K. Orakzai, I. Shamsher and A. Adnan "Urdu Nastaleeq Optical Character Recogniotion", *World Academy of Science, Engineering and Technology*, (32), pp.249-252, 2007.

[9] M. I. Razzak, S. A. Hussain, M. Sher and Z. S. Khan "Combining Offline and Online Preprocessing for Online Urdu Character Recognition", *Proceedings of the International MultiConference of Engineering and Computer Science*, (1). ISBN: 978-988-17012-2-0., 2009.

[10] M. Akram and S.Hussain, "Word Segmentation For Urdu OCR System",Proceeding of the 8th Workshop on Asian Language Resources,Beijing, China, August,2010.

[11] S. T. Javed, S. Hussain, A. Maqbool, S.Asloob, S. Jamil and H. Moin "Sagmetation Free Nastalique Urdu OCR " *Word Academy of Science, Engineering and Technology*, 70, 2010

[12] S. S. Bukhari, F. Shafait and T. M. Breuel "High Performance Layout Analysis of Arabic and Urdu Document Images" 11th *International Conference on Document Analysis and Recognition, ICDAR'11*. Beijing, China, September 2011.

[13] I. Shamsher, Z. Ahmad, J. K. Orakzai, and A. Adnan "OCR For Printed Urdu Script Using Feed Forward Neural Network", World Academy of Science, Engineering and Technology,34,2007.

[14] A. Gulzar and S. Rehman. "Nastaleeq : a Challenge Accepted by Omega", *TUGboat, XVII European TEX Conference*, 29(1), pp.89-94, 2007.

[15] N. Shahzad, B. Paulson and T. Hammond, "Urdu Qaeda: Recognition System for Isolated Urdu Characters" IUI 2009 Workshop on Sketch Recognition, Sanibel Island, Florida, February 8, 2009

[16] J. Tariq, U.Nauman and M.U.Naru "A novel approach to construct OCR for printed Urdu isolated characters", Second international conference on Computer Engineering and Technology (ICCET), 3, pp. 495-498, June 2010.

Electronic Resources:

[17] National Language Authority, "Urdu Alphabet", [On line: *www.nla.gov.pk]*, Retrieved: (07.05.2010).

[18] Rauf Parekh, "Controversy over number of letters in Urdu alphabet", Dawn (English Newwspaper), Karachi/Islamabad, Pakistan, (07.15.2009).

Books:

[19] A. Durrani. "Urdu Informatics", *National Language Authority,* Islamabad, Pakistan, 2008

[20] M. W. Sagheer, C. L. He , N. Nobile and C. Y. Suen " A New Large Urdu Database for Off-Line Handwriting Recognition" *Image Analysis and Processing – ICIAP 2009,* Springer Berlin / Heidelberg , pp. 538-546, 2009.

Error Analysis of Single Speaker Urdu Speech Recognition System

Saad Irtza, Dr. Sarmad Hussain
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore, Pakistan
Saad.lrtaza@kics.edu.pk,Sarmad.Hussain@kics.edu.pk

Abstract

Speaker independent, spontaneous and continuous speech recognition system (ASR) can be integrated to other technologies like mobile to create an interface between technology and illiterate people so that they can use modern technologies. One of the major hurdles in such ASR is unacceptable word error rate. The paper explores the possibility of analyzing the Urdu speech corpus based on recognition results to improve word error rate (WER).

1. Introduction

This paper describes the implementation of single speaker ASR system and process employed in the analysis of speech corpus based on recognition results. The speech corpus has been recorded based on earlier corpus design [1]. It consists of mixture of read and spontaneous speech and divided into two portions for training and testing [2]. Based on the test results confusion matrix has been generated indicating the correctly matched and confused phonemes. Speech corpus has been updated based on the results. The next section describes the previously work done on Urdu ASR and of other languages.

2. Literature Review

Many ASR systems has been developed using different methods on different languages. Brief survey of ASR in Urdu and different languages will be presented in this section.

Performance of English speech recognition has been evaluated in noisy condition by using HTK toolkit [3]. Data of fifty two male and female speakers, 8440 utterance and connected digits has been recorded at different places having different SNR's. Average word accuracy is 87.81%.

Performance of English ASR has been analyzed based on word recognition error rate on subset of Malach [4] corpus. Noise compensation technique has been implemented that results in 1.1% reduction of WER [5].

Speaker adaptive training [6] (SAT) and discriminative training with minimum phone frame error (MPFE) criterion has been used to decrease the errors in Finish Morph based continuous recognition system [7]. Error analysis based on acoustic model has been performed in continuous Chinese speech recognition system Easy talk [8], [9].

Speech recognition system using subspace Gaussian mixture model approach has been developed by having sixteen Gaussian per state. The system has been trained and tested on English, German and Spanish languages of 15.1, 16.5, 14.7 and 1.8, 2.0, 3.7 hours of data respectively. CallHome [10], corpora has been used for evaluation of training and decoding of recognition performance. Phoneme error rate for English, German and Spanish language has been reduced from 54.9, 46.2, 56.3% to 51.7, 44.0, 53.4%.

Hindi (Swaranjali) speech recognizer has been developed for two male speakers [11]. Rcognition vocabulary consists of isolated hindi digits from zero to nine and trained with twenty utterance of a word for each speaker. Recognition result for two speakers has been found to be 84.49% and 84.27%. Hindi speech recognition system by using HTK toolkit has been developed for eight speakers. Recognizer is based on acoustic word model. Rcognition vocabulary consists of thirty isolated Hindi words. Word accuracy has been found to be 94.63% [12].

Robust Urdu speech recognition system by using Sphinx 3 toolkit has been developed in which three language models have been developed incrementally, one model consist of data from 40 female speakers only, one from 41 male speakers only, and one with both male and female speakers (81 speakers). The error rate was 60.2% [2]. An Urdu SR system using by using

Pattern-Matching and Acoustic Modelling approaches to SR for Urdu language has been proposed with a 55-60% accuracy rate [13]. They have used ANN (Artificial Neural Network) to convert a set of frames into phonetic based categories. They used Viterbi search algorithm to search the best sequence path for the given word to be recognized. A single speaker SR system for isolated Urdu digits by using ANN approach has been developed [14]. A mono-speaker database system for Urdu digits by using ANN approach in Multilayer Perceptron (MLP) has been proposed [15]. This system is implemented by using Matlab toolkit. Urdu ASR system has been developed by using sphinx4. This system is based on small vocabulary (fifty two isolated spoken Urdu words). Training set consists of speech data from ten speakers having total of 5200 utterances. The mean word error rate was 5.33% [16]. An Urdu ASR system of single speaker medium vocabulary, 800 utterances consisted of read and spontaneous speech data are mixed together in various ratios, has been developed and the system is tested using spontaneous speech data only [17].

3. Methodology

Two experiments have been developed. 1) Experiment-1 (Baseline) 2) Experiment-2 (Revised). Training and testing data for baseline experiment is described in Table-1.

Table 1: Baseline Data

No. of training utterances	620
Duration of Data	56 minutes
No. of test utterances	45
Read speech utterances	351
Spontaneous speech utterances	269

The speech corpus on Urdu language for the testing and training has been developed described in [2]. Training and testing data is non-overlapping. The transcription of speech files will be done manually and orthographically in Urdu script. The transcription rules will be based on [19]. In the transcription of speech files non-speech areas in the segments will be represented by the Silence, Vocalization and Breath tags manually. All the files will be converted to the Sphinx format using the Sphinx Files Compiler described in [17].

Based on the recognition issues data of revised experiment-2 has been modified as described in Table-2.

Table 2: Revised Data

No. of training utterances	671
Duration of Data	65 minutes
No. of test utterances	60
Read speech utterances	400
Spontaneous speech utterances	269

Data has been added on incremental basis such that amount of training data does not remain a significant factor in decreasing word error rate. This additional data has been selected to increase the amount of training phoneme. This data has been added in form of full read sentences. Words have been chosen that consists of these phonemes. Sentences have been selected such that it contains maximum of these words. The aim is to analyze the effect of increasing training data on phoneme accuracy.

3.1. Toolkit

Speech data has been recorded on laptop in wav format at 16 kHz sampling frequency. Praat [18] has been used to record the speech files over the microphone channel. Sphinx, speech recognition toolkit has been used to develop and test the acoustic model. The Latest nightly release of Sphinx train Sphinx3 has been used to develop the ASR system.

