Proceedings of the Conference on

LANGUAGE & TECHNOLOGY

2016

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

CONFERENCE COMMITTEES

Organizing Committee

General Chair:	Sarmad Hussain, CLE, KICS-UET, Lahore, Pakistan
Chair, Technical Committee and Co-Chair, Script Processing Track:	Faisal Shafait, NUST, Islamabad, Pakistan
Co-Chair, Language Processing Track:	Tahira Naseem, IBM, USA
Co-Chair, Speech Processing Track:	Tania Habib, UET, Lahore, Pakistan
Chair, Publication Committee:	Agha Ali Raza, IT University, Lahore, Pakistan
Chair, Program Committee:	Sana Shams, CLE, KICS-UET, Lahore, Pakistan

Technical Committee

Adnan ul Hasan	National Engineering and Scientific Commission, Pakistan
Agha Ali Raza	Information Technology University, Pakistan
Annette Hautli-Janisz	University of Konstanz, Germany
Athar Khurshid	Lahore Leads University, Pakistan
Faisal Shafait	National University of Sciences and Technology, Pakistan
Farhat Jabeen	University of Konstanz, Germany
Gernot Kubin	Technical University of Graz, Austria
Hassan Sajjad	Qatar Computing Research Institute, Qatar
Imran Siddiqi	Bahria University, Pakistan
Karthick Narasimhan	Massachusetts Institute of Technology, USA
Kashif Javed	University of Engineering and Technology, Pakistan
Miriam Butt	University of Konstanz, Germany
Muhammad Kamal Khan	Shaheed Benazir Bhutto University, Pakistan
Muhammad Shehzad Hanif	King Abdul Aziz University, Saudi Arabia
Muhammet BAŞTAN	Turgut Özal University, Turkey
Muzzamil Luqman	University La Rochelle, France
Nadir Durrani	Qatar Computing Research Institute, Qatar
Nibal Nayef	University of Hamburg, Germany
Rudolf Rabenstein	University of Erlangen-Nuremberg, Germany
Saad Irtza	University of New South Wales, Australia
Saba Urooj	Center for Language Engineering, Pakistan
Sarmad Hussain	University of Engineering and Technology, Pakistan
Sebastian Lory	University of Konstanz, Germany
Sharmeen Tarar	University of the Punjab, Pakistan
Sheikh Faisal Rashid	University of Engineering and Technology, Pakistan
Tafseer Ahmad	DHA Suffa University, Karachi, Pakistan
Tahira Naseem	International Business Machines, USA
Tania Habib	University of Engineering and Technology, Pakistan
Yoon Keok Lee	International Business Machines Research, USA
Young-Suk Lee	International Business Machines Research, USA
Yuan Zhang	Massachusetts Institute of Technology, USA

Publication Committee

Agha Ali Raza	Information Technology University, Pakistan
Hira Ejaz	Information Technology University, Pakistan

Program Committee:

Aisha Yousaf	Center for Language Engineering, Pakistan
Atif Ali Javed	Center for Language Engineering, Pakistan
Benazir Mumtaz	Center for Language Engineering, Pakistan
Dilshad Ali	Center for Language Engineering, Pakistan
Hina Khalid	University of Engineering and Technology, Pakistan
Kashif Javed	University of Engineering and Technology, Pakistan
Muhammad Kamran Khan	Center for Language Engineering, Pakistan
Sahar Ruaf	Center for Language Engineering, Pakistan
Sana Shams	Center for Language Engineering, Pakistan
Tania Habib	University of Engineering and Technology, Pakistan
Touqeer Ehsan	Center for Language Engineering, Pakistan

FOREWORD

On behalf of the Organizing Committee, we welcome the authors and participants to the sixth 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 2016 (CLT16). As the sixth CLT, this conference is helping mature the language technology research in Pakistan and providing the intended platform for researchers to interact.

Thirty eight papers have been submitted for CLT16, of which twelve have been accepted for presentation and four for posters, through a rigorous review process by an international technical committee. The papers cover Mewati, Pashto, Sindhi and Urdu language specifically, and a host of areas, including linguistics and computational aspects of phonetics, phonology, syntax and semantics. This CLT also presents exciting talks including recent research in syntax, prosody and frontiers in machine translation.

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 Higher Education Commission of Pakistan, Punjab Higher Commission, 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 CLT16 and a pleasant stay in Lahore.

Sarmad Hussain
On behalf of the Organizing Committee

TABLE OF CONTENTS

Analysis and Development of Resources for Urdu Text Stemming

Abdul Jabbar
Department of Computer Science
Institute of Southern Punjab, Multan, Pakistan
a.jabbar73@hotmail.com

Sajid Iqbal
Department of Computer Science
Bahauddin Zakariya University, Multan, Pakistan
Sajid.iqbal@bzu.edu.pk

Muhammad Usman Ghani Khan
usman.ghani@kics.edu.pk
Department of Computer Science and Engineering
University of Engineering and Technology, Lahore

Abstract

Urdu has been facing various challenges in Natural Language Processing (NLP) due to its morphological richness. Stemming, as a preprocessing technique used in different applications of natural language processing, is one of the basic morphological operation applied in written text. The technique is used differently in its various applications like machine translation, query processing, question answering, text summarization and information retrieval. This paper presents the Urdu resources for Urdu text stemming such as affixes list, stop word list, stem word list and stemming rules to remove the infixes letter(s) and recoding to extract correct stems. Here, we collect 1211 affixes, 1124 stop words, 40904 stem word list and 35 rules with their various variations to remove the infixes.

Keywords: *Stemming, Urdu stem words, Urdu affixes, Urdu stop words, Natural Language Processing*

1. Introduction

Urdu language is different from other languages like English in terms of its linguistics and phonetic rules. It was developed in the 12th century from the regional Apabhramsha of northwestern India[1]. In the following paragraph, we list the prominent differences as compared to English language with respect to text stemming process.

Urdu script writing orientation is from right to left. In Urdu word alphabets are connected and do not start with capital letter as in English. Furthermore, most of the characters change their shape based on their position in the word and adjunct letters. Table 1 below demonstrates some of the Urdu characters morphology.

Table 1 Different shape of Urdu Character

آخری شکل Final	درمیانی شکل Medial	ابتدائی شکل Initial	حرف Letter
ع،ﻊ	ﻋ	ع	ع
مرتفع، محلوقوع	تعارف، تعظیم	عابد، عندلیب، عروس بہار	مثالیں

In English each word is separated by a hard space, but in Urdu words are not always separated by hard

[1] http://www.britannica.com/topic/Urdu-language

space, e.g. اولے کا بدلا. Further, two or more words could be written as a single word like کیساتھ، ہمنواز. This is a challenging issue in Urdu text segmentation In English proper nouns always start with capital letter but in Urdu proper nouns do not start with a capital letter because Urdu has no letter casing. Due to this, it is a challenging task to extract proper nouns from Urdu text. English language researchers have used rules and tools like named entity recognition (NER) to mark proper nouns in given text. On the other hand, such tools for Urdu are either rare or have low accuracy rate [32],[33].

In English, inflectional or derivational forms are created by attaching affixes to either or both sides of the root. For example, words *readable* and *unreadable* are created from the root word *read.* In this example, "*un*" is a prefix and "*able*" is a suffix. However, in Urdu language, affix letter(s) could be found anywhere in the middle of the word. For instance, رسوم derived from رسم and اکابر derived from اکبر. This is a challenging issue in Urdu text stemming because, it is difficult to distinguish between affix letters and actual part of the root letters. Conversion from singular to plural is also different in Urdu language. There are multiple rules to do this. For example, a singular could be converted to plural by adding some suffix such as کتاب سے کتابیں, by adding some co-fix with suffix -Further there may be multiple rules to convert a word into its plural کتاب سے کتابیں /کتابوں/کتب. The case of broken plural words is also different, because they do not follow the normal morphological rules. Urdu broken plural words are somewhat like irregular English plurals. The difference is that English singular and plural words resemble to each other, such as man to men, but in the case of Urdu both may be non-identical مشہور سے مشاہیر . Another difference between two languages is that English has only uni-gram words even after derivation. In Urdu there are uni-gram, bi-gram and tri-gram words that are obtained after derivation for example غیر تربیت یافتہ ، شادی شدہ and کتابیں.

Resource development is basic and fore most step of Natural Language Processing (NLP) field. Urdu resource development starts with Urdu Zabta Takhti (UZT) that is standard code for Urdu characters, approved by Government of Pakistan (GOP) [8].

A text corpus consisting of 18 million Urdu words is collected by the Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan [10]. CLE at University of Engineering and Technology, Lahore has also produced different resources for Urdu NLP. This paper focuses on construction of Urdu resources which can be used especially for stemmer development for Urdu text and evaluation, and generally for research in computational linguistics. This paper describes the lists of resources produced during my MS thesis and are available online[2].

The remainder of this paper is organized as follows: Section 2 gives the overview of stemming algorithms. Section 3 describes the development of stop words list. Section 4 introduces the Urdu affix list. The Development of the stem word list is described in section 5 and development of infix rules are described in section 6. In section 7 conclusion and discussion of future work can be found.

2. Stemming algorithms

In stemming, affixes are chopped off from derivational and inflectional forms of Urdu words to extract stem. For example, Urdu words خوشگوار، ناخوش in which نا and گوار are affixes and its common stem is خوش. According to Lovins [15] "*a stemming algorithm is a computational procedure that reduces all the words with same root by stripping each word of its derivational and inflectional suffixes*". Khoja and Garside [13] define the stemming algorithm in Arabic language context as "*Stemming is the process of removing all of a word's prefixes, suffixes and infixes to produce the stem or root*".

Anjali Ganesh [23] analyzed the English stemmer and classified them into three: truncating, Statistical and mixed. Moral [24] grouped stemming algorithms into algorithmic-based approaches and linguistic-based approaches. Moghadam [25] classify Persian stemmer into three classes: structural stemmers, table lookup stemmers and statistical stemmers. Basically, there are two types of stemmers and third is a combination of these two, these are described in detail below.

a) Rule base stemmer

This is most commonly used stemming technique. First stemmer [5] that is found in literature was a rule base stemmer. This stemmer is suitable for those

[2] https://sourceforge.net/projects/resource-for-urdu-stemmer/

languages that have well defined linguistic structure. Rule base approach is also a better choice for infixes removal. This stemmer can give promising results if excellent rules are developed. On the other hand, too much rules make it more complex with inclusion of contrary and recursive set of rules. To apply one rule you have to remove other (opposite) rule. This situation can easily lead toward deteriorated performance.

b) Statistical stemmer

Another popular approach is the use of statistical techniques. In such approaches, the system learns from available corpus and decides on unseen data using knowledge gained through its experience, i.e. statistical measures obtained through experience are used to remove prefixes and affixes. However, the use of statistical methods is popular among the language processing community. N-gram based statistics are used by W. B. Frakes [26]. Melucci [27] used HMMs and YASS: Yet another suffix stripper by Majumder [28] is a few of the examples. It can easily be seen that such approaches are limited to gain experience. In other words, such approaches rarely cover the complete grammar of the language [29]. For example, N-gram approach is better to remove suffixes only however, Urdu language has suffixes, co-suffixes, prefixes, infixes and circumfixes (prefixes and suffixes simultaneously) [30] and use of HMM requires that every word must start with the prefix [27]. This review shows that statistical approaches are insufficient in stemming process. A comprehensive review of stemming technique applied in Urdu, Persian and Arabic is present in [39].

c) Hybrid stemmers

Hybrid stemmer is combination of the rule base stemmer and statistical stemmer [40].

3. Development of stop words list

Usually a sentence contains stop words and content words as shown in figure 1. Stop words have no linguistic meaning, so useless in queries and can generally be safely ignored, e.g. in Urdu "کا","پر","سے" and in English "on", "in", "to", and "the". On the other hand, content words are keywords of any sentence and have lexical meaning. Content words list usually contains nouns, verbs, or adjectives. Urdu stop word list usually contains postpositions, determiners, pronouns, and conjunctions [7]. [2] Used 400 translated Urdu stop words from English stop words. In these translated Urdu stop words, some are not actually stop words e.g. طریقوں،لمحات وغیرہ. 1200 closed words list is provided in [22] and all the close words are not stop words, for instance بےشک, تسلیمات are not stop words but closed words, and these words are stem able words.