4. Experiment-1 Recognition Results

Baseline recognition results are described here. To perform error analysis on above recognition results, algorithm has been developed to find the frequency of training, matched and confused phonemes and tabulated in the form of matrix. The following graphs show the relationship between the percentage error rate and amount of training data for every phoneme.

Table 3: Baseline Recognition Results

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	18%

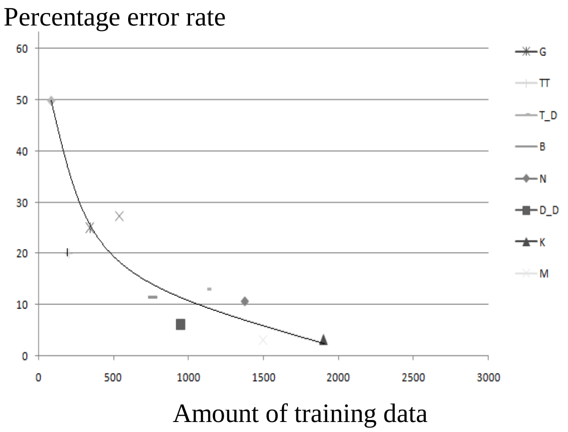


Figure1: Graph for Stops

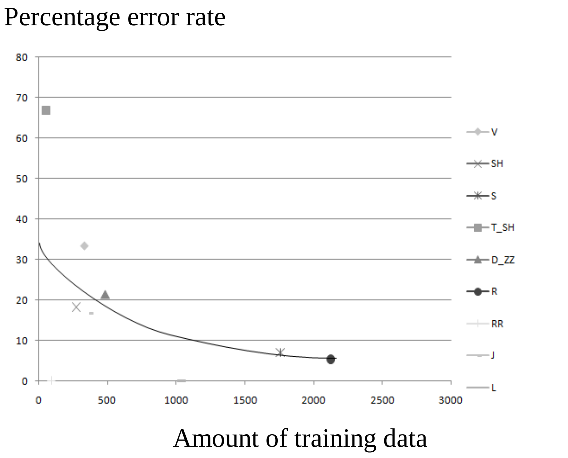


Figure 2: Graph for Fricatives, Trills, Flap, Approximants

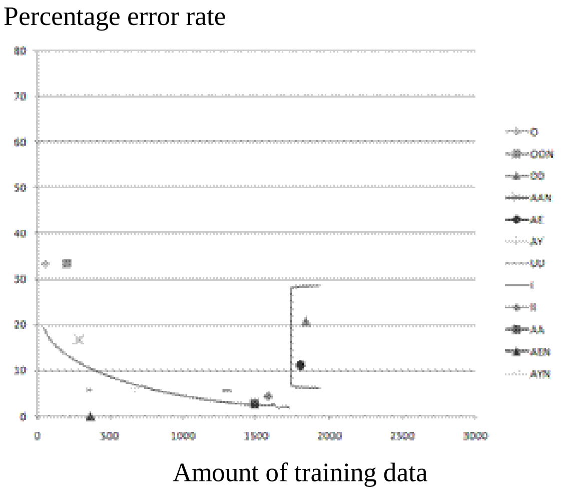


Figure 3: Graph for Vowels

The following Table shows the original phoneme, the confused ones and confusion frequency.

Table 4: Confusion Matrix

Phone	Confusion	Frequency
P	Sil	3
TT	Sil	1
T_D	Sil	3
T_D	D_D	1
N	Sil	3
K	Sil	2
K	P	1
K	B	1
M	Noise	1
V	R	1
Z	D_D	2
Z	R	1
Z	Noise	1
F	Sil	2
SH	K	1
SH	H	1
S	Noise	1
H	Sil	1
T_SH	AA	1
D_ZZ	Z	1
D_ZZ	Noise	2
R	Noise	2
J	Noise	2
O	OON	1
OO	O	2
OO	AE	1
AE	Sil	1
U	AA	1
U	Sil	2
I	II	1
I	Sil	2
AA	OO	2
AA	Sil	2
AA	Noise	4

4.1. Experiment-1 Discussion

Following are the conclusions extracted from the Figure-1, 2, 3 and Table4. Large y-axis value and small x-axis value gives large error rate due to the reason that training data was not sufficient. Large y-axis value and large x-axis value gives large error rate. As the training data was sufficient so there might be two reasons for it:

i) Tagging error
ii) Phoneme is problematic

As described in section-3, the Silence, Vocalization and Breath tags will be defined manually to represent non-speech areas in the segments. From Table-4, there are a lot of confusion between phonemes and silence. Following four solutions have been used to overcome the above problem:

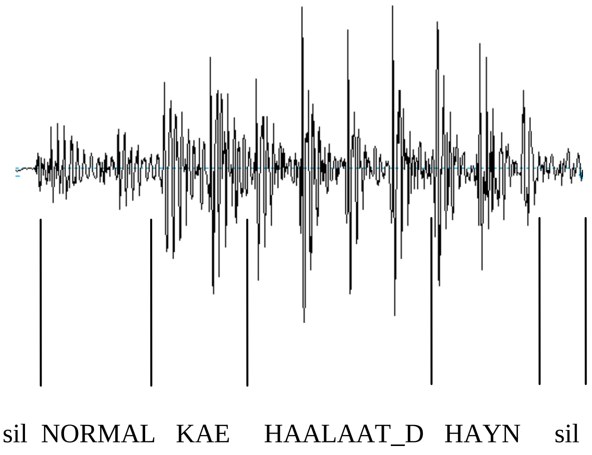
1- Carefully analyze the transcription to check tagging error.
2- Add more data such that phonemes are balanced.
3- The non-speech areas in the segment should be identified automatically.

There are some phonemes whose training data is sufficient but error rate is high. As described in section-3, the segmented speech files have been transcribed orthographically in Urdu script manually. These phonemes have been analyzed in the transcription and original sound files and following are some issues are:

1- Stops phoneme (T_D) was not completely uttered by the speaker.
2- Vowels (DD, AE) were not correctly transcribed at some places.

These issues are solved according to original wave files. There are some phonemes whose training data was not sufficient. Based on results from figure-1, 2, 3, data has been added to previous speech corpus. The non-speech areas in the segment have been identified automatically by using force alignment algorithm. It aligns the transcribed data with the speech data [21].

Original Transcript: <s> <sil> NORMAL KAE HAALAAT_D HAYN <sil> </s>
Force-aligned Transcript: <s> NORMAL KAE <sil> HAALAAT_D HAYN </s>



5. Experiment-2 Results

Revised recognition results are described here. Following graphs show the improved results of above approaches.

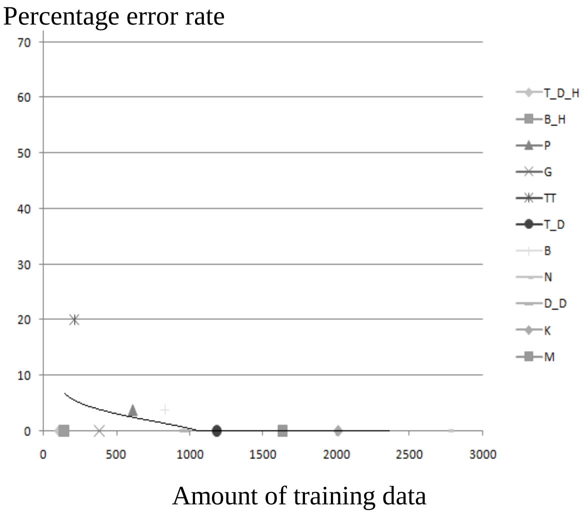


Figure 4: Graph for Stops

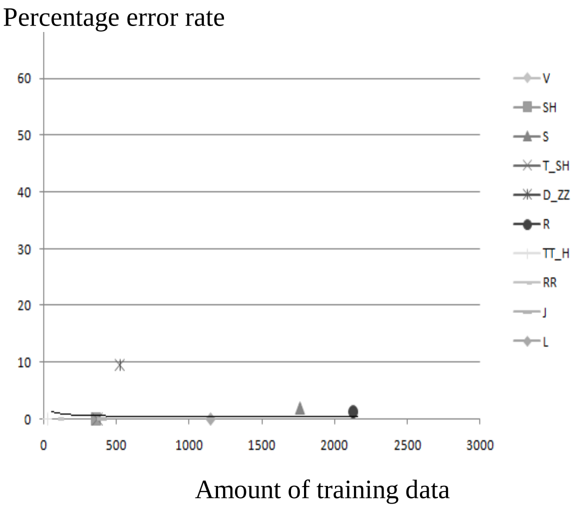


Figure 5: Graph for Graph for Fricatives, Trills, Flap, Approximants

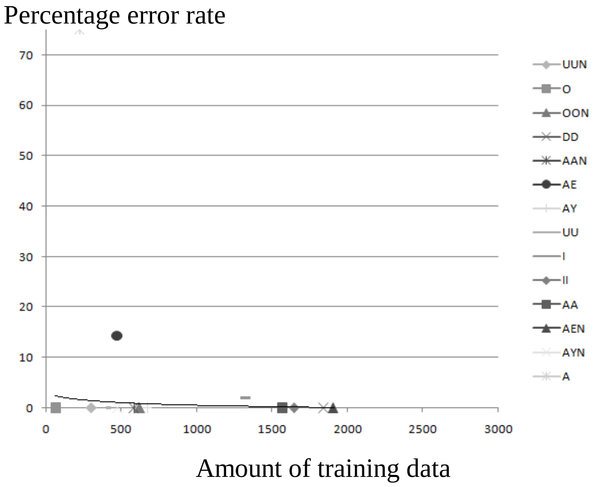


Figure 6: Graph for Vowels

Analysis of transcription improves the phoneme accuracy as described in Table5.

Table 5: Improved Phoneme Accuracy

Phone-me	Training Data	Previous Error rate (%)	Improved Error rate (%)	Percentage Improvement (%)
T_D	1127	13.04	6.52	50
DD	1842	20.69	6.89	66.67
AE	1804	11.11	6.67	39.69

Effect of increasing training data improves the phoneme accuracy as described in Table6.

Table 6: Improved Phoneme Accuracy

Phoneme	Original training data	Increased training data	Improved accuracy from (%)-to (%)
B_H	82	142	50-0
P	540	608	27.3-3.7
G	342	415	25-0
SH	276	360	18.1-0
T_SH	55	515	66.3-0
D_ZZ	485	524	21.4-9.5
O	25	101	33.3-0
OON	203	621	33.3-0
AAN	285	585	16.6-0
AY	572	675	5.3-0
TT	290	974	20-20

Improved Recognition results are described in Table7.

Table 7: Revised recognition Results

No. of tied states	100
Beam width	1e-120
Language weight	23
Word error rate	3.9%
Percentage Improvement	78.3%

5.1. Experiment-2 Discussion

Analysis described in section-3 gives the information that how much times each phoneme appears in the training data. Percentage error rate (PER) has been found by using the formula

$$PER = [(f2-f1) / f2]*100$$

Where
f2= # of times phoneme appear in test data
f1= # of times phoneme correctly decoded

The phoneme 'T_D_H' did not appear in figure-1 but in figure-4 because it did not appear in test data. Same is the case with phoneme 'TT_H' in figure-2, 'UUN' and 'A' in figure-3. Test data has been increased to add the above phonemes.

Training data has been increased based on the technique described in section-3. Table-6 shows the original training data, increased training data and improved accuracy. From Figure-1, 2 and 3, phonemes having relatively low training data and large error rate have been short listed in Table-6. The amount of increased training data is different for every phoneme because it has been increased in form of full sentences. Training data of phonemes other than those listed in Table-6 has also been increased because of adding full sentences e.g. training data of phoneme 'K' has been increased from approximately 1800 (Figure-1) to 2000 (Figue-4). Its error rate has also been decreased to 0%. Some phoneme has no effect on increasing training data and their error rate is also not alarming. From Table-6 error rate of some phoneme has been reduced to 0% and for others to some numerical value. It depends on the context in which phonemes appear in training data. The last phoneme 'TT' in Table-6 shows that there is no effect of increasing training data because utterance/pronunciation of this phoneme was not correct in original wav file. The saturation value of training data for each phoneme is different for single speaker. It may be different for different speaker (male/female) or number of speakers (male/female).

6. Conclusion

Section-5 shows that word error rate can be improved by refining training data, applying force alignment algorithm on transcription and increasing training data for selected phonemes. However these analysis methods will improve WER if and only if utterance/pronunciation of phonemes under observation is correct. For example the phoneme ‘TT’ had no effect of increasing training data (as shown in Table-6) because pronunciation was not correct in original data. The word that contains this phoneme has been difficult to pronounce.

References

[1] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, Rahila Parveen, Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System *in proc. O-COCOSDA*, Kathmandu, Nepal, 2010.

[2] Huda Sarfraz, Sarmad Hussain, Riffat Bokhari, Agha Ali Raza, Inam Ullah, Zahid Sarfraz, Sophia Pervez, Asad Mustafa, Iqra Javed, Rahila Parveen, Large Vocabulary Continuous Speech Recognition for Urdu, *in the Proceedings of International Conference on Frontiers of Information Technology (FIT),* Islamabad, Pakistan, 21-23 December 2010.

[3] Pearce david, Hans-günter Hirsch, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ISCA ITRW*, September 18-20, 2000.

[4] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajiˇc, D. Oard,M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing,* 2004.

[5] Olivier Siohan, Bhuvana Ramabhadran, Geoffrey Zweig, Speech Recognition Error Analysis on the English MALACH Corpus, *ICSLP 8th international conference on spoken language processing*, Jeju island, Korea, October 4-8, 2004.

[6] Teemu Hirsimaki and Mikko Kurimo, Analysing Recognition Errors in Unlimited-Vocabulary Speech Recognition, in *the proc. of NAACL HLT*, Boulder, Colorado, June, 2009.

[7] C. K. Raut, K. Yu and M. J. F. Gales Cambridge University Engineering Department, Adaptive Training using Discriminative Mapping Transforms, *ISCA 9th international conference of international speech communication*, Brisbane, Australia, September 22-26, 2008.

[8] Luo Chunhua, XU Mingxing, Zheng Fang, Center of Speech Technology, Acoustic Level Error Analysis in Continuous Speech Recognition, *ISCSLP*, Beijing, China, October 13-15, 2000.

[9] Fang Zheng, Zhangjiang Song, MingXing Xu, Jian Wu, Yinfei Huang, Wenhu Wu, Cheng Bi., EasyTalk, A Large-Vocabulary Speaker-Independent Chinese Dictation Machine, *EuroSpeech’99, Vol.2, pp.819-822*, Budapest, Hungary, Sept.1999.

[10] Canavan A, and G. Zipperlen, CALLHOME Spanish Speech, Linguistic Data Consortium, 1997, 1997 Contarra systems, 2001
http://www.contarra-systems.com/

[11] Pruthi T, Saksena, S and Das, P K Swaranjali, Isolated Word Recognition for Hindi Language using VQ and HMM, *International Conference on Multimedia Processing and Systems (ICMPS),* IIT Madras, 2000.

[12] Kumar kuldeep, r. k. aggarwal, Hindi speech recognition system using htk, *Internation journal of computing and business ISSN(online) :2229-6166, volume 1*,May 2011.

[13] Akram M. U. and M. Arif, Design of an Urdu Speech Recognizer based upon acoustic phonetic modelling approach, IEEE *INMIC 2004, pp. 91-96*, 24-26 December, 2004.

[14] Azam S. M., Z.A. Mansoor, M. Shahzad Mughal, S. Mohsin, Urdu Spoken Digits Recognition Using Classified MFCC and Backpropgation Neural Network, *IEEE Computer Graphics, Imaging and Visualisation CGIV*, Bangkok, 14-17 August, 2007.

[15] Ahad Abdul, Ahsan Fayyaz, Tariq Mehmood, Students Conference, Speech Recognition using Multilayer Perceptron, *ISCON apos’02. IEEE Volume 1*, 16-17 August, 2002.

[16] Ashraf Javed, Naveed Iqbal, Naveed Sarfraz Khattak, Ather Mohsin Zaidi, Speaker Independent Urdu Speech Recognition Using HMM, *INFOS IEEE*, Cairo, 28-30 March, 2010.

[17] Ali Raza, Sarmad Hussain, Huda Sarfraz, Inam Ullah and Zahid Sarfraz, An ASR System for Spontaneous Urdu Speech, *In the Proc. of Oriental COCOSDA*, Kathmandu, Nepal. 24-25 November 2010.