Stop words

پاکستان میرا وطن ہے

Content words

Figure 1: Example of stop word and content word

Up to now, no significant work has been carried out to find the stop words from Urdu text. After studying the various Urdu grammar books and literature [3], [4], [18], [22], [34], [35], [36], [37]. We developed stop word list consisting of 1124 Urdu words

4. Development of Affixes lists

The affixes are a morpheme or set of morphemes/words that are frequently attached to other words and create new words. Internal structure of Urdu word is shown in figure 2.

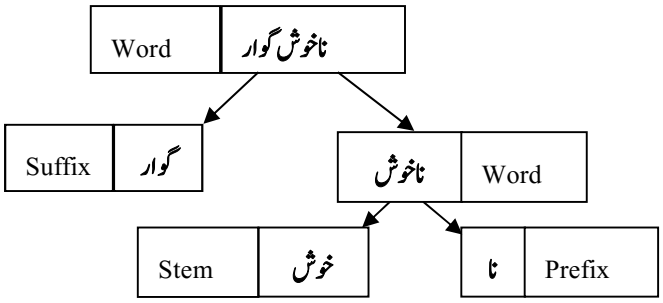


Figure 2 Internal structure of Urdu words Figure

A prefix is a morpheme or word that attaches with the start of the words and changes its meaning, e.g. ناپاک in which نا is a prefix morpheme and words خوش اخلاق in which خوش is a prefix word. On the other hand, suffix is such morpheme or word that comes at the end of the words and it does not change the meaning of the word, e.g. لڑکیاں in which ان are a suffix morpheme and the word دل پسند in which پسند is suffix word. Infixes are letters that can be anywhere in the middle of words, e.g. اکابر with which letter ا is an infix and root word is اکبر.

Urdu not only borrows the words from other languages, but also borrows affixes. Sometimes loan words are also used as affixes to make a hybrid word by Hybridization [31]. In table 2, some examples of affixes/words from the Hindi, Persian, Arabic and English are listed. In Urdu language, loan affixes word/letters from one language can be attached to other language loan affixes, words/letters to create a new single or compound word.

Table 2 Sample affixes list

Language	Prefixes	Words	suffixes	Words
Persian affixes in Urdu	تہ	تہ بند	انہ	مردانہ
Hindi affixes in Urdu	ک	کراہ	ک	بیٹھک
Arabic affixes in Urdu	ال	القران	فی	فی صد
English affixes in Urdu	ڈبل	ڈبل روٹی	سٹور	کریانہ سٹور

Qurat-ul-Ain [1] identified 174 prefixes and 712 postfixes. Khan [11], [12] collected 180 prefixes and 750 suffixes for Urdu text. Mubashir [16] mentions 60 prefixes and 140 suffixes. 122 suffixes and 15 prefix are identified by [7]. After studying the various Urdu grammar books and literature [3], [4], [9], [18], [22], [34], [35], [36], [37], we constructed prefixes 643, suffix 568; then arrange them according to their length.

5. Development of Stem word list

Stem words list is essential to validate the extracted stem. Khan [12] construct a stem words list of 3500 words for Urdu. Mubashir [16] developed a stem words list of about 10000 words. After studying the various Urdu grammar books and literature [3], [4], [9], [18], [22], [34], [35], [36], [37] we construct a root word list containing 40904 words. Examples of stem words are تدبیر، حکم.

6. Development of infixes rules

After studying the various Urdu grammar books and literature [9], [18], [34], [35], [36], [37]. We chop off the infixes letters from the Urdu words using orthographic pattern. We construct 10 rules for Urdu word length 4 letters with variations of rules, 12 rules for Urdu word length 5 letters with variations of rules and 13 rules for Urdu word length 6 letters with variations of rules. A sample list of infixes removal rules are given in table 3 and complete rules are available online.[3]

7. Conclusion

Generally, Information Retrieval (IR) system used variant forms of the query word by stemming process. We have pointed out the differences between Urdu and English language with respect to stemming process. We have also described different types of stemming approaches and borrowed affixes from Arabic, Persian, Hindi and English languages. In this paper, we presented required linguistic resources for Urdu text stemming. Evaluation of these language resources are given in [38].

[3] https://sourceforge.net/projects/resource-for-urdu-stemmer/

.

Table 3 Sample infixes removal rules with variation

Set of Rules: Words Length 4 and Stem Words Length 3 Characters					
Rule No. 1					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	امور	ر	و	م	ا
Stem Word	امر	ر		م	ا
Rule No. 1 Variations A					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	خطوط	ط	و	ط	خ
Invalid Stem	خطط	ط		ط	خ
Deletion	ط			ط	خ
Stem Word	خط			ط	خ
Rule No. 1 Variations B					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	حصول	ل	و	ص	ح
Invalid Stem	حصل	ل		ص	ح
Insertion	ا	ل	ص	ا	ح
Stem Word	حاصل	ل	ص	ا	ح
Rule No. 1 Variations C					
Index		**3**	**2**	**1**	**0**
Orthographic pattern		ـ	و	ـ	ـ
Input word	سجود	د	و	ج	س
Invalid Stem	سجد	د		ج	س
Insertion	ه	ه	د	ج	س
Stem Word	سجده	ه	د	ج	س

References

[1] Akram, Qurat-ul-Ain, Asma Naseer, and Sarmad Hussain. "Assas-Band, an affix-exception-list based Urdu stemmer." In Proceedings of the 7th Workshop on Asian Language Resources, pp. 40-46. Association for Computational Linguistics, 2009.

[2] Burney, Aqil, Badar Sami, Nadeem Mahmood, Zain Abbas, and Kashif Rizwan. "Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors." International Journal of Computer Applications 46, no. 19 (2012).

[3] BBC Urdu (2016): News and research articles retrieved from http://www.bbc.com/urdu

[4] DAWN News(2016): News and research articles retrieved from http://www.dawnnews.tv/

[5] Ethnologue Languages of the World (2015). "Urdu." Retrieved 29 November, 2015, from https://www.ethnologue.com/language/urd.

[6] ENCYCLOPAEDIA BRITANNICA (2015). "Urdu language." Retrieved 01 DECEMBER, 2015, from http://www.britannica.com/topic/Urdu-language.

[7] Gupta, Vaishali, Nisheeth Joshi, and Iti Mathur. "Design & development of rule based inflectional and derivational Urdu stemmer 'Usal'." In Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on, pp. 7-12. IEEE, 2015.

[8] Hussain, Sarmad, and Muhammad Afzal. "Urdu computing standards: Urdu zabta takhti (uzt) 1.01." In Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, pp. 223-228. IEEE, 2001.

[9] Hussain, Sara. "Finite-state morphological analyzer for urdu." PhD diss., National University of Computer & Emerging Sciences, 2004.

[10] Hussain, Sarmad. "Resources for Urdu Language Processing." In IJCNLP, pp. 99-100. 2008.

[11] Khan, Sajjad, Waqas Anwar, Usama Bajwa, and Xuan Wang. "Template Based Affix Stemmer for aMorphologically Rich Language." International Arab Journal of Information Technology (IAJIT)12, no. 2 (2015).

[12] Khan, Sajjad Ahmad, Waqas Anwar, Usama IjazBajwa, and Xuan Wang. "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language." In 24th International Conference on Computational Linguistics, p. 69. 2012.

[13] Khoja and Garside, "Stemming Arabic Text" 1999. Available online at URL: http://zeus.cs.pacificu.edu/shereen/research.htm#stemming [accessed 27/12/2015].

[14] Kwintessential (2015). "The Urdu Language." Retrieved 02 December, 2015, from http://www.kwintessential.co.uk/language/about/urdu.html.

[15] Lovins, Julie B. Development of a stemming algorithm. MIT Information Processing Group, ElectronicSystems Laboratory, 1968.

[16] Mubashir Ali, Shehzad Khalid, Muhammad Haneef Saleemi "A Novel Stemming Approach for UrduLanguage" ISSN: 2090-4274, Journal of Applied Environmental and Biological Sciences, J. Appl.Environ. Biol. Sci., 4(7S)436-443, 2014, www.textroad.com

[17] Qureshi, Anwar & Awan" Morphology of the Urdu Language", International Journal of Research in Linguistics and Lexicography, INTJR-Volume 1-Issue 3, September 2012,

[18] Ruth Lail Schmidt (1999). URDU: AN ESSENTIAL GRAMMER.

[19] Daily Pakistan (2015). "Urdu declared second most popular language among 2301 others." Retrieved 01 December, 2015, from http://en.dailypakistan.com.pk/pakistan/urdu-declared-second-most-popular-language-among-2301-others/.

[20] García, María Isabel Maldonado. "Comparación del léxico básico del Español, el Inglés y el Urdu." Unpublished doctoral dissertation-UNED, Madrid 500 (2013).

[21] Urdu words list got from http://www.cle.org.pk/software/ling_resources/wordlist.htm Retrieved 02 DECEMBER, 2015,

[22] Urdu closed words list retrieved on 09 march 2016 from http://cle.org.pk/software/ling_resources/UrduClosedClassWordsList.htm

[23] Jivani, Anjali Ganesh. "A comparative study of stemming algorithms." Int. J. Comp. Tech. Appl 2, no. 6 (2011): 1930-1938.

[24] Moral, Cristian, Angélica de Antonio, Ricardo Imbert, and Jaime Ramírez. "A survey of stemming algorithms in information retrieval." Information Research: An International Electronic Journal 19, no. 1 (2014): n1.

[25] Moghadam, Fatemeh Momenipour. "Comparative Study of Various Persian Stemmers in the Field of Information Retrieval." Journal of information processing systems 11, no. 3 (2015): 450-464.

[26] W. B. Frakes. (1992). Information Retrieval: Data Structures & Algorithms, Chapter 8,Retrieved 01 October,2015, from http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap08.htm

[27] Melucci, Massimo, and Nicola Orio. "A novel method for stemmer generation based on hidden Markovmodels." In Proceedings of the twelfth international conference on Information and knowledge management, pp. 131-138. ACM, 2003.

[28] Majumder, Prasenjit, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra, and KalyankumarDatta. "YASS: Yet another suffix stripper." ACM transactions on information systems (TOIS) 25, no. 4 (2007): 18.

[29] Anzai, Yuichiro. Pattern Recognition & Machine Learning. Elsevier, 2012.

[30] Husain, M. S., Ahamad, F., & Khalid, S. (2013). A language Independent Approach to develop Urdustemmer.

In Advances in Computing and Information Technology (pp. 45-53). Springer BerlinHeidelberg.

[31] Qureshi, Anwar & Awan" Morphology of the Urdu Language", International Journal of Research in Linguistics and Lexicography, INTJR-Volume 1-Issue 3, September 2012,

[32] Singh, U., Goyal, V. and Lehal, G.S., 2012. Named Entity Recognition System for Urdu. In COLING (pp. 2507-2518).

[33] Riaz, Kashif. "Rule-based named entity recognition in Urdu." In Proceedings of the 2010 named entities workshop, pp. 126-135. Association for Computational Linguistics, 2010.

[34] Board, P. T. (2010). "اردو قواعدوانشاء" for Class-10th. Lahore: Punjab Textbook Board.

[35] Bloch, Dr. Sohail Ahmad (2012), " بنیادی اردو قواعد " , Muqtadrah Qumi Zuban Pakistan, Islamabad.

[36] Haq, Molvi Abdul (1996), "قواعد اردو", Anjaman Tariqi e Urdu, New Dehli (Hind)

[37] UEP (2014)," تخلیق اردو گرائمر", for class 8th, Unique Education Publisher, Urdu bazar Lahore.

[38] A. Jabbar et al. " Effective Urdu Stemmer Based Hybrid Approach". Information Processing & Management (under major revision).

[39] A.Jabbar et al. "A survey on Urdu and Urdu like Language Stemmers and Stemming Techniques" Artificial intelligence review (under minor revision)

[40] Lehal, RohitKansal Vishal Goyal GS. "Rule Based Urdu Stemmer." In 24th International Conference on Computational Linguistics, p. 267. 2012.

Urdu Text Genre Identification

Farah Adeeba, Sarmad Hussain, Qurat-ul-Ain Akram[1]
Center for Language Engineering
[1]ainie.arkram@kics.edu.pk

Abstract

Automatic genre identification of document is becoming ever-increasing important since the availability of more and more text in digital form. This study describes automatic Urdu text genre identification by evaluating state of the art classifiers on different sets of features. The features are extracted by structural and lexical analysis of Urdu text. In addition, term frequency and inverse document frequency of features are computed. Different types of experiments are performed on two types of Urdu text corpora to evaluate the features set and classifiers for automatic Urdu text genre identification system. SVM classifier outperforms irrespective of the features set . The experimental analysis reveals that lexical features are more effective than structural features, and significantly improve the genre identification accuracy.