[18] Praat, doing phonetics by computer, www.fon.hum.uva.nl/praat, accessed June 2010.

[19] S. Hussain , Letter to Sound Rules for Urdu Text to Speech Sytem, *in proc. of workshop on computational*

approaches to Arabic script-based languages, Geneva, Switzerland, 2004.

[20] Chai Wutiwiwatchai, Patcharika Cotsomrong, Sinaporn Suebvisai, Supphanat Kanokphara, Information Research and Development Unit National Electronics and Computer Technology Center, Phonetically Distributed Continuous Speech Corpus for Thai Language, *COCOSDA*, 2003.

[21] Photina Jaeyun Jang and Alexander G.Hauptmann, Improving Acoustic Models with Captioned Multimedia Speech, Multimedia computing system, *IEEE*, Florence, Italy, July, 1999.

A Dictionary Based Urdu Word Segmentation Using Dynamic Programming for Space Omission Problem

Rabiya Rashid, Seemab Latif
Department of Computer Software Engineering,
College of Telecommunication Engineering,
National University of Sciences and Technology (NUST), Islamabad, Pakistan.
rabiya.rashid@gmail.com, seemab@mcs.edu.pk

Abstract

Text segmentation is a process of dividing a sentence into its constituent words. For Natural Language Processing, Word Segmentation is an initial and obligatory step. Research in word segmentation has been done in different languages like English, Dutch, Chinese, Norwegian, Swedish and much more but this research focuses on Urdu language. Unlike English language, words in Urdu language are not always separated by spaces and spaces are not consistently used, which gives rise to both space omission and space insertion errors in Urdu. Space omission and space insertion error is the major challenge for segmentation task. This paper discusses the problems of Urdu Word segmentation and also suggests a solution to the space omission problem and space insertion problem. First, the clustered words are segmented and then each clustered word is divided into valid word. We use dictionary for marking word boundaries and for validating that the word is segmented correctly. This technique can be used for any application of Urdu text. This work has been tested on words collected from Geo[1], Jang[2], BBC[3] news sites and other online documents available on internet. The proposed solution is tested on 11,995 words and the result is around 97.2%.

1. Introduction

First step for developing any application of Natural Language Processing is word tokenization i.e. separate the words of sentences. Word segmentation of Urdu Language is different from English language in that spaces not always mark the word boundary, hence it has space omission problem. While word segmentation in Urdu is different from Chinese language in that space marks boundary of some of the words, hence it has space Insertion problem.

2. Related Research

A very thorough study has been done by Kashif Riaz [4], [5], [7], [8], [9] in Urdu morphology. A comprehensive work on Urdu Word segmentation has been carried out by Nadir Durrani and Sarmad Hussain [3].

Recently, Urdu segmentation has been done for printed Urdu text [6].

Gupreet Sign Lehal [1], [2] has developed a technique to cater with space omission problem for Urdu word segmentation with application to Urdu-Devanagri transliteration system. He used bilingual corpus to tokenize Urdu words.

3. Urdu Language

Urdu is a widely spoken language (the third most widely-spoken language in the world after English and Mandarin), its linguistics characteristics and its resemblance with other spoken languages like Hindi (Hindi is written in a different script than Urdu, but both have same grammar) and Persian (both Urdu and Persian languages have same script). Urdu script derived from a Persian modification of Arabic script written from right to left. Urdu language contains 28 alphabets, 25 consonants and 12 vowels.

[1] www.geo.tv/urdu.htm
[2] http://www.jang.com.pk
[3] http://www.bbc.co.uk/urdu

3.1. Urdu Morphology

This section describes the morphology of Urdu text with the help of examples. Urdu alphabets can be divided into two groups; joiners and non-joiners. Joiners have four shapes e.g. alphabet ب is a joiner and it has four shapes; ﺑ , ﺒ, ﺐ , ب , and non-joiners have two shapes, e.g. alphabet ر is a non-joiner and it has two shapes ﺮ, ر . Table 2 shows the joiners and the non-joiners in Urdu alphabets.

If a word ends in a joiner alphabet, space has to be inserted, e.g. while writing a word like میری, space has to be inserted else the next word will merge with this word. E.g. if sentence contains two words (میری بات) it will be written like this میری space بات . If space is not inserted it will be like this;میریبات, which is an visually incorrect. Similarly, if there is a multi-term word e.g. ضرورت مند, space has to be inserted, if space is not inserted its shape will be ضرورتمند, which is not acceptable in Urdu. Therefore we can say that space is inserted in Urdu words for marking the boundary of the words as well as for a word to have a proper shape. Now consider a part of sentence دوہزارسال, contains three words (دو, ہزار, سال), the words دو and ہزار have non-joiners as their last alphabet i.e. و and ر are non-joiners, they do not need space for word to have a proper shape. So a word that ends at a non-joiner will not have a space from its next word. Segmenting words that ends with a non-joiner is a major challenge.

Table 1 shows different shapes of joiners and non-joiners as they come in the start, middle and end of a word. As it is clear from the table that joiner has different start and ending shapes, so it is relatively easy to identify words that end at joiners. While a non-joiner has same start and end shape therefore it is hard to identify the boundary of a word that ends at a non-joiner.

We have taken the term ligature from Lehal [3]. Urdu word consists of many ligatures e.g. consider a word نام, it consists of two ligatures i.e. نا and م. So every word is a cluster of ligatures. In this paper, we call those words as compound words which consists of more can one word like دنیا جہاں, سیخ کباب.

4. Segmentation Algorithm

After going through Urdu morphology it is clear that either a word can end at joiner or at non-joiner. The algorithm we proposed has four modules. It is a pipeline model in which sentences goes through different modules. First, it enters tokenization module.

Table 1: Shapes of joiners and non-joiners

	Start Shape	Medial Shape	End Shape
Joiner : ب	باجی	مبنی ،دبا	کتاب، کتب
Non-Joiner : ر	رات	مرحلہ ارادہ،	بیمار، امبر

Table 2: Non-joiners and Joiners in Urdu

Joiners	ا د ڈ ذ ر ڑ ز ژ و ے
Non-Joiners	ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ہ ی ھ

4.1. Tokenization Module

As we discussed above, space is typed while writing Urdu words, either for correct shaping or for word boundary. When sentence enters tokenization module it divides the sentences into sub sentences whenever space is encountered. E.g. a sentence ہم سب امیدسے ہیں will be converted into

- ہم
- سب
- امیدسے
- ہیں

4.2. Valid Word Checker Module

It checks whether a word is in the dictionary or not. First each clustered word is checked in dictionary, if it is present it is a single and a valid word, if not that means it's a cluster of words.

In the given example امیدسے will not be a valid word in dictionary that means it consists of two or more words.

4.3. Space Omission Module

Those clustered words which were not present in Valid Word Checker Module are sent to Space Omission Module, in which words are separated from each other. Marking boundary in this case is a challenge because it can be segmented in a number of ways. Now we divide into ligatures. In our case these will be:

- ا
- مید
- سے

We then combine these ligatures to form different combinations. Valid combinations for a clustered word with three ligatures can be 4. First we make different combinations, and then we take each combination at a time and check the words of those combinations in dictionary. The combination in which all words are present in dictionary is selected. It can be possible that two combinations contain all the valid words. E.g. جاکراپنی can be segmented in two ways; جا+ کر + اپنی or جا+ کرا + پنی . Both combinations are valid which one to select? To solve this problem we have used list of frequent words (present on CRULP). The combinations having maximum frequent words will be selected. Figure 1 shows the overall architecture and Figure 2 presents the pseudo code of the proposed algorithm.

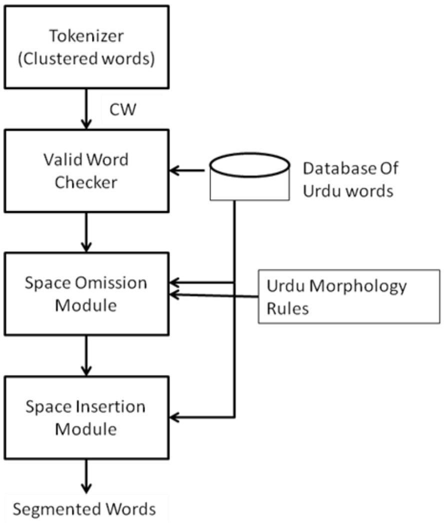


Figure 1: Architecture of Proposed System

```
if(sentence contains space)
{
   Divide sentence into CW
}
if(CW is present in dictionary )
{
   Add to words list
}
else
{
  Divide the CW into ligatures
       if(dictionary contains ligatures)
      {
         Add to word list
      }
     else
      {
            Combine ligatures and then check in      dictionary
      }
}
if(two consecutive words in word list are valid)
{
Merge the into a single word
}
```

Figure 2: Pseudo-code of Proposed System.

4.4. Space Insertion Module

Last module is Space Insertion Module in which multi-term words are detected. This is done by using a very simple technique.

We merge two consecutive words and check whether the merged word is present in dictionary or not , if it does it will be added in the valid word list and those two words will be deleted from the valid word list e.g. consider word شادی شدہ till now شادی and شدہ are different words. We will combine these words and check if they are present in dictionary, if they are present these words will be merged into a single word.

5. Results

This algorithm is tested on a very small data of Urdu words. It is tested on about 11995 Urdu words and 97.2% of words are segmented correctly. Out of 11995 words, 272 words were not segmented correctly. 112 words are segmented incorrectly because of the proposed algorithm while 160 words were not present in dictionary. The efficiency of this algorithm depends solely on the dictionary. If a word is valid and is not present in the dictionary it will not be considered a valid word. The Urdu word list is taken from CRULP (www.crulp.org). It is quite comprehensive but still some words are missing, so those missing words must be added.

5. Future Work

A better Urdu database must be developed which must be complete in every perspective, as dictionary plays a very critical role in segmentation. The more the words dictionary contains, the better the result of the segmentation algorithm. For English words, a proper list should be developed which must cater the problem of English words.

6. Conclusion

In this paper we presented a technique that solves both Space Omission problem and Space Insertion problem. This proposed technique uses dictionary to cater these problems. We have used frequently used words list for better segmentation results.

References

[1] G. S. Lehal, "A Word Segmentation System for Handling Space Omission Problem in Urdu Script**",** *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing* (WSSANLP), *the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, August 2010, p. 43–50.

[2] G. S. Lehal, "A Two Stage Word Segmentation System For Handling Space Insertion Problem In Urdu Script", *Proceedings of World Academy of Science, Engineering and Technology*, Bangkok, Thailand, , 2009, p. 321-324.

[3] N. Durrani and Hussain Sarmad, *Urdu Word Segmentation*, 2010, Retrieved in August 2011.
Available:
http://www.crulp.org/Publication/papers/2010/Urdu Word Segmentation NAACL.pdf

[4] D Becker, K. Riaz, "A study in Urdu Corpus Construction." *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computer Linguistics*. August 2002.

[5] K. Riaz "Baseline for Urdu IR evaluation" *Proceeding of the 2nd ACM workshop on Improving Non English Web Searching, iNEWS 2008, Napa Valley, California, USA,* October 30, 2008. ACM 2008

[6] Malik. H, Faheim. M A, "Segmentation of Printed Urdu Scripts Using Structural Features" *VIZ '09 Proceedings of the 2009 Second International Conference in Visualisation, pages 191-195,* IEEE Computer Society Washington, DC, USA 15-17 July 2009

[7] K. Riaz, "Concept Search in Urdu", *Proceedings of the 2nd PhD workshop on Information and knowledge management*, Napa Valley, California, USA, , October 30, 2008, p. 33-39.

[8] K. Riaz, "Baseline for Urdu IR Evaluation", *Proceedings of the Conference on Information and Knowledge Management, 2nd ACM workshop on Improving Non-English Web Searching*, 2008, p. 97-100.

[9] K. Riaz, "Stop Word Identification in Urdu", *Proceedings of the Conference of Language and Technology*, Bara Gali/Peshawar, 2007.

Complexities and Implementation Challenges in Offline Urdu Nastaliq OCR

Danish Altaf Satti, Dr. Khalid Saleem
Quaid-i-Azam University, Islamabad
dasatti@gmail.com , ksaleem@qau.edu.pk

Abstract

The complexity in machine recognition of Arabic language due to its cursive nature is well known. Urdu is a popular language which is written in Arabic based script but uses a special calligraphic style of writing known as Nastaliq. The calligraphic nature of Nastaliq and other linguistic properties of Urdu introduce many other complexities which must be kept in mind in the development of OCR. This paper introduces all those complexities and open issues which are unique to Urdu language and Nastaliq style or writing from OCR point of view.

1. Introduction

Optical Character Recognition (OCR) is a branch of Pattern Recognition which is used to recognize printed text, normally in form of digitally scanned images or live text coming from drawing by a user through some digital input device. The goal is to convert the input into machine recognizable form. The recognition process is highly dependent upon the source script and writing style. The process used in the recognition on one type of script will usually not be able to recognize the text in some other script, or it may be only partially applicable. This is because machine recognition processes are highly dependent upon the geometrical features, font and writing rules of the script. This makes Optical Character Recognition a vast field and requires research for each type of the writing styles.

Urdu language possess some of the properties which are considered most challenging in character recognition world. The most common of them is cursiveness which it inherits from Arabic. This is why most researchers tackle the problem of Urdu Character Recognition in the same way as that of Arabic Character Recognition which is wrong. The complexities of recognition of Urdu script are much more than that of Arabic script thus require much more attention and out of the box thinking. This paper is written specifically to introduce all the complexities, to the best of our knowledge, which are unique in recognition of Urdu Nastaliq script.

Section 2 describes the characteristics of Urdu script such as its character set and its basic differences from Arabic and Persian. Section 3 presented a summary of current research done in the field of Urdu Character Recognition. Section 4 describes different complexities found in Urdu Nastaliq style of writing and in the end the conclusion is presented in Section 5.

2. Characteristics of Urdu Script

Urdu is written in Arabic script from right to left while numbers in Urdu are written from left to right [16]. Urdu character set is a super set of Arabic and Persian character set, with some additional characters to express Hindi phonemes [13]. Arabic script contains 28 characters while Persian contains 32 characters. Urdu has 38 characters with a few extensions to basic characters which makes total of 41 characters when written in isolated form. The extensions include *alif-madaa* (آ) to *alif* (ا), *Nun-gunna* (ں) to *nun* (ن) and *choti-hey* (ة) to *hey* (ه). This is only the case for the isolated characters however Urdu inherits the property of cursiveness from Arabic script. Cursiveness means that characters are joined with each other while written and take a new shape. This characteristic of Arabic language makes it very difficult for the machine to segment each character separately and recognize it. Not every character in Urdu and Arabic connects with the other characters and some connect only from one side. Some of the characters in the character set are also used as a diacritic marks. These include *Toy* (ط) and *Hamza* (ء). Separate diacritics are also used in Urdu like Arabic such as *zer* (ـِـ), *zaber* (ـَـ), *pesh* (ـُـ), *shadd* (ـّـ) etc but are much less common than in Arabic text. Dots are also very common and significant. In Urdu a character may contain up-to three dots above, below or inside it. 17 out of 38 characters in Urdu have dots, 10 of which have 1 dot, 2 have 2 dots and 5 characters have 3 dots. In

Urdu script height of characters vary a lot as compared to Latin script. Characters in Urdu may also overlap each other vertically.

Urdu is written in Nastaliq style unlike Arabic/Persian which are written in Naskh style. Nastaliq is a calligraphic version known for its aesthetic beauty which originated by combining two styles, Naskh and Taliq [3]. A less elaborate version of style is used for writing printed Urdu. The credit of computerizing Nastaliq goes to Mirza Ahmed Jameel who created 20,000 Nastaliq ligatures in 1980, ready to be used in computers for printing [8]. He called it Noori Nastaliq. Many people followed and created their own Nastaliq style fonts among which Jameel Noori Nastaliq, Alvi Nastaliq and Faiz Lahori Nastaliq are popular. All the Nastaliq fonts fulfill the basic characteristics of Nastaliq writing style.

3. Previous Work

Unfortunately very less work has been persuaded in the field of Character Recognition of Urdu text. One of the reason the amount of complexity associated with Urdu script. These complexities will be discussed in detail in the next section. Another reason could be lack of government and private sector interest and funding. Current work in the field of Urdu Optical Character Recognition can be divided into two major sub-categories which are *i) Online* and *ii) Offline*. Offline recognition means attempting to recognize text which is already present in the form of printed or handwritten material. Thus offline recognition can be further divided into two categories: *i) Printed* and *ii) Handwritten*. Online recognitions refers to real-time recognition as user moves the pen two write something. Thus online recognition only involves handwritten text. Online recognition is considered less complex as compared to offline recognition because in online recognition temporal information of pen traces are available, which is not the case in offline recognition.