1. Introduction

Automated genre identification deals with prediction of genre of an unknown text, independent of its topic and style. With the tremendous increase in digital data, automated genre identification is becoming important for information retrieval to classify huge text into different categories. In addition, automatic genre classification of a document also helps to retrieve relevant documents according to the user requirements. It also plays an important role to improve the performance of Natural Language Processing (NLP) applications including part of speech (POS) tagging, parsing and word sense disambiguation system [1].

Urdu belongs to Arabic script and is national language of Pakistan, spoken by more than 100 million people[4]. It has rich morphology i.e. word has many surface forms, requires five agreement (case, gender, respect, number, person**).** In this paper, Urdu text genre classification technique is presented which automatically classifies text into culture, science, religion, press, health, sports, letters and interviews genres. This article investigates the impact of structural and lexical information for the genre identification. In addition, different machine learning techniques including Support Vector Machines(SVM), Naive Bayes and Decision trees are also evaluated. The proposed systems are objectively evaluated by using two different corpora to explore the effect of corpora size in genre identification. A series of experiments show that lexical level approach is quite effective and significantly improves the identification accuracy.

The rest of this paper is organized as follows. Section 2 gives an overview of state of the art techniques for automatic genre identification. Section 3 presents the methodology of the genre identification of Urdu documents, which involves two main phases;(1) extraction of structural and lexical features, and (2) classification of Urdu documents based on extracted features. In addition, text pre-processing is also discussed in Section 3. Urdu benchmark corpora used in this study are described in Section 4. System experiments along with results are presented in Section 5. Finally, Section 6 concludes the research findings of this study.

2. Related Work

Efforts have been carried out for automatic text genre identification using rule based and statistical techniques. The documents have different features which are used to segregate the genre of these documents. These features are used to classify the genres using machine learning algorithms. Classifiers suggest the genre based on the computed features. In literature structural, lexical, sub lexical and hybrid features are being used for genre identification.

Lexical features are usually computed in terms of word frequencies in context of a document, and other documents (TF-IDF). Stamatatos et al. [2] developed a method of text genre identification by using common word frequencies. Wall street journal of the year 1989 is used as dataset consisting of 2560K words, which is further divided into training and testing files. 640k text

[4] http://www.ethnologue.com/language/urd

files are used in training phase and 16k text files are used for testing. For genre classification, 30 most frequent words of British National Corpus (BNC) are used. The Discriminant analysis is performed to extract most frequent words from training corpus. The experimental results show that 30 most frequent words of BNC corpus play important role by giving 2.5% error rate, as compared to the words extracted from training corpora.

Lee and Myaeng[3] proposed genre classification by using word statistics from different class sets, genre and subject class. Goodness value of a term for a genre is computed in two ways; (1) term's document frequency ratios for genre and subject class, and (2) term frequency (TF) ratios. These term's document frequency ratios and term frequency are used to compute the probability of a document belonging to a genre. Dataset of English and Korean languages are used for this study. A total of 7,615 of English and 7,828 documents of Korean are collected from web consisting of reportage, editorial, technical paper, critical review, personal home page, Q&A, and product specification. Naive Bayesian and similarity approaches are used for genre identification. Results show that term frequency (TF) ratios performs well as compared to the document frequency ratios.

Lexical features are also used for Urdu text classification [4] to categorize the text document into six genres. They compute different words based statistical features from corpus. These features are classified using Naive Bayes and SVM. The experimental results show SVM performs well with reasonable genre identification accuracy. Zia et al.[5] also used lexical level information for Urdu text classification using different classifiers including k-NNs, SVM and decision trees. The system has 96% f-measure to label genre of document among one of four genres.

Structural features are usually extracted from POS, phrase related and chunk related information. Lim et al.[6] used structural features for genre identification of web documents. A total of 1,224 documents are used to extract POS, phrase and chunk level features. K-Nearest-Neighbor algorithm is used to classify the genre by using these structural features. The accuracy of the system is 36.9%, 38.6% and 37% for POS, phrase and chunk level features respectively. Such Structural features cannot be applied for resource scarce languages which have very limited annotated language resource.

Classification accuracy has direct relationship with dataset size. The change in sample size results into direct change in classification accuracy. Sordo and Zeng[7] investigated the dependency between sample size and classification accuracy. It is observed that classification accuracy increases with the increase in dataset size. Irrespective of size, dataset itself plays an important role for genre classification.

Accurate genre classification needs minimal overlapping between genres' lexical and contextual information. Kanaris and Stamatatos[9] used two different corpora for genre identification. Both corpora have different accuracies.

In addition, number of genres has significant impact on genre classification. Limited number of genres results into higher accuracy. Moreover, significant training data also improve the genre identification accuracy.

Ali and Ijaz [4] used lexical features for Urdu text classification using SVM and Naive Bayes methods. Maximum accuracy i.e. 93.34% is achieved by using SVM. Presented system classify document in one of six categorize. Zia et. al. [5] used different feature selection approaches for Urdu text classification. Lexical features are used for classification of document into four genres. The available techniques for Urdu are mainly focused on the evaluation of machine learning classification approaches by using lexical features. In this paper, the effect of lexical and structural features is observed by applying different machine learning classification approaches. In addition, two different text corpora are used to investigate the impact of training dataset size variability on genre classification.

3. Methodology

In this paper, impact of lexical and structural features is analyzed for the development of Urdu text genre identification system. Urdu words unigram and bigrams are extracted as lexical features whereas words POS and words sense information is used as structural information. These features are classified using state of the art machine learning classification approaches. The details of each phase of the presented technique are discussed in sub sequent sections.

3.1 Text pre-processing

In this paper, two different datasets are used which are discussed in detail in Section 4. Dataset-1 is cleaned, POS tagged, sense tagged corpus which is manually distributed into respective genres by expert linguists. Dataset-2 is large corpus in size which is only manually distributed into genres by linguists.

Therefore, preprocessing is applied to extract features from these corpora. The details are given below.

3.4.1. Corpus cleaning.

As mentioned above, Dataset-1 is cleaned therefore Urdu words unigram and bigram extraction is straight

forward. Words are extracted by tokenizing the corpus on space. But, in Dataset-2 the extracted words list has some issues. This list is manually analyzed and an automatic corpus cleaning is developed by applying different heuristics to clean the corpus. Some examples of Urdu space insertion issues are discussed below. Heuristics to resolve these issues are also discussed.

1. عرصہ۳۰ سال سے پی ٹی سی ٹیچر
2. دنیا بھر میں موجود 65کے لگ بھگ باقاعدہ اوپن یونیورسٹیوں کے ساتھ
3. برطانیہ میںUKOU کے قیام کے محض تین سال بعد
4. HKEY_LOCAL_MACHINE\SOFTWARE\Microsost\Windows\CurrentVersion\Explorer\BitBucketاب

Due to missing space between Urdu digits and text, the two words e.g. ۳۰ سال are treated as single word. Same type of issue is observed between sequence of Latin digits and Urdu text e.g. 65کے.The space is inserted at start and end of sequence of Urdu/Latin digits to resolve this issue. There are also examples of joined Latin character sequence with Urdu text e.g. میںUKOU in Example 3. These Latin strings can be URLs as shown in example 4.

Hence to resolve this issue, space is automatically inserted at start and end of Latin character sequence. In addition, space is inserted between normal text and punctuation marks so each punctuation mark is considered as individual word.

After resolving space insertion issues, Dataset-1 and Dataset-2 is further processed to add tags which are useful to process lexical features. The complete URL is replaced with special word tag. To give tag to the web URL, the corpus is processed and web URLs are extracted by using regular expression. The complete web URL are replaced with special tag i.e. "httpaddr". In the same manner, email address is extracted using regular expression and replaced with "emailaddr" tag. Moreover, Latin cardinal number strings are extracted and replaced with a tag as "CD". These special word tags help to manipulate the lexical information for Urdu genre identification.

3.4.2. Stemming.

Urdu is morphological rich language i.e. a word may have more than one surface forms resulting into need of large amount of data containing all surface forms of the words so that machine learning classifier can better learn all forms. The requirement of this huge amount of dataset due to multiple surface form is resolved by applying Urdu stemmer. Before extracting words unigrams and bigrams from Dataset-1 and Dataset-2, Urdu stemming algorithm [10] is applied on both datasets. For better results of Urdu stemmer [10], all closed class words are extracted from both datasets using Urdu closed class list[5].

3.4.3. POS tagging.

Dataset-1 is manually annotated with POS tags [11] (details are given in Dataset Section). Dataset-2 is large corpus as compared to Dataset-1 and is not annotated with POS tagged. Manual annotation of this corpus with POS tags is troublesome task. Hence, an automatic Urdu POS tagger is used [11] to automatically tag the Dataset-2 so that word POS features can be extracted.

3.2. Features extraction

In any machine learning based NLP application, features play an important role to improve the accuracy of the application. In the same way, for the development of automatic genre identification, text document is processed to extract useful feature set which better distinguishes the genre of the respective document. For the development of Urdu genre identification system, two different types of features are extracted; (1) lexical features, and (2) structural features. To extract lexical features, words unigram and bigrams are computed along with their Term Frequency (TF) and Inverse Document Frequency(TF-IDF). These lexical features are separately extracted from both datasets.

To compute structural features, POS and sense information of word is used. Word along its POS feature set is computed from both datasets where as word with its sense information is only extracted from Dataset-1 as only this dataset is manually annotated with sense tags. Structural level information is used for experimentation of Dataset-1 to investigate the impact of word POS and sense on genre identification accuracy.

For dimensionality reduction low frequent terms are discarded. To see the impact of features set on genre identification, separate systems are developed for each feature, presented in Table 1. Details of number of features in each feature set are given in Section 5.

3.3. Classifiers

[5]http://cle.org.pk/software/ling_resources/UrduClosedClassWordsList.htm

The features are used to train the machine learning classification algorithms. The features are extracted from training data. These features along with label of the genre class are fed to the classifier in the training phase.

Table 1: Systems for the Urdu genre identification

System	Features
System 1	Word Unigram
System 2	Word Bigrams
System 3	Word/POS
System 4	Word/Sense

In the recognition, the features of the input document are computed and then based on the learning model, classifier predicts genre. In this study, state of the art classification algorithms including Support Vector Machines, Naive Bayes and C4.5 are used for Urdu genre identification system. Each classifier is separately trained on each of the features-based system (Table 1). The results are discussed in Section 5.

4. Dataset

Two different datasets are used for Urdu text genre identification. These are (1) CLE Urdu Digest 100K [8]named as Dataset-1, and (2) CLE Urdu Digest 1 Million named as Dataset-2.

Dataset-1 is a balanced corpus having 100K Urdu words which is collected from multiple genres of Urdu Digest corpus. This corpus is manually cleaned by linguist to resolve space insertion and space deletion issues. In addition, same corpus is manually annotated with POS tags by linguist [11]. The complete Urdu POS tagset along with guidelines for corpus annotation is defined. A total of 35 POS tags are defined in Urdu POS tagset.

Dataset-2 is 1 million Urdu words corpus, distributed into multiple domains. This corpus is automatically cleaned using the heuristic discussed in Corpus Cleaning Section. In addition, Dataset-2 is automatically annotated with Urdu POS tagset using POS tagger [11] which has 96.8% accuracy. The motivation behind using two different version of CLE Urdu digest is to investigate the effect of dataset size on the accuracy.

Eight genres i.e. culture, science, religion, press, health, sports, letters and interviews of both datasets are used for classification experiment. Details of training and testing documents of Dataset-1 and Dataset-2 against each genre are presented in Table 2.

5. Experiments and results

The lexical and structural features from Dataset-1 and Dataset-2 are extracted. Each feature set is labeled with different system and is evaluated separately to analyze its impact on genre identification accuracy. The number of features extracted from training data of Dataset-1 and Dataset-2 for each system are provided in Table 3. For Dataset-1, 228 training and 56 testing documents are used. While, Datset-2 includes 686 and 160 documents for training and testing, respectively (Table 2).