Most of the people who worked in Urdu character recognition only attempted to recognize the isolated characters. This is because the cursiveness and some other properties of Nastaliq writing style makes it very difficult to segment words into individual characters. Two major approaches followed for recognition of complete Urdu text found in the literature are: *i) Segmentation based* and *ii) Segmentation free*. Thus far no promising attempt have been made for the segmentation of printed Nastaliq text and segmentation based approaches have been applied on the Naskh style only.

S. Malik and S.A. Khan [11] used “a rule based slant analysis and conversion” for online Urdu handwriting recognition. Their system is able to recognize isolated Urdu characters, numbers, and 200 two character Urdu words with a recognition rate of 93% for isolated characters and numbers and 78% for two character words. S.A. Hussain et al. [5] used a segmentation free approach with 20 different structural features for recognition of 850 single character, 2 character and 3 characters ligatures enabling recognition of 18000 common words of Urdu dictionary. They used BPNN (back-Propagation Neural Network) as a classifier with accuracy of 93% for base ligatures and 98% for secondary ligatures. M.I. Razzak and S.A. Hussain [13] presented a segmentation free approach for recognition of online Urdu text using a hybrid classifier of HMM and Fuzzy Logic. Authors report a recognition rate of 87.6% and 74.1% for Nastaliq and Naskh styles respectively. K.U. Khan and I. Haider [9] applied various classifier such as correlation based classifier, back propagation neural network classifier and probabilistic neural network based classifier on isolated online handwritten Urdu characters and found that probabilistic neural network based classifier works best. A database of 110 instances of handwritten Urdu characters from 40 individuals of different age groups was used and recognition rate of 94% to 98% was reported for 4 different groups of Urdu characters set classified on the bases of number of strokes. M. I. Razzak et al. [14] applied combined online and offline pre-processing techniques on Urdu text for improving efficiency of the Urdu character recognition process. Z. Ahmed and J. K. Orakzai [1] used feed forward neural network for recognition of offline Urdu text. Size of the input text was kept constant and text was assumed to be diacritic free. They report a recognition rate of 93.4%. T. Nawaz et al. [12] applied pattern matching technique on the chain code for the recognition of isolated Urdu characters in Naskh style. They report a recognition rate of 89%. I. Shamsher et al. [17] also used feed forward neural network for recognition of isolated Urdu characters. They report accuracy of 98.3%. S. A. Hussain et al. [6] used Kohonen Self-organizing Map (K-SOM) for pre-segmented Urdu characters in Naskh style. Their system can handle 104 segmented character ligatures with 80% accuracy. S. Sardar and A. Wahab [15] used K-Nearest Neighbour (KNN) algorithm for isolated online and offline Urdu characters using 5 features. They report a recognition rate of 97.12%.

4. Complexities in Urdu Script Recognition

Urdu inherits all the complexities of Arabic script as well as introduces new complexities based upon its Nastaliq writing style. The calligraphic nature of Nastaliq makes recognition of Urdu text much more difficult as compared to Arabic/Persian script which is written in Naskh. Further more linguistic properties of Urdu such as more larger character set also increase the level of complexity. In this section we will take a close look at differences between Nastaliq and Naskh style and other complexities involved in the recognition process. For more details on process for recognition of Arabic text please see [2] and [10].

4.1. Number of Character Shapes

In Arabic each letter can have up-to 4 different shapes depending on its position i.e. initial, middle, ending or isolated. Some letters join with other letters from both sides, some join from only one side and some do not join at all. Each connected piece of characters is also known as ligature or sub-word. Thus a word can consist of one or more sub-words. In Urdu the shape of the character not only depend on its position but also on the character to which it is being joined. The characters change their glyph shape in accordance with the neighboring characters. This feature of Nastaliq is also known as context-sensitivity [3, 5, 16] . Thus in Urdu the possible shapes of a single character are not limited to 4 but it can have many more shapes depending on the preceding and following characters.

We have identified 21 classes in Urdu based upon the characters shape similarity. These classes have unique property that all members of a class when joined with some character in an other class make the same primary shape. Primary shape means that secondary elements such as dots and other diacritics are not considered. Figure 1 shows these classes as well as different glyphs of a character at different positions. Among these classes character *hamza* (ء) do not join from any side and make only one primary shape while all other characters connect form either right or both sides. So the shape of a character will depend on 20 classes as well as three positions which can make up-to 60 different shapes for a single character. The character can exist in isolated form as well, so the number can go up-to 61. This is however upper bound and actual number of shapes are much less because in many cases characters share their glyphs at same positions for different classes. Figure 2 shows different shapes of charter *bey* (ب) when joined with characters from different classes at different positions. From the figure we can count up-to 25 different shapes of bey however the actual number may vary to some extend depending upon the actual font used. This context-sensitive nature of Nastaliq is one of the major property which distinguishes it from Naskh.

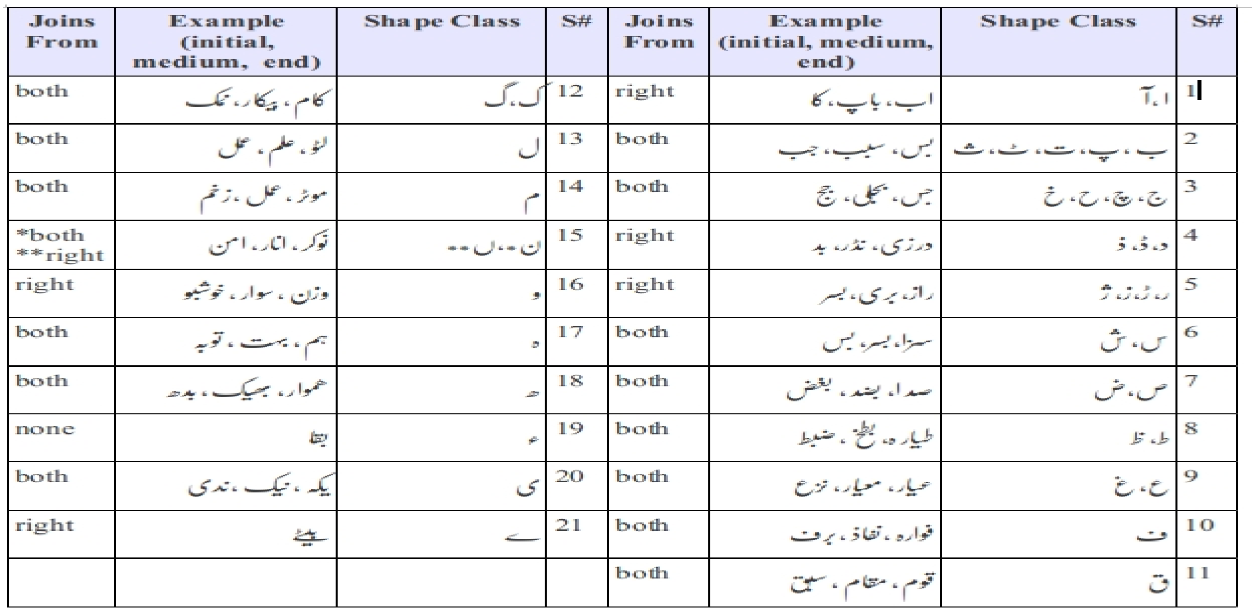
4.2. Slopping

The calligraphic nature of Nastaliq also introduces slopping in the text. Slopping mean that as the new letters are joined with previous letters, a slope is introduced in the text because the letters are written diagonally from top-right to bottom-left [3]. This property of Nastaliq is also named as diagonality by some authors [7]. This means that vertical starting and ending positions of a character in Urdu script are less significant in determining the character. So their vertical positions can not be the determining factor for the characters. One of the major advantages of slopping is that it conserves a lot of writing space as compared to Naskh.

Slopping also means that characters no more join with each other on the baseline which is an important property in Naskh. It is utilized in the character segmentation algorithms for Arabic/Persian text. So the character segmentation algorithms designed for Arabic/Persian text can not be applied on the Urdu text. Number of character shapes and slopping makes Nastaliq character segmentation most challenging task in the whole recognition process and till now in our knowledge not a single algorithm exists which promises decent results in segmentation of sub-words into individual characters. This is also one of the main hurdle which keeps most of the researchers away from accepting the challenge of Nastaliq character recognition. Figure 3 shows the slopping property of Nastaliq text.

4.3. Stretching

Another very important property of the Nastaliq style is stretching. Stretching means that letters are replaced with a longer versions instead of their standard version [3]. Some characters even change their default shape when stretched i.e. *seen* (س) however some only change their width. The purpose of stretching is not only to bring more beauty into the glyph of the character but it also serves as a tool for justification. Justification means



Joins From	Example (initial, medium, end)	Shape Class	S#	Joins From	Example (initial, medium, end)	Shape Class	S#
both	کام ، پکار ، نمک	ک ، گ	12	right	اب ، باپ ، کا	ا ، آ	1
both	لو ، علم ، عل	ل	13	both	بس ، سبب ، جب	ب ، پ ، ت ، ٹ ، ث	2
both	موڑ ، عمل ، زخم	م	14	both	جس ، بجلی ، جج	ج ، چ ، ح ، خ	3
*both **right	نوکر ، انار ، امن	ن * ، ں **	15	right	درزی ، نڈر ، بد	د ، ڈ ، ذ	4
right	وزن ، سوار ، خوشبو	و	16	right	رات ، بری ، بسر	ر ، ڑ ، ز ، ژ	5
both	ہم ، بہت ، توبہ	ہ	17	both	سزا ، بسر ، بس	س ، ش	6
both	ھموار ، بھیک ، بدھ	ھ	18	both	صدا ، بصد ، بغض	ص ، ض	7
none	بچا	ء	19	both	طیارہ ، بطخ ، ضبط	ط ، ظ	8
both	یک ، نیک ، ندی	ی	20	both	عیار ، معیار ، نزع	ع ، غ	9
right	بیٹے	ے	21	both	فوارہ ، نفاذ ، برف	ف	10
				both	قوم ، مقام ، سبق	ق	11

Figure 1: Urdu character shape classes

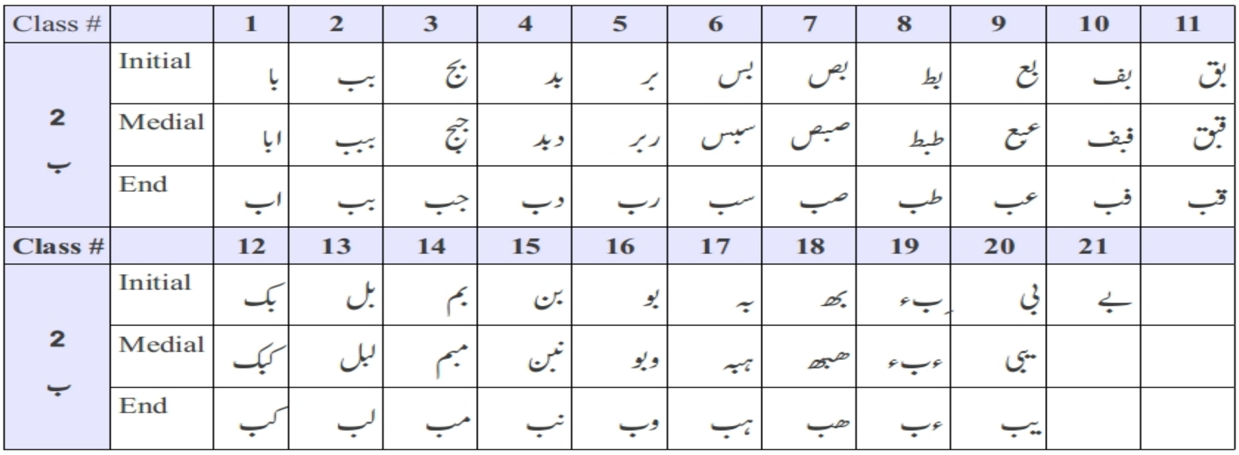
that the text meets the boundaries of the bounded area irrespective to the varying length of the sentences. However it should be noted that not every character in Urdu can be stretched. For example *alif* (ا), *ray* (ر), *daal* (د) can not be stretched but *bey* (ب), *seen* (س) and *fay* (ف) can be stretched. It should also be noted that stretching works closely with the context-sensitive property of Nastaliq and certain class of characters can only be stretched when joined with another character of a certain class or written at a certain position (initial, medial, end, isolated). All these attributes of stretching show that stretching is a complex procedure and it also increases the complexity in machine recognition tremendously. Standard Nastaliq fonts used in the prints normally do not support stretching. However it is commonly used in the titles of the books and calligraphic art. So if we are dealing only with machine printed Nastaliq text, we normally do not need to worry about stretching, but if we are dealing with calligraphic or handwritten Nastaliq document, there is a huge possibility that we have to deal with stretched version of characters. Figure 4 shows Urdu text in which stretched version of character *jeem* (ج) and *qaaf* (ق) are used.

4.4. Positioning and Spacing

Like stretching, positioning and spacing [5] are an important tool for justification in Nastaliq and are also used for the beautification of text. Positioning means the placement of ligatures and sub-words in Nastaliq and spacing means the space between two consecutive ligatures. In normal situations the ligatures are written to right of previous ligature with a small standard spacing. But positioning allows the ligatures to be placed at different positions such as new ligature is started somewhere from the top of previous ligature or it can be placed right above it even if it is a part of another word. Positing will not care even it had to overlap and connect two ligatures if the need arises. Unlike stretching, positioning is quite common and used extensively in the news heading in the Urdu print media industry because of its extreme power to accommodate long and big headings in small spaces in the paper. All these flexibilities and strengths of Nastaliq make it real challenge for the machine recognition. On one hand context-sensitivity and sloping makes the character segmentation a very difficult task and on the other hand positioning makes even the ligature and sub-word segmentation equally more difficult.

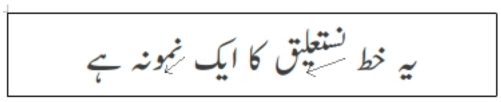
4.5. Intra Ligature and Inter Ligature Overlapping

Another important characteristic of Nastaliq script is intra ligature and inter ligature overlapping [13]. Intra ligature overlapping means that different characters within same ligature may vertically over-



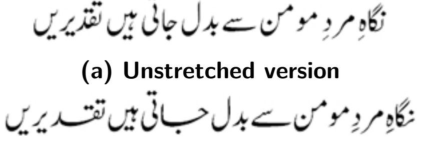
Class #		1	2	3	4	5	6	7	8	9	10	11
2 ب	Initial	با	بب	بج	بد	بر	بس	بص	بط	بع	بف	بق
	Medial	ابا	ببب	جبج	دبد	ربر	سبس	صبص	طبط	عبع	فبف	قبق
	End	اب	بب	جب	دب	رب	سب	صب	طب	عب	فب	قب
Class #		12	13	14	15	16	17	18	19	20	21	
2 ب	Initial	بک	بل	بم	بن	بو	بہ	بھ	بء	بی	بے	
	Medial	کبک	لبل	مبم	نبن	وبو	ہبہ	ھبھ	ءبء	یبی		
	End	کب	لب	مب	نب	وب	ہب	ھب	ءب	یب		

Figure 2: Different shapes of "bey" when joined with different character classes



یہ خط نستعلیق کا ایک نمونہ ہے

Figure 3: Slopping



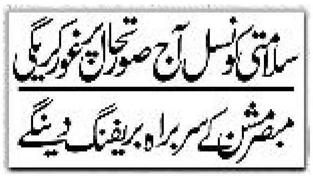
نگاہِ مردِ مومن سے بدل جاتی ہیں تقدیریں

(a) Unstretched version

نگاہِ مردِ مومن سے بدل جاتی ہیں تقدیریں

(b) Stretched version

Figure 4: Stretching



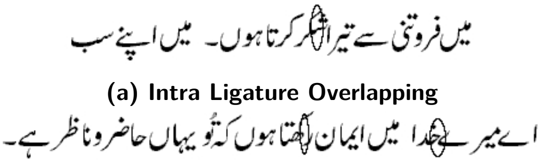
سلامتی کونسل آج صورتحال پر غور کریگی

مبصر مشن کے سربراہ بریفنگ دینگے

Figure 5: Positioning and Spacing

lap each other and inter ligature overlapping means that individual characters from different sub-words may also vertically overlap each other. It is different from positioning because positioning is used only in extreme cases and it positions the whole ligature at completely different position however intra ligature and inter ligature are a part of standard Nastaliq writing and they do not reposition the ligature but only cause a part of the ligature(s) to overlap. Another difference is that intra ligature and inter ligature overlapping will not cause ligature connectivity unlike positioning. These two features are found every where in the Nastaliq text irrespective to whether positioning is used or not.