The features extracted from training data are used to train each classifier for Dataset-1 and Dataset-2 separately. Testing data is used to test the performance of each system trained by respective classifier.

Table 2: Datasets

Genre	Data Set 1		Data Set 2	
	Training Document	Testing Document	Training Document	Testing Document
Culture	34	8	120	30
Science	45	10	98	21
Religion	23	6	95	20
Press	23	6	94	24
Health	23	6	129	31
Sports	23	6	25	6
Letters	28	7	90	21
Interviews	30	7	35	7
Total	**229**	**56**	**686**	**160**

Table 3. Number of features in Dataset-1 and Dataset-2

System	Features	No. of features for Dataset-1	No. of features for Dataset-2
System 1	Word Unigram	156	1,665
System 2	Word Bigrams	316	4,798
System 3	Word/POS	1,037	6,548
System 4	Word/Sense	1,570	.....

The accuracy measures including Precision(P), Recall(R) and F-measure(F) are computed to evaluate accuracy results. Recall(R) is the number of correctly classified documents divided by the number of total documents. Precision(P) is the number of correct classifications divided by the number of classification made whereas F-measure(F) is computed by using the following equation

F = 2 * (Precision*Recall) / (Precision + Recall).

Each classifier is trained separately on a system of each dataset, excluding System 4 which is trained and tested only for Dataset-1. The accuracy results of SVM, Naive Baye and C45 classifiers are presented in Table 4, Table 5, and Table **6**, respectively. Dataset-2 having more training examples gives better results as compared to the Dataset-1 for each system and each classifier. Moreover, results reveal that lexical features provide higher accuracy as compared to the structural features. The System 2 i.e. word bigrams yields higher precision, recall and f-measure as compared to the other systems. It has been observed from the results that SVM outperforms the other classifiers irrespective of feature type.

Table 4. Systems results using SVM

System	Dataset-1			Dataset-2		
	P	**R**	**F**	**P**	**R**	**F**
System 1	0.50	0.50	0.48	0.68	0.68	0.67
System 2	0.38	0.33	0.35	**0.74**	**0.70**	**0.70**
System 3	0.63	0.62	0.62	0.72	0.68	0.68
System 4	0.53	0.35	0.38	...	...	...

Table 5. Systems results using Naive Bayes

System	Data Set 1			Data Set 2		
	P	**R**	**F**	**P**	**R**	**F**
System 1	0.45	0.37	0.37	0.68	0.67	0.66
System 2	0.37	0.39	0.37	**0.70**	**0.7 0**	**0.69**
System 3	0.59	0.58	0.58	0.67	0.65	0.63
System 4	0.34	0.35	0.32	...	...	...

Table 6. Systems results using C4.5

System	Dataset-1			Dataset-2		
	P	**R**	**F**	**P**	**R**	**F**
System 1	0.34	0.32	0.32	0.45	0.45	0.45
System 2	0.44	0.41	0.42	**0.47**	**0.45**	**0.46**
System 3	0.46	0.44	0.43	0.44	0.44	0.43
System 4	0.171	0.179	0.161	...	...	...

In addition, after doing detailed analysis, it has been observed that some texts in the genres are overlapping e.g. science and health which results in misclassification. The genre which is not overlapping e.g. sports has 100% genre identification accuracy.

6. Conclusion

In this paper, automatic Urdu text genre identification system has been presented by evaluating the impact of lexical and structural features along with different state of the art classifiers. From the results, it is observed that lexical features are most appropriate for identifying the genres of Urdu documents. In addition, results indicate that SVM has an advantage over other classifiers irrespective of feature type. The size of the training data also affects the accuracy. It is reinforced that significant amount of training data improves the document classification accuracy.

References

[1] B. Kessler, G. Bumberg and H. Schütze, "Automatic detection of text genre," in *in proc. European Chapter*

of the Association for Computational Linguistics (EACL), Madrid, Spain, 1997.

[2] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Text genre detection using common word frequencies," in *proc. Conference on Computational linguistics*, Stroudsburg, PA, USA, 2000.

[3] L. Y. and M. S. H., "Text genre classification with genre-revealing and subject-revealing features," in *proc. international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2002.

[4] A. R. Ali and M. Ijaz, "Urdu text classification," in *pro. International conference on Frontiers of Information Technology*, 2009.

[5] T. Zia, Q. Abbas and M. P. Akhtar, "Evaluation of Feature Selection Approaches for Urdu Text Categorization," *International Journal Intelligent Systems and Applications,* pp. 33-40, May 2015.

[6] C. S. Lim, K. J. Lee, G. C. Kim, K. Su, J. Tsujii, J. Lee and O. Kwong, "Automatic genre detection of web documents," in *Natural Language Processing-IJCNLP* , Berlin, Heidelberg, Springer Berlin Heidelberg, 2004, pp. 310-319.

[7] S. M. and Z. Q., "On sample size and classification accuracy: a performance comparison," in *proc. International conference on Biological and Medical Data Analysis* , Berlin, Heidelberg, 2005.

[8] S. Urooj, S. Hussain, F. Adeeba, F. Jabeen and R. Parveen, "CLE Urdu Digest Corpus," in *proc. Conference on Language and Technology*, Lahore, Pakistan, 2012.

[9] K. I. and S. E., "Webpage genre identification using variable-length character n-grams," in *proc. IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, 2007.

[10] Q. Akram, A. Naseer and S. Hussain, "Assas-band, an Affix-Exception-List Based Urdu Stemmer," in *in the Proc. of the 7th Workshop on Asian Language Resources*, Suntec City, Singapore, 2009.

[11] T. Ahmed, S. Urooj, S. Hussain, A. Mustafa, R. Parveen, F. Adeeba, A. Hautli and M. Butt, "The CLE Urdu POS Tagset," in *poster presentation in Language Resources and Evaluation Conference(LERC 14)*, Reykjavik, Iceland., 2014.

Urdu Phonological Rules in Connected Speech

Mahwish Farooq and Benazir Mumtaz
Center for Language Engineering,
University of Engineering and Technology, Lahore

Abstract

The present work deals with the phonological rules in Urdu language. All these rules have been reported by considering the multiple pronunciations of a word, which has same spellings and parts of speech (POS). For the confirmation of multiple pronunciations, firstly a word list of 13717 words has been extracted from 10 hours speech corpus of a female native Urdu speaker. Secondly, in order to confirm whether these multiple pronunciations are speaker dependent or language dependent, data from 9 more native speakers have been collected for the confirmation of multiple pronunciations. In this paper, phonological rules related to the segment alternation, segment deletion and segment insertion have been investigated. Analysis reports that (i) segment alternation occurs due to stress, (ii) unstressed articulation causes segment deletion and (iii) segment insertion emerges to break consonant cluster at coda position.

Keywords—Urdu phonological rules, multiple pronunciations, segment deletion, segment insertion, segment alternation, syllabification and restructuring

1. Introduction

Urdu is an Indo-Aryan language and it has 100 million speakers in all over the world and they have multiple pronunciations and accents

[**1**]. In this study, Urdu phonological rules are reported based on multiple pronunciations of a word, which has same spellings and part of speech. For example, a word محبت (love /məhəbbət̪/)

[**2**] has two more alternative pronunciations /mʊhəbbət̪/ and /mohəbbət̪/ and both of these pronunciations are equally intelligible among native Urdu speakers. The motivation of this study is to investigate the phonological phenomena behind these alternative pronunciations. Phonological rules mean the information of possible and non-possible combinations of sounds in a language. The phonological rules also give information about the alternative or multiple pronunciations of a word

[**3**]. In other words phonological rules deal with the words' morphology and concern with the way in which morphemes combine to form a meaningful word.

[**4**]. Studies reported that phonological variations are inevitable and unconsciously used by the native speakers

[**5**]. Sometimes the multiple pronunciations raised controversy and confusion for the language users. Therefore, there is a big need to find out the types of variations and their possible reasons based on Urdu phonology. In connected speech production, Urdu has sound change rules almost similar to other languages i.e. alternation, assimilation, deletion, vowel lengthening, etc

[**6**]. These sound change rules have produced multiple pronunciations of different surface form of an already existed phonetic script. However, the present study deals with only three important phonological rules; (i) segment alternation, (ii) segment deletion and (iii) segment insertion. All these phenomena have created multiple pronunciations therefore these issues are resolved by finding out the backend strategy of language and language users. It is also reported that few variations are speaker dependent and some are context dependent.

This paper proposes Urdu phonological rules in connected speech. The remaining paper has been arranged accordingly: studies on the phonological rules of different languages are reported in Section 2, Section 3 presents the methodology of this study, Section 4 reports results, section 5 is about data analysis and discussion of the proposed phonological rules in Urdu language, while the conclusion and future discussion are presented in Section 6.

2. Literature review

All languages (such as English, Czech, Japanese, Shona, Hungerian, Finish, Setswana, Dutch, Russian,

etc) have different phonological rules based on voice quality

[**7**]. In order to acoustically analyze phonological rules of a speech, voice quality is an important factor because it is directly affected by the habitual variation of speakers' vocal apparatus. These variations contributed in accent variations and multiple pronunciations of a language. According to American National Standards Institute (ANSI) voice quality could be used for differentiating the speech variations which are based on the momentary actions of speech segments

[**8**]. Every language has a unique and stereotypical speech segments i.e. consonant, vowel and approximant [**4**]. In connected speech production, these segments lose their individual features because one segment is coarticulated with other segments like a connected string of sounds.

[**9**] By taking or losing its individual features. The connected speech production is a complicated phenomenon because different segmental and supra-segment factors are involved in articulation [**10**]. Moreover, Vander reports that the motivation behind multiple pronunciations is also based on the attitude of the language users i.e. hypercorrection and overgeneralization [**11**].

According to sound change theory, it is inevitable to control sound changes in an utterance [**12**] as these are inherent variations and are called "non-programmed features" of alternative pronunciation [**5**]. This is clear by a research where a single speaker has repeated a single word 10,000 times but he would not been able to produce an exactly similar utterance. Although these utterances were similar based on the natural sound class but were different from each other based on the discrete sound features [**3**]. These segmental features are not enough because connected speech production is a complicated process [**10**]. Therefore, auditory transcription has a drawback that it cannot generate exact reproduction of human speech by using traditional phonetic symbols. These pronunciation differences are the part of phonetic grammar of a language [**3**]. The phonetic grammar is based on the phonological rules of a given language. There are number of phonological rules existed in different languages of the world i.e. deletion, insertion, alternation, assimilation, nasalization, aspiration, voicing, etc [**13**]. But only segment alternation, segment deletion and segment insertion in different languages have been discussed in the subsequent sections.

2.1. Segment alternation

Segment alternation is a basic principle for multiple pronunciations. In connected speech, shuffling of one sound with another sound is called segment alternation [**4**]. In Hindi language, according to one rule, nasal consonant converts preceding oral vowel into a nasal vowel [**14**].

In German and Czech languages, word final voiced obstruents converted into voiceless stops i.e. /hʊnd/ as /hʊnt/. In some Spanish dialects, voiced stops become fricatives if surrounded by vowels [**7**]. In Turkish language, syllable final devoicing of voiced consonant has also been reported but this is not equally applicable to voiced fricative and sonorant [**15**]. In Farsi (Persian) language, /r/ phoneme appears in three allophonic forms [r], [r̥] and [ɾ]. These forms are dependent to the phonological environment where the sound comes [**4**]. In Lithuanian language, assimilation of voicing and devoicing is also common [**3**].

2.2. Segment deletion

Deletion of a phonemic unit is called segment deletion. It is a very common phenomenon in connected speech production [**13**]. It occurs due to the laziness of people in articulation process [**16**]. In Hindi language, schwa and a nasal consonant are deleted when they are preceded by an oral vowel or a nasal vowel respectively [**14**]. In English language, according to the relative functional load (RFL) phenomenon, if the syllable final consonants /t, d, n/ are followed by an unstressed /l/ or /n/ then the latter consonants would take the quality of a complete syllable by the deletion of preceding schwa [**17**]. Moreover, it reports word final /ə/ deletion if it is followed by a stressed syllable [**18**]. In Turkish language, syllable medial and syllable final velar /g/ phoneme is deleted by converting the preceding short vowel into a long vowel [**15**].