Inter ligature overlapping makes ligature segmentation more difficult, whereas intra ligature overlapping introduces complexities in character segmentation. Figure 6 shows overlapping example in Nastaliq.



میں فروتنی سے تیرا شکر کرتا ہوں۔ میں اپنے سب

(a) Intra Ligature Overlapping

اے میرے خدا میں ایمان رکھتا ہوں کہ تو یہاں حاضر و ناظر ہے۔

(b) Inter Ligature Overlapping

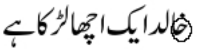
Figure 6: Overlapping

4.6. Filled Loop Characters

While above presented properties of Urdu script makes it a nightmare for character segmentation stage there is another property which makes recognition and distinction of some characters from others equally difficult. This property is called the filled loop property which is also unique to Nastaliq. Some characters in Urdu language have small loops in them which when written in Nastaliq style are filled from inside. This makes these character extremely identical to some other characters which are very similar to them in shape but do not contain loops. This difference is easily visible for naked eye but for machines it becomes very difficult to distinguish between two. These characters include *wao* (و), *meem* (م) and *qaaf* (ق). For example *wao* (و) when written in Nastaliq will be very difficult to distinguish from *daal* (د), specially after performing thinning, which can be a major step in recognition process, these two characters will look exactly the same and *meem*, *qaaf*, *saad* etc will loose their distinguishing feature and their rounded heads will be reduced to lines.

4.7. False Loops

Starting point of some characters when written in Nastaliq joins the base resulting a false loop which is not part of the actual structure of the character. The characters from Shape Class 3 (Figure 1) inhabit the property of false loops and include *jeem* (ج), *chey* (چ), *hey* (ح) and *khey* (خ). There can be two approaches to tackle this issue: 1) Identifying the false loops and break them. 2) Recognize them without breaking the loop. Both approaches are challenging because for machines it is very difficult to recognize false loops or distinguish them from characters with real loops. Not proper tackling of false loops issue will increase the error rate by misclassification of Shape Class 3 characters with Shape Class 7 and 8 characters. Even dots can not be final determining factor but can be helpful in reducing the misclassification rate. Example of false loop for character “khey” is shown in Figure 7.



خالد ایک اچھا لڑکا ہے

Figure 7: False Loop

4.8. Varying Stroke Width

Nastaliq has varying stroke width. Same character has different stroke widths at different positions. Stroke width is an important feature which can be utilized in different stages of recognition process for example it plays an important role in detection and removal of secondary components such as dots and other diacritic marks from the primary components. It is very useful that all the secondary components are removed and handled separately and then accommodated in the recognition stage for the final recognition. This simplifies the procedure by reducing the classes and makes handling of primary text components more easier. Dots can easily be extracted from text if correct stroke width is known because height and width of a single dot is equal to the stroke width of the text. Similar relationships between other components and stroke width can also be found. In Nastaliq due to its varying stroke width property it is difficult to find the exact stroke width. Nastaliq needs more intelligent algorithm for stroke width extraction than used in Latin and Arabic text.

4.9. Complex Dot Placement Rules

In Urdu a character can have up-to three dots placed above, below or inside it. However slopping and context sensitivity can alter the rules for the standard positions of dots. In many situations due to slopping and context-sensitivity, there won’t be enough space for the dots to be placed at standard position such as inside or right below the character. In that case the dots will be moved from their standard position to some other position nearby. The characters whose presence can influence standard dot placement rules are *beri-ye* (ے), *jeem* (ج), *chey* (چ), *hey* (ح), *khey* (خ), *fey* (ف), *qeaf* (ق), *ain* (ع) and *qaaf* (ک) [4]. Due to Inter Ligature and Intra Ligature overlapping the dots might be positioned where it is difficult to decide the actual component to which they belong to. Simple nature of Naskh do not face this issue and dots will always be found at specific locations for the character. However, in case of Nastaliq, situation becomes more complex where it is more difficult to associate dots to the correct primary component.

5. Conclusion

The standard style for writing Urdu, Nastaliq, is inherently complex for machine recognition due to its calligraphic nature. The Challenge of Urdu Character Recognition is different from Arabic Character Recognition because of these complexities. Various issues need to be resolved for Nastaliq Character Recognition among which more important are context-sensitivity, slopping, positioning, overlapping, filled loops and false loops. All the issues presented in this paper are yet to be resolved thus require special attention. We believe that these issues are complex and need to be considered individually by the researchers. Once solved, it will lead to a robust solution to Urdu Nastaliq OCR. So this paper can be taken as a road map to the solution of Urdu Nastaliq OCR problem.

References

[1] Z. Ahmad, J.K. Orakzai, I. Shamsher, and A. Adnan. Urdu nastaleeq optical character recognition. In *Proceedings of world academy of science, engineering and technology*, volume 26, 2007.

[2] B. Al-Badr and S.A. Mahmoud. Survey and bibliography of arabic optical text recognition. *Signal processing*, 41(1):49–77, 1995.

[3] M. Asad, A.S. Butt, S. Chaudhry, and S. Hussain. Rule-based expert system for urdu nastaleeq justification. In *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International*, pages 591–596, 2004.

[4] A. Gulzar and Shafiq ur Rahman. Nastaleeq: A challenge accepted by omega. *XVII European TEX Conference*, 29(1):83–94, 2007.

[5] SA Husain, A. Sajjad, and F. Anwar. Online urdu character recognition system. In *MVA2007 IAPR Conference on Machine Vision Applications*, 2007.

[6] S.A. Hussain, S. Zaman, and M. Ayub. A self organizing map based urdu nasakh character recognition. In *Emerging Technologies, 2009. ICET 2009. International Conference on*, pages 267–273, 2009.

[7] S.T. Javed and S. Hussain. Improving nastalique specific pre-recognition process for urdu ocr. In *Multitopic Conference, 2009. INMIC 2009. IEEE 13th International*, pages 1–6, 2009.

[8] Prof Dr. Syed M. Abdul Khair Kashfi. *Noori Nastaliq Revolution in Urdu Composing*. Elite Publishers Limited, D-118, SITE, Karachi 75700, Pakistan, 2008.

[9] K.U. Khan and I. Haider. Online recognition of multi-stroke handwritten urdu characters. In *Image Analysis and Signal Processing (IASP), 2010 International Conference on*, pages 284–290, 2010.

[10] MS Khorsheed. Off-line arabic character recognition–a review. *Pattern analysis & applications*, 5(1):31–45, 2002.

[11] S. Malik and S.A. Khan. Urdu online handwriting recognition. In *Emerging Technologies, 2005. Proceedings of the IEEE Symposium on*, pages 27–31, 2005.

[12] T. Nawaz, S.A.H.S. Naqvi, H. ur Rehman, and A. Faiz. Optical character recognition system for urdu (naskh font) using pattern matching technique. *International Journal of Image Processing (IJIP)*, 3(3):92, 2009.

[13] M.I. Razzak, F. Anwar, SA Husain, A. Belaid, and M. Sher. Hmm and fuzzy logic: A hybrid approach for online urdu script-based languages' character recognition. *Knowledge-Based Systems*, 23(8):914–923, 2010.

[14] M.I. Razzak, S.A. Hussain, M. Sher, and Z.S. Khan. Combining offline and online preprocessing for online urdu character recognition. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 1:18–20, 2009.

[15] S. Sardar and A. Wahab. Optical character recognition system for urdu. In *Information and Emerging Technologies (ICIET), 2010 International Conference on*, pages 1–5, 2010.

[16] S.A. Sattar, S. Haque, M.K. Pathan, and Q. Gee. Implementation challenges for nastaliq character recognition. *Wireless Networks, Information Processing and Systems*, pages 279–285, 2009.

[17] I. Shamsher, Z. Ahmad, J.K. Orakzai, and A. Adnan. Ocr for printed urdu script using feed forward neural network. *the Proceedings of World Academy of Science, Engineering and Technology*, 23, 2007.