2.3. Segment insertion

In a connected speech articulation, adding up of a phoneme is called segment insertion [**19**]. Articulation time [**7**] and speakers' attitude are the major elements for insertion [**11**]. Turkish language has complex consonant clusters both at onset and coda position. Same language behavior has been observed even in the articulation of English by the native Turkish people as in the word 'group' (collection of entities /grʊp/) which became /gʊrʊp/ [**15**]. In Armenian language, initial consonant cluster have been splitted by the insertion of /ə/ vowel. In Lomongo language, /j/ insertion took place in compound words [**3**].

2.4. Urdu phonological rules

In Urdu language, different phonological rules have already been discussed by other researchers. According to Hussain, Urdu has (i) nasal assimilation, (ii) velar assimilation, (iii) bilabial assimilation and (iv) /h/ deletion [**6**]. Akram has reported /ə/ insertion in order to control multiple onsets in a syllable [**20**] and Nawaz reported /ʔ/ deletion in Urdu speech articulation [**18**]. The previous studies on Urdu phonological rules only focus on the segmental feature analysis. However, this study reports that segment features are not enough to deal with the actual pattern of multiple pronunciations in Urdu language. Phonology of connected speech in Urdu is dependent on different factors i.e. segmental features, context, stress pattern, syllabification and restructuring [**10**].

3. Methodology

Urdu phonological rules are extracted from the speech of 10 speakers. Multiple pronunciations have been observed in their speeches which are different from the standard pronunciation of the words. In order to confirm; whether these pronunciations are mispronunciations or multiple pronunciations, firstly, 10 hours speech corpus of a native Urdu female speaker is studied for the initial analysis. This speech corpus is extracted from three corpora i.e. 35 million words' corpus; CLE Urdu Digest Corpus 1M and 2.6 million words' corpus of Urdu news [**21**]. The speech obtained from the corpora is annotated at multiple levels i.e. phoneme, syllable, word, phrase, stress, utterance or sentence levels using Case Insensitive Speech Assessment Phonetic Alphabet (CISAMPA) method [**22**].

Secondly, these phonological variations have been confirmed by obtaining the data from 9 native (7 males and 2 females) speakers of Urdu. All these speakers were graduates and use Urdu and Punjabi in their daily routine. Educated native speakers are deliberately selected in order to confirm; whether literacy plays any role in standard pronunciation or not.

10 hours corpus is comprised of 103902 words containing 9852 unique words, 13717 duplicates and 80333 English loan words and Urdu functional words. For in this research, only duplicate words list is used for further research in the alternative pronunciations of standard vocabulary.

The word list of duplicates provides multiple instances of a word with same spellings including their transcription, parts of speech, number of syllables in a word, stress pattern and file ID.

Analysis of word list highlights that variations may occurred due to four reasons; (i) it might be an annotation error, (ii) mismatches may occur due to homographs or homophones having different parts of speech, (iii) mismatches may occur due to different stress patterns of a word in different files and (iv) variation may occur due to alternative pronunciations. In this research, first two types have been ignored but only third and fourth types have been considered for the confirmation of multiple pronunciations. The standard pronunciations of vocabulary have been confirmed by using "Urdu Lughat: Tarixi Usuul Per" [**2**] and the meanings in English have been incorporated by consulting Oxford Urdu-English Dictionary [**23**].

4. Results

After analyzing the word list of multiple pronunciations of words with reference to the text-grid files, it is concluded that speaker has articulated same words with multiple pronunciations. Detailed results are given in the table 1; where TW means total words, SP stands for standard pronunciation and AP is used for alternative pronunciation.

Table 1: Single Speaker Speech Analysis Report

Total Number of Alternative (Duplicate) Words = 13717																		
	Segment Alternation							Vowel Deletion				Consonant Deletion			Vowel Insertion			
	Short-to-Short Vowel T=2530			Short-to-Medial and Medial-to-Short Vowel T=458		Medial-to-Medial Vowel	Long-to-Long Vowel	Disyllabic Word		Tri-syllabic Word		Polysyllabic Words			Monosyllabic Words T=633 Insertion=317			
								Short Vowel T=202		Short Vowel	Medial Vowel							
	ə→ ɪ	ə→ ʊ	ɪ→ ə	e→ ə	e→ ɪ	e→æ	e:→ æ:	ə→ φ	ɪ → φ	ə→φ	e → φ	/j/	/h/	/v/	before /l/	before /r/	before /s/ or /z/	
TW	850	850	830	165	293	448	124	157	45	79	77	4403	4742	29	210	200	211	
SP	550	300	779	95	154	124	28	93	21	15	15	1578	806	3	100	100	106	
AP	300	550	51	70	139	324	96	64	24	64	62	2825	3936	26	110	100	105	

In order to confirm, these variations are speaker dependent or context dependent, 9 more native speakers have been selected for recordings. For this purpose, a mini corpus of 75 words' list (25 words of each category) has been selected for analyzing multiple pronunciations.

Recording of these words have is carried in a carrier sentences to avoid stress and boundary effects. The speech is recorded and annotated in PRAAT software using same methodology presented in [**22**]. Later on, these words and their multiple pronunciations are cross checked among 9 native speakers' speech. Detailed calculations are given in table 2.

Table 2: Nine Speakers Speech Analysis Report

Total Number of Alternative Words = 75																	
	Segment Alternation							Vowel Deletion				Consonant Deletion			Vowel Insertion		
	Short-to-Short Vowel T=25			Short-to-Medial and Medial-to-Short Vowel T=25		Medial-to-Medial Vowel	Long-to-Long Vowel	Disyllabic Word Short Vowel T=25		Tri-syllabic Word Short Vowel	Tri-syllabic Word Medial Vowel	Polysyllabic Words			Monosyllabic Words T=25 Insertion		
	ə→ ɪ	ə→ ʊ	ɪ→ ə	e→ ə	e→ ɪ	e→æ	e:→ æ:	ə→ φ	ɪ → φ	ə→ φ	e → φ	/j/	/h/	/v/	before /l/	before /r/	before /s/ or /z/
SP 2	13	15	10	7	13	24	06	14	24	14	12	15	16	17	17	25	15
SP3	13	15	11	17	12	20	12	22	7	7	9	13	7	7	19	18	25
SP4	24	14	11	11	19	13	22	22	17	17	15	19	17	17	23	22	22
SP5	14	18	12	23	15	13	11	19	11	18	7	18	11	11	20	22	12
SP6	17	22	10	22	11	16	21	12	23	23	17	7	23	23	23	23	24
SP7	18	20	11	12	16	18	18	18	22	22	11	17	21	22	22	22	14
SP8	19	22	13	11	17	10	13	17	12	12	23	11	24	24	19	24	18
SP9	22	20	11	12	15	10	12	22	18	9	17	23	13	23	19	23	19
SP10	5	11	19	14	14	10	14	24	12	15	19	22	18	23	20	25	14
TN	145	157	108	129	132	134	129	170	146	137	130	145	150	167	182	204	163
%age	64	70	48	56	59	60	57	75	65	61	58	64	67	74	81	91	72

5. Data analysis and discussion

Like many other languages, Urdu also has sound change rules, which become the cause of multiple pronunciations of an already existed phonetic script. Data analysis confirms that there are three main categories of alternative pronunciations of the same vocabulary. Those are:

1. Segment Alternation
2. Segment Deletion
3. Segment Insertion

5.1. Segment alternation

Phonemic alternation occurs due to the shuffling of one sound with another. The first principle for multiple pronunciation is segment alternation; "except in case of suppletion, every morpheme has only one phonological form. Any variation in the phonetic shape of a morpheme results from the operation of regular phonological rules". According to the definition, morphology does not allow alternative pronunciations of a segment but phonology supplies the information at which context a segment could alternate its stereotypical features. These phonologically variant segments are called "alternants" [**4**].

Urdu also has different "alternants" but native speaker articulated one "alternant" at a time. According to the present data analysis, Urdu native speakers switch between multiple pronunciations by substituting one vocalic segment with another. This alternation occurs at four levels;

(i) Short to short vowel alternation
(ii) Short to medial and medial to short vowel alternation
(iii) Medial to medial vowel alternation
(iv) Long to long vowel alternation

All these alternations are discussed in the subsequent sessions. However, the reasons of first two types are not discussed, as data indicates they might be speaker dependent variations.

5.1.1. Short to short vowel alternation

First condition is short-to-short vowel alternation; it occurs when one short vowel alternates with another short vowel e.g. in the word بلند (high /bələnd̪/), /ə/ is converted into /ʊ/ and formed an alternative pronunciation /bʊlənd̪/.

5.1.2. Short to medial and medial to short vowel alternation

Second condition is medial to short vowel and short to medial vowel alternation; it occurs when a medial vowel substitutes with a short vowel or a short vowel alternates with a medial vowel e.g. the word اختلاف (conflict /ɪxt̪ela:f/) has two multiple pronunciations /ɪxt̪əla:f/ and /ɪxt̪ɪla:f/. The word شاعر (poet /ʃa:ɪr/) has two multiple pronunciations; the

standard pronunciation /ʃa:ɪr/ and other alternative pronunciation with the medial vowel /ʃa:**e**r/.

In the first example, medial vowel substitutes with a short vowel. In the second example, short vowel alternates with a medial vowel. Sometimes phonemic alternation also causes change in syllabification of the word by taking diphthong form. Mostly short and medial vowels substitute in this order i.e. /e/ medial vowel substitutes with short vowel [ə/ɪ] while /ʊ/ short vowel substitutes with medial vowel /o/ and vice versa.

5.1.3. Medial to medial vowel alternation

Third condition is; medial vowel alternates with another medial vowel e.g. the word احترام (respect /ehʈera:m/) has another alternative pronunciation /æhʈera:m/. In polysyllabic words, if the letters الف، ح come together at word initial place as in the word احسان (good deed /ehas:n/) such type of words have standard transcription with /e/ medial vowel but speakers have alternated /e/ medial vowel with /æ/ medial vowel and the same is the case with the word احتجاج (protest /ehʈeʤa:ʤ/) which has another alternative pronunciation /æhʈeʤa:ʤ/. Condition for their alternations is;

i. In polysyllabic words, if the word is articulated with stress then /e/ medial vowel would be substituted with /æ/ medial vowel.

5.1.4. Long to long vowel alternation

Fourth condition is the long vowel alternation with long vowel as in the word تیئیس (twenty three /ʈe:i:s/). It has two pronunciations; one is the standard one /ʈe:i:s/ and the other is the alternative pronunciation /ʈæ:i:s/ of the same word (for more examples see appendix).

In polysyllabic words, this phenomenon has been commonly observed both at word initial and word medial positions. Especially, if the letters الف and ع co-occur at word initial position as in the word اعتبار (Trust /eʔʈeba:r/), this would not be wrong if we take /e:/ long vowel as a standard segment [**18**]. The conversion of long vowel /e:/ with the long vowel /æ:/ occurs in the polysyllabic words;

i. When stress /e:/ long vowel is substituted with /æ:/ long vowel.

5.2. Segment deletion

In a connected speech, segment deletion of a phoneme is also called elision. It is common in casual connected speech [**13**] which causes re-syllabification [**24**]. According to Waqar and Waqar speakers deleted phonemic segments due to their laziness which is another factor, responsible for the change in pronunciation [**16**] e.g. the word بسر (to live /bəsər/) has another alternative pronunciation as /bəsr/. Vowel deletion reduces number of syllables as well. Different types of phonemic deletions are observed in this research; (i) short or medial vowel deletion, (ii) /h/ deletion, (iii) /j/ deletion and (iv) /v/ deletion.

1. Segment deletion always occurs at word medial or word final syllable but never at word initial position.
2. Sometimes consonantal deletion converts its preceding short vowel into long vowel e.g. in the word حصہ /portion/, hıssəh/ changes into /hıssa:/
3. Long vowel deletion is not possible.
4. Short or medial vowel deletion has been observed in disyllabic and tri-syllabic word.
5. Stress plays an important role in segment deletion.

i. Unstressed articulation causes vowel deletion in bi and tri-syllabic (polysyllabic) word.
   a. By reducing stress in disyllabic words, firstly short vowel deletion occurs in the last syllable then syllabic reformation takes place. The re-syllabification occurs due to consonant clusters at coda position. For example, the word امر (eternal /ə.mər/) converts into /əmr/.
   b. Vowel deletion occurs in tri-syllabic (polysyllabic) words due to unstressed articulation of the penultimate syllable of the word, which not only causes segment deletion but also becomes reason for reformation of syllables in the word. This phenomenon is called vowel syncope [**25**]. Vowel is the nucleus of the syllable therefore vowel deletion demands re-syllabification [**9**]. It is a complicated process as it follows phonotactic rules of the language [**24**] e.g. Urdu phonotactic rules do not allow consonant cluster (/df/, /ʈm/, /ʈb/, /xr/, etc) at word initial position [**20**] and same is the case at syllable initial position. For example the word آخرت (hereafter /a:xɪrəʈ/) converts into /a:xrəʈ/.

ii. /h/ deletion occurs at word final position if it is articulated in connected speech without stress as the word بادشاہ (king /ba:ɖʃa:h/) turned into /ba:ɖʃa:/ and بچہ (child /bəʧʧəh/) converted into /bəʧʧa:/ [**6**].

iii. Usually, /j/ deletion occurs word medially to form a diphthong e.g. the word کیوں (why /kɪjũ:/) as /kɪũ:/ and کیا (what /keja:/) as /kæa:/ [**26**]. However in some cases /j/ deletion occurs without

making diphthong as in the word حیثیت (status /hæːsɪjjəʈ/) as /hæːsɪːəʈ/ and لیے (for /lɪje:/) as /lɪeː/

iv. /v/ deletion occurs by the substitution of /v/ consonant with the vowel. /v/ deletion occurs inter vocalically in two ways; by making diphthongs i.e. the word ہوئی (was /hʊvi:/) converts into a monosyllabic word /hu:i:/ [**26**]. While on the other hand, unstressed articulation also causes /v/ deletion, without making diphthong as in the word ہندووں (Hindues /hɪnɖʊvõ:/) v deletion occurs without making a diphthong /hɪnɖuːõ:/

5.3. Segment insertion

The addition of a phonemic segment in a word is called insertion or epenthesis [**19**]. Articulation time of articulators is the major reason for the segment insertion [**7**] and it may be speakers' attitude i.e. hypercorrection and generalization about rules because people overdo things when they like and dislike them [**11**]. In Urdu connected speech, the segment insertion, especially the insertion of /ə/ has been commonly observed phenomenon among ten speakers' speech. Multiple pronunciations of monosyllabic words occur due to the insertion of a short vowel which ultimately increases number of syllables in a word.

Syllable is factorable unit of the word which associates the linear string of segments in a structure [**20**]. For example, the word امر (work) has two multiple pronunciations; one is the standard pronunciation /əmr/. The other is the alternative pronunciation /əmər/ with /ə/ insertion and syllabic reformation. This insertion might be the effect of over generalization of the word امر (eternal, /əmər/).

1. Vocalic segment insertion (only short vowel /ə/) takes place in order to break word final consonant cluster and this insertion happens in three contexts which are as follows;
   a. If consonant is followed by a liquid sounds /l/ or /r/ e.g. قبر (grave /qəbr/) as / qəbər/ and اصل (original /əsl/) as /əsəl/.
   b. If consonant is followed by a bilabial nasal sound /m/ e.g. in the word کرم (fate /kərm/) as /kərəm/.
   c. If consonant is followed by an alveolar fricative consonant /s/ or /z/ e.g. in the word حبس (congestion /həbs/) as /həbəs/.

It is confirmed after analyzing speech corpus that multiple pronunciations of words occur due to different phonological rules in Urdu language. All these reported rules are discussed and marked after taking consents from Urdu native speakers.

It is observed that in connected speech production (i) phonological variations occur only in open class words i.e. noun, adjective etc (ii) unstressed articulation causes segment deletion of /ə/, /h/, /j/ and /v/, (iii) segment deletion always occurs in disyllabic or trisyllabic words (iv) segment deletion always occurs at word medial or word final position (v) sometimes consonantal segment deletion converts preceding short vowel into long vowel and (vi) long vowel deletion is not possible. Moreover, (vii) segment insertion took place in consonant clusters at coda position when a consonant is followed by liquid sound, bilabial nasal sound or an alveolar fricative. It is also noticed that (viii) segmental alternations have occurred due to stress, (ix) speakers' education is not the guarantee for the articulation of standard pronunciation.

6. Conclusion and future discussion

This research presents phonological rules related to segment alternation, deletion and insertion in Urdu speech. Using these rules, the existed Urdu lexicons can be updated as they give only morphological information of the word without incorporating new language changes. Incorporation of phonological information will be help in finding out alternative pronunciations of the word.

There are other issues as well which have not been discussed here but would be investigated in future research. This includes study of short vowel insertion in polysyllabic Urdu words; alternative selection of short or medial vowel in a word, /h/ deletion at word medial position and multiple pronunciations of proper nouns. Moreover, the role of socio-cultural and educational background of the person in multiple pronunciations would also be studied in future.

7. References

[1] M. Farooq, *An Acoustic Phonetic Study of Six Accents of Urdu in Pakistan*, M. Phil Thesis, Department of Language and Literature,: University of Management and Technology, Johar Town, Lahore, Pakistan, 2015.

[2] *Urdu Lughat: Tarixi Usuul Per*, 1st ed, 3rd Publication, Vol. 1 Muheet Urdu Press Karachi, Pakistan, June 2013.

[3] D. Odden, "Feature Theory," in *Introducing Phonology*. Cambridge, , NewYork: Cambridge University Press, United States, 2005, ch. 7, pp. 129-168.

[4] J. T. Jehsen, "Distinctive Features," in *Principles of Generative Phonology*, iv ed., Konrad E.F. Korener, Ed. The Netherlands, USA: John Benjamins Publishing Company, Amsterdam/Phila Delphia, 2004, ch. 1, pp. 79-106.

[5] O. J. John,.: Phonology Laboratory, Department of

Linguistics, University of California, Berklay, California, pp. 75-94.

[6] S. Hussain, "Phonological Processing for Urdu Text to Speech System," *Localisation in Pakistan, in Localisation Focus: The International Journal for Localisation*, vol. 3(4), 2005.

[7] J. Panevov and J. Hana, Intro to Linguistics _ Phonology, October 13, 2010.

[8] K., Jody; Sidtis, D. Vanlancker; Gerratt, Bruce, "Defining and Measuring Voice Quality," in *Sound to Sense*, New York, USA, June 2014, pp. 163-168.

[9] P. Roach, *English Phonetics and Phonology: A Practical Course*, 4th ed.: Cambridge University Press, 2009, vol. 1.

[10] C. K. Hall, "Defining Phonological Rules over Lexical Neighbourhoods: Evidence from Canadian Raising," in *Proceeding of the 24th West Coast Conference on Formal Linguistics: Somerville MA: Cascadilla Proceedings Project*, USA, 2005, pp. 191-199.

[11] H. V. Hulst, "Rule Conversion in Phionology," *Dutch Lexicological Institute Leiden*, 1979, pp. 336-349.

[12] J. J. Ohala, "The Application of Phonological Universals in Speech Pathology," *Speech and Language: Advances in Basic Research and Practice* , vol. 3, no. 0-12-608603-6, pp. 75-94, 1980.

[13] G. Finch, "Phonetics and Phonology," in *Linguistic Terms and Concepts*, 1st ed., John Peck and Martin Coyle, Eds. New York: St. Martin's Press, INC. vol. 1, , 2000, ch. 3, pp. 33-76.

[14] Professor Trigo, LX 513 Phonology.

[15] H.V. Hulst and J. V. Weijer, Topics in Turkish Phonology.

[16] A. Waqar and S. Waqar, "Identification of Diphthongs in Urdu and their Acoustic Properties," , Lahore, Pakistan, 2002, pp. 16-26.

[17] M. C. Murcia, D. M. Brinton, and M. Janet, *Teaching Pronunciation Hardback with Audio CDs (2): A Course Book and Reference Guide*, 2nd ed.: Cambridge University Press, 2010.

[18] S. Nawaz, "Deletion Rules in Urdu Language," CRULP, Center for Research in Urdu Language Processing, Lahore,.

[19] Dr. R. Mendoza. Phonological Progresses.

[20] B. Akram, "Analysis of Urdu Syllabification Using Maximal Onset Principle and Sonority Sequence Principle," , 2002, pp. 160-166.

[21] W. Habib, R. Hijab, S. Hussain, and F. Adeeba, "Design of Speech Corpus for Open Domain Urdu Text to Speech System Using Greedy Algorithm," in *Conference on Language and Technology (CLT14)*, Karachi, Pakistan, 2014.

[22] B. Mumtaz et al., "Multitier Annotation of Urdu Speech Corpus," in *CLT14 - 5th Conference on Language and Technology*, Karachi, 2014.

[23] *Oxford Urdu-English Dictionary*, 1st ed, Oxford University Press, Karachi, Pakistan, 2013.

[24] D. Kahn, Syllable Based Generalizations in English Phonology, August 9, 1976.

[25] V. T. Genannt Nierfeld, "Restructuring," *Lingua*, vol. 33, , 1974, pp. 137-156.

[26] R. Bhatti, "Identification of Diphthongs in Urdu and their Properties," in *Conference for Language and Technology, CLT16*, Lahore, 2016.

[27] J. J. Ohala, "Coarticualtion and Phonology," *Language and Speech*, vol. 2, no. 36, pp. 155-170, 1993.

[28] J. Panevova and J. Hana. Intro to Linguistics - Phonology:. (2010, October). [Online] https://ufal.mff.cuni.cz/~hana/teaching/2013wi-ling/04-Phonology.pdf

Appendix

Short Vowel Alternation			
Words	**English**	**SP**	**AP**
بلند	high	bələnd̪	bʊlənd̪
محبت	love	məhəbbət̪	mʊhəbbət̪

Short to Medial and Medial and Medial to Short Vowel Alternation				
Words	**English**	**SP**	**AP**	
ارتکاب	committing an offence	ɪrt̪eka:b	ɪrt̪ɪka:b	ɪrt̪ək a:b
استعمال	use	ɪst̪ema:l	ɪst̪ɪma:l	ɪst̪əma:l
محمد	Proper noun	mʊhəmməd̪	mohəmməd̪	

Medial to Medial Vowel Alternation (e → æ)				
Words	**English words**	**SP**	**AP**	
			SS	**US**
احترام	respect	eht̪era:m	æht̪era:m	eht̪era:m
احتجاج	Protest	eht̪ɛʤa:ʤ	æht̪ɛʤa:ʤ	eht̪ɛʤa:ʤ
احتیاط	Care	eht̪ija:t̪	æht̪ija:t̪	eht̪ija:t̪
احرام	Unstitched white cloth for Hajj	ehra:m	æhra:m	ehra:m
احساس	Feeling	ehsa:s	æhsa:s	ehsa:s
احسان	Good deed	ehsa:n	æhsa:n	ehsa:n
ارکام	pillar	ehka:m	æhka:m	ehka:m
اختمام	phlebotomy	eht̪ema:m	æht̪ema:m	eht̪ema:m

Long to Long Vowel Alternation (e: → æ:)				
Words	**English words**	**SP**	**AP**	
			SS	**US**
اعجاز	miracle	eʔʤa:z	æ:ʤa:z	e:ʤa:z
اعتماد	Trust	eʔt̪ema:d	æ:t̪ema:d	e:t̪ema:d
اعتراض	objection	eʔt̪era:z	æ:t̪era:z	e:t̪era:z
اعلان	to announce	eʔla:n	æ:la:n	e:la:n
اعتبار	Trust	eʔt̪eba:r	æ:t̪eba:r	e:t̪eba:r

Vowel Deletion			
Words	**English Words**	**SP**	**Deletion**
اعتماد	Trust	e:t̪ema:d	e:t̪ ma:d
اعتراض	objection	e:t̪era:z	e:t̪ ra:z
آخرت	hereafter	a: xɪ rət̪	a:x rət̪
احتجاج	Protest	eht̪ɛʤa:ʤ	eht̪ ʤa:ʤ
امر	eternal	ə mər	əmr
جبل	mountain	ʤə bəl	ʤəbl
اعتراض	objection	e: t̪e ra:z	e:t̪ ra:z
آخرت	hereafter	a: xɪ rət̪	a:x rət̪

Short Vowel /ə/ Insertion before Liquid Sounds					
List of Words	**English Words**	**SP**	**Word final Consonant Cluster**	**Manners of Articulation**	**AP**
اصل	Original	əsl	Alveo-Fricative + Lateral	Any Consonant followed by Laterals /l/ or /r/ consonant; triggered	əsəl
غسل	Bath	xʊsl	Alveo-Fricative + Lateral		xʊsəl
مثل	Example	mɪsl	Alveo-Fricative + Lateral		mɪsəl
فضل	Bounty	fəzl	Alveo-Fricative + Lateral		fəzəl
عدل	justice	əd̪l	Dental + Lateral		əd̪əl
عقل	Wisdom	əql	Uvular + Lateral		əqəl
ذکر	account/talk	zɪkr	Velar + trill		zɪkər

عصر	Time Period	əsr	Alveo-Fricative + trill	schwa insertion	əsər
قبر	Grave	qəbr	bilabial + trill		qəbər
کفر	unbelief	kʊfr	Labiodental + Lateral		kʊfər
قدر	Value	qəɖr	Dental + trill		qəɖər
جبر	cruelty	ʤəbr	bilabial + trill		ʤəbər

Short Vowel /ə/ Insertion before /m/					
Words	**English Words**	**SP**	**Consonant Cluster**	**Articulation Manners**	**AP**
قسم	Kind	qɪsm	Alveo-Fricative+ Bilabial Nasal	Any Consonant followed by bilabial nasal /m/	qɪsəm
حکم	Order	hʊkm	Velar + Bilabial Nasal		hʊkəm
جرم	sin	ʤʊrm	liquid+ Bilabial Nasal		ʤʊrəm
کرم	fate	kərm	liquid+ Bilabial Nasal		kərəm
علم	education	ɪlm	liquid+ Bilabial Nasal		ɪləm

Short Vowel /ə/ Insertion before Alveo-fricative Consonants				
Words	**English Words**	**SP**	**Word final Consonant Cluster**	**AP**
اخذ	extract	əxz	velar + alveo-fricative	əxəz
حبس	congestion	həbs	Bilabial stop + alveo-fricative	həbəs
لفظ	word	ləfz	labiodental + alveo-fricative	ləfəz
قرض	loan	qərz	trill + alveo-fricative	qərəz
قبض	constipation	qəbz	bilabial stop + alveo-fricative	qəbəz

Morphology of Mewati Language.

Nadia Fareed.
University of Management and Technology Lahore
Email:Nadiafareed1111@gmail.com

Abstract

This paper focuses on the morphology of Mewati language. Mewati language is among the Indo- Aryan languages and is the mother tongue of Meo people. Meo people are the inhabitations of Mewat which is an ancient region in India. So, Mewati is the vernacular of Meo or Mayo people living in the Mewat. It is reported that after 1947 a lot Meos migrated towards Pakistan and settled in different areas here. So, aim of this paper is to examine the structures of Mewati language and different formats behind the construction of the novel words. In this regard a qualitative and descriptive study have been done in order investigate the different word formation process in Mewati language. Hocket's (1958) morphological models Item –and – Arrangement models or (IA) and Item- and – Process models or (IP) implemented in methodology for the analysis of the current data.

The data was collected from book, magazines, and newspapers on Mewati language. The data is arranged into nouns, cases, verbs, and adjectives. This paper documented properties of script and grammar of Mewati language along with brief history origin of Mewati language. The contrastive study of Mewati and urdu have been done in the proceeding section while keeping in mind the end results of the research. The data is represented systematically in tables with italic Mewati words and meanings in English. This demonstration will help the non- native readers in understanding of the Mewati grammar

1. Introduction

1.1. Meo, Mayo and Mewati (میواتی)

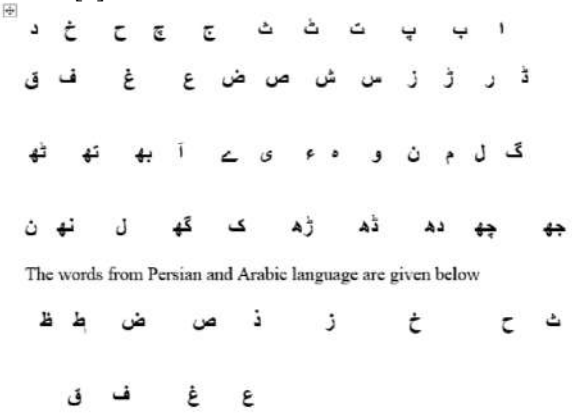
Meos are the inhabitants of Mewat which is an ancient region in India. The language of the Meo people living in Mewat is called Mewati.

So, Mewati is the mother tongue and vernacular of Meo or Mayo people. Meo people use Mewati in their informal setting context. It is reported that Mewati is the dialect of Rajasthani language and has no standard official script. Hence, literary data of Mewati language transferred orally from generation to generation [1].

It is stated that Meos migrated towards Pakistan from Mewat and settled in different cities in Pakistan e.g. Kasur, Sialkot, Multan, near Wagha boarder, Sialkot, and in Lahore. The reason behind the immigration of the Meos from the Mewat is based chiefly upon two combative Hindu movements, who forced the Muslim Meos to reconvert into Hinduism. In this case, refusal to them forced the Meos to leave Mewat [2]. Now, it is also evaluated that approximately 12 million people are living in various areas of Pakistan. Majority is living in Sindh, Khyber Pakhtun (KPK) and Baluchistan [3].

In addition to this, the word Meo is described as "*a singular masculine word which means a brave and illiterate nation of Mewat"* [4]. Mewati language is a morphologically rich language with 39 alphabets, 9 vowels, 31 consonants and 2 diphthongs. The supra segmental features are rare in this language. There are nine cases in this dialect: nominative, vocative, agentive, accusative, instrumental, locative, and genitive [5].

In Mewati language there are a lot of words are adopted from other languages of the world like Arabic, Persian, Urdu, and English. Those words are used in modified forms or in close to their original expression. For example, those words are niwaj, hajj, sadak, kitab, kalam etc. [6].



While taking into account various aspects of Mewati language, it is demonstrated that in Mewati language, that Mewati language is classified into five kinds:

.*khari Mewati* (میواتی کھڑی), 2). *Burj Mewati* (میواتی رج ب).3). *Rathi Mewati*
(میواتی ٹھی را), 4). *Nahera Mewati* (میواتی نہیڑا), 5). *Kathera Mewati* (کٹھیڑا میواتی). It is also stated that Mewati language is divided into two major types which *khari Mewati* (میواتی کھڑی) and *Pari Mewati* (میواتیپڑی) [7].

It is also depicted that for the understanding of accent and structure of any language it is very important to communicate with the native speaker of that language. The observation and practice of required language will reveal the exact word structure and its pronunciation [8].

Furthermore, it is perceived that the accent of Mewati language is quite rough. As, Mewati language is the dialect of Rajasthani language and there is also some differences found in both the languages like there is no presence of "L" (ل) sound found in Rajasthani language like Mewati language. In Mewati language both "L" and 'D' sound is replaced by ''R'' sound e.g. [9]

Table: 1.1

Rajasthani Language	Burj Basha	Mewati
Palol	Paror	Parol
Sadak (Road)	Sarak	Serak (Road)

Similarly, the syntactic arrangement of the words in Mewati language resembles with the Indo- Aryan languages. This dialect also follows the SOV formula for the description of complete sentence expression [Srivast]. For example

✓ *Beto iskul javey go*

"Son will go to school"

1.1.1. Aim of the study

To explore the strategies of word formation in Mewati language?

1.1.2. Research Questions

1. What are the strategies of word formation processes in Mewati language?
2. What is the procedure of gender and numbers formation in Mewati language?
3. What is the difference between Urdu and Mewati marking of case?

1.1.3. Purpose of the Study

The purpose of the study is to explore the morphological structure of Mewati language. The rules of word construction in Mewati language.

1.1.4. Scope of the Study

Morphology being the sub branch of linguistics which study the internal structure of the words. The objective of the current study is to explore the rules and arrangement of the words and its constituents.

So, the structure of nouns, verbs, adjectives and case markers are explored which revealed the distinct morphological characteristics of Mewati language. Mewati language being an indo Aryan language also exhibits similarity of scripts with Urdu and Hindi language. Hence, the study of patterns of word formation of Mewati language is important in its sense of first investigatory report of its morphological structure.

1.1.5. Delimitation of the Study

The existing study is narrowed to the morphology of the Mewati language which is one of the sub-discipline among the other disciplines of linguistics like phonology, syntax, semantics, and pragmatics etc. So, the focus of the research is on investigation of patterns of word formation in Mewati language.

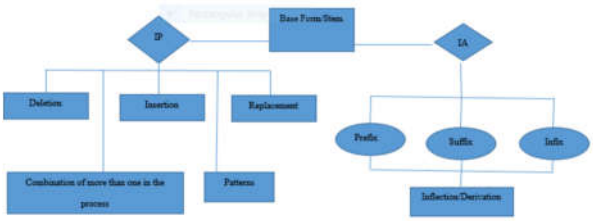
1.1.6. Data Collection Tools

The present study is descriptive and qualitative in nature. Books, magazines, newspapers are used for the documentations of data.

1.1.7. Theoretical Frame Work

Hocket (1958)'s morphological models are implemented in the current study. Hocket (1958)'s classified morphological models into item- and-arrangement (IA) and item – and – Process (IP) model.

Item – and – Arrangement (IA) model is named as morphemic and non- procession model. It is also entitled as affixes vs. morphemes model. In which, items are *morphs* or *morphemes* while arrangement symbolizes *sequence* of morphemes. So, IA is the exhibition of linear arrangement of morphemes. For example, the word chairs is a linear combination of two morphemes *chair* +/*s*/.



Item – and – process or (IP) model is called non-linear and processual model. In which items are titled as morphemes while process is the representation of rules. In this model morphemes can be added, deleted, converted and replaced by applying rules. For example, in the formulation of plurals of the word 'fans' the morphemes 'fan' attached with the morpheme /s/, /z/ [10].

2. Literature Review

The anatomical study of words and their structures is called morphology. This study base on the analysis of morphemes, affixation, reduplication and compounding. It is communicated that a variety of morphemes are existed as: prefix, stem, infix, and suffix. Even a single morphemes '*lailmi*' in Urdu word which means '*unawareness*' can be broken down into three morphemes. These three morphemes are prefix la '*un*' and Arabic stem '*ilm*' 'awareness' and Urdu suffix –i. the morphological analysis of this word reveals that the prefix la concerts the noun ilm into an adjective 'lailm' which is then changed into anther noun '*lailmi'* by the addition of the suffix –i[11].

While it is also states morphology deals with the morphemes and how they arrange to form variant words [12]. In addition to, it is illustrated that for studying and understanding of language, it is essential to investigate the "word formation processes'' and "rules" of constructing new words. The study of arrangement and fundamental procedures in the coinage of novel words which enhance the lexical vocabulary of the language. Hence, these rules and processes are the formatives which are productive in nature [13].

Morphology is sub-categorized into two kinds: inflectional and derivational. Many morphologists make a distinction between inflection and derivation. The difference between both the types depends upon the attachment of a morpheme to the stem or root of the words. As, inflection has capability of forming the form of the same word without changing thecategory of the word. For example, basta 'bag' baste 'bags' and basto(plural oblique). In this example, the word plate is singular noun, which changes into plural by the additions of two different plural markers '-e', and '–o'. This shows the oblique and vocative forms of the noun as well. Hence, inflected forms are variants of the same word [14].

Derivation on the other hand, is anti-parallel in processing, in which novel words are developed by altering the category of the base. For example, in Mewati language the derivation process is dhaan 'wealth' which is a noun and changes into an adjective by the addition of the suffix 'i' into dhaani ' wealthy/ rich person' which is an adjective in characteristics[15].

In addition to the categories of the morphology, it is also illuminated that he words are made up of small grammatical units called morphemes, which has an essential impact on the semantic and syntactic behavior of the words. Those morphemes are the smallest linguistics signs which are helpful in the study of the words. They are arranged in arbitrary sounds for establishing new words with different meanings. Morphemes are further divided into three types : root, stem and base, the root is defined as the core of the word which instantiates the various other forms of word e.g. *talk, talks, talking, talked* etc. in many language like English many roots can stand alone these types of roots are called free morphemes[16].

Although, the acceptance of any affix is called stem or when a root accepts an affix it is in the form of stem. For example, the word ''*foolishness''* is a combination of root, stem and suffix. It is worth mentioning here that the affixes are the bound morphemes which can attach to the root, stem and base as well e.g. *un-,dis-,-ful,-ness,-s,-es* etc. The morphemes are sub classified into three types, prefix, suffix, and infix

Prefix: when a bound morpheme attached at the staring of the root or stem e.g. *un-happy, in-decent, il-legal*, and *disapprove*. Here in these examples *un-, in-. il- ,dis-* prefixes. Suffix: in this case the morpheme attaches at the end of the root or stem e.g. *respect-full, honor-able, foolish-ness, help-less* etc. *–full, -able, -ness, -less* etc. is types of suffix.

Infix: when a morpheme is inserted in the mid of the root it is called infix. Arabic language is the best example of infix contrary to English which has no infix system. The word *kitab*
'book;' changes into '*ktb*' and '*intkb*'. [17]

3. Method

3.1. Research Design

The present research is qualitative in nature. So, it follows Hocket's morphological models as descriptive research design.

3.1.1. Data Analysis

Hocket (1958)'s morphological models: item- and – arrangement model or (IA) and item – and- process

model or (IP) are used in the analysis of data of the present study. These models provide the base for the construction of novel words through inflection and derivation.

3.1.2. Formation of Noun according to Number

To describe the name of any place, thing and object is called noun. In Mewati language the construction of noun is divided into two categories e.g. formation of noun according to numbers and genders.

3.1.3. Numbers (Singulars and plurals)

The formation of numbers (formation of plurals from singulars) follow four different rules in Mewati language which are given below

Case no. 1

1. In this the sound /o/ at the end of the words show singular slot. Which is made plural by the replacement of vowel */o/* at the end of the stem by vowle */a/* e.g. beto *(son)* ~ beta (*sons*), darwajo (*door*)~ darwaja (*doors*), bhanjo (*nephew*)~bhanja (*nephews*).

Table: 3.1

Mewati Gloss (sg.)	**Meaning**	**Mewati Gloss(pl.)**	**Meanings**
Chacho(sg)	Uncle	*Chacha (pl.)*	Uncles
Beto	Son	*Beta*	Sons
Poato	grandson	*Poata*	grandsons
Salo	Brother in law	*Sala*	Brothers in law
Choro	Boy	*Chora*	Boys

Case no.2

In this case plurals are made by the addition of consonant /n/ at the end of those singulars which sounds as /i/ e.g. chori (*girl*)~ chorin (*girls*), ghoari (*mare*)~ ghoarin *(mares*), chaabi *(house*)~ chaabin (*houses*).

Table: 3.2

Mewati Gloss (sg.)	**Meaning**	**Mewati Gloss (pl.)**	**Meaning**
Gali	Street	*Galin*	Streets
Ghari	Watch	*Gharin*	Watches
Murgi	Hen	*Murgin*	Hens
Tokri	Basket	*Tokrin*	Baskets
Chaabi	Key	*Chaabin*	Keys

Case no.3

In this case plurals are developed from the singulars by the addition of */an/* at the end of the singular words e.g. mulk (*country*) ~ mulkan (*countries)*, khaet (*field*) ~ khaetan *(fields*), baag (*garden*) ~ baagan (*gardens*).

Table: 3.3

Mewati Gloss (sg.)	**Meaning**	**Mewati Gloss(pl.)**	**Meanings**
Baat	Thing	*Baatan*	Things
Raatt	Night	*Phoolan*	Nights
Mulk	Country	*Haaran*	Countries
Phool	Flower	*Phoolan*	Flowers
Haar	Necklace	*Haaran*	Necklaces

3.1.4. Gender formation.

In Mewati language the construction of genders (masculine and feminine) consists of three cases as well.

Case no.1

In the first case the masculine ends in vowel sound /o/. This is replaced by /i/ for the formation of feminine e.g. choro (*boy*) ~chori (*girl*), goaro (*horse*) ~goari (*mare*).

Table: 3.4

Mewati Gloss (masc.)	**Meaning**	**Mewati Gloss (femi.)**	**Meaning**
Choro	Boy	*Chori*	girl
Syano	Wise male	*Syani*	Wise female

Nano	Grandfather	*Nani*	Grandmother

Case no. 2

In this case feminizes are formed by replacing /a/ of masculine by /i/ e.g. phoopa (*uncle*)~phoopi (*aunt*), kaka (*baby boy*)~kaki (baby *girl*) etc.

Table: 3.5

Mewati Gloss (masc.)	**Meaning**	**Mewati Gloss (fem.)**	**Meaning**
phoopa	Uncle	*phoopi*	Aunt
Ghoara	Horse	*Ghoari*	Mare
Beta	Son	*Beti*	Daughter

Case no. 3

Addition of */ni/* at the end of words *e.g.* mor *(peacock)~morni (peacock hen).*

Table.3.6

Mewati Gloss(masc.)	**Meaning**	**Mewati Gloss (fem.)**	**Meaning**
Ustaj	Teacher (male)	*Ustani*	Teacher (female)
Dewar	Brother in law	*Dewrani*	Wife of brother in law
Dactar	Doctor (male)	*Dactarni*	Doctor (female)

Cases or Case Markers in Mewati Language

It is stated that case is "*system of marking dependent nouns for the type of relationship they bear to their heads*" [18]

In Mewati language three cases experienced which are nominative, vocative and agentive case. Other cases are foundbut in this study we will discuss the above three.

3.1.5. Nominative case

The nominative case of the Mewati language is unmarked

✓ **Chori ayee**

Chori (noun), ayee (came, past participle)

"*The girl came''*

✓ **Balkan roato**

Balkan (children, noun pl.), raoto (Weep, present participle)

"*The Children weep''*

3.1.6. Agentive case

This case is marked by /–ne/. This case shows the subject and object agreement in the sentence e.g.

✓ ***Bhai ne roat khai***

Bhai (N), roat (O), khai (Verb, past participle), ne (agn. case)

"Brother ate bread''

✓ ***Tum ne dud piyo***

Tum (N), ne (agn. case), dud (O), piyo (verb, past participle)

"You drank milk''

3.1.7. Vocative case

In this case the use of /re/ or /ore/ is marked for addressing

✓ ***re rabb mu maryo***

Re-rabb (expression), mu (N), maryo (dead, verb, past participle)

"Oh God! I am dead''

✓ ***ore amma kithe soano***

ore(listen),amma(mother,N),kithe(where,H.v),soano(sle e ping,present progressive)

"Listen! Where is mother sleeping''.

3.1.8. Accusative case

In Mewati language accusative case is not morphologically marked. It occurs with both the cases without any marked condition. For example,

✓ ***ai -nE bula*** ai (mother, noun sg.), -nE(accus.case), bula(verb)

"Call the mother"

✓ ***Chacho –nE bhejo***

Chacho (uncle, noun sg.), -nE(accus. Case), bhejo (verb)

"Sent to uncle".

3.1.9. Genitive case

It is said in Mewati language the genitive case is marked for gender numbers like in Urdu language. In

this case the case markers are /-k/, /-r/ are named accusative case markers. For example,

✓ ***Raja ko ghoaro***

Raja (noun), ko (agentive, case), ghoaro(object)
"Raja's horse"

✓ ***Lugai ki chaabin***

Lugai (wife, noun), ki (agentive case), chaabin (keys, object)
"Wife's keys"

3.1.10. Locative case

In Mewati language locative case markers are marked as –mẽ , -mã̃ , -par. These locative markers of Mewati language are like Urdu. For example, in Urdu the locative case markers are –mai, and -par for indicating location.

Example:

✓ ***bhai khaat par leto ho*** bhai (brother, noun), khaat(cot, object), par
(Locative case), leto (
verb), ho(preposition)
"Brother was sleeping on the cot"

✓ ***rah mã̃ nadi***

rah (way, noun), mã̃ (locative case), nadi(river, object)
"River on the way"

3.1.11. Instrumental case

In Mewati language instrumental case is marked by saĩ̃, -se, and -seĩ̃. For example,

✓ ***chamcho se kha***

chamcho (spoon, noun), se(instr. case), kha (verb)
"Eat with spoon"

Verbs

Verb defines the state of action, being and state of being. In Mewati language the verbs are constructed on different conditions. Two types of verbs on the basis of presence of object is perceived which is called transitive (presence of object) and intransitive (without object).

The intransitive verb converted into transitive by following the rules given below in the cases

Case no. 1

The insertion of morpheme */o/* at the end of the stem e.g. bas (*live*)~ baso (*make one to live*), sun (*listen*)~ suno (*make one to listen*) etc.

Table: 3.7

Transitive	Meaning	Intransitive	Meaning
Likh	Write	*Likho*	Make to write
Chup	Hide	*Chupo*	Make to hide
Ud	Fly	*Udo*	Make to fly

Case no.2

The insertion of morpheme /a/ at the end of the stem e.g. bas (*live*)~ basa (*make one to live*), sun (*listen*)~ suna (*make one to listen*).

Table: 3.8

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Pooch	Ask	*Poocha*	Make someone to ask
Bol	Speak	*Bola*	Make someone to speak
Khol	Open	*Khola*	Make someone to open
Pis	Grind	*Pisa*	Make someone to grind
Khod	Dig	*Khoda*	Make someone to dig

Case no. 3

The addition of cluster of vowels */ayo/* by deleting sound */o/* e.g. baso (*live*)~ basayo (help someone in living), suno (*listen*)~ sunayo (*make someone to live*) etc.

Table: 3.9

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Likh	write	*likhayo*	Make someone to write
Dekh	see	*dekhayo*	Make someone to see
pak	Cook	*pakayo*	Make someone to cook

Case no. 4

Attachment of morpheme /+ao/ at end of the stem of root e.g. tehal(walk)~ tehalao (help in movement), sun(listen)~ sunao(help in listening)

Table: 3.10

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Kat	Cut	*katao*	Help in cutting
Rakh	keep	*rakhao*	Help in keeping
Char	climb	*charao*	Help in climbing

Case no. 5

The addition of morpheme /+to/ at the end position of the stem in order to make the progressive verb e.g. parh (study)~parhto (studying), ja (go)~jato (going) etc.

Table: 3.11

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Likh	Wrtite	*likto*	Writing
Kha	Eat	*khato*	Eating
Aa	Come	*aato*	Coming

Case no. 6

Addition the insertion of / +wayo/ while deleting the last /o/ sound at the end of the stem e.g. poocho (ask) poochwayo (make someone to ask), daro(fear) darwayo(make someone to fear) etc.

Table: 3. 12

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Poochto	asking	*Poochwayo*	Make someone to ask
Bolto	speaking	*Bolwayo*	Make someone to speak
Kholto	Open	*Kholwayo*	Make someone to open
Pisto	Grinding	*Piswayo*	Make someone to grind
Khodto	Digging	*Khodwayo*	Make someone to dig

3.2. Adjectives

Adjective is the expression of quality of any person and thing. Which makes it different form the others. In Mewati language the construction of adjective is by two different ways one is from the noun base and second is from verb base. For example, **Case no .1**
The addition of morpheme */+i/* convert a noun into adjective e.g. *kaam~kaami worker), phoaj~phoaji.*

Table: 3.13

Mewati gloss	Meaning	Mewati Gloss	Meaning
Niwaj	prayer	*Niwaji*	Piou man
Khael	Game	*khaelari*	Player
Des	Country	*Desi*	Indigenous

Case no. 2

The insertion of morpheme */+to/* at the end of the verb base form the adjective e.g. mar~ *marto naag, naach ~naachto choro* etc.

Table: 3.14

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Khaa	eat	*Khato Balkan*	Eating children
Jaa	Go	*Jato log*	Going people
So	sleep	*Soato chacho*	Sleeping uncle

Case no. 3

The addition of the morpheme */+o/* at the end of the verb stem also form the adjectives e.g. *chakh~ chakho aam, toot~tooto darwajo* etc.

Table: 3.15

Mewati Gloss	Meaning	Mewati Gloss	Meaning
Beth	Sit	*Betho kaag*	Sitting crow
Khol	open	*Kholo bari*	Opening window
Ur	Fly	*Urto baaj*	Flying eagle

Degree of Adjectives.

In Mewati language comparative and superlative degree of adjective is found in which use of */su/* and */ sab su/* mentions the moderation of the quality.

1. The use of */su/* in the sentence marks the comparative degree of adjective e.

✓ ***Balkan su piyaro kon lagy he.***

"Who wll be dearest than children?"

2. The presence of */sab su/* is the representation of the superlative degree of the adjectives e.g

.

✓ ***oow chori sab su piyari he.***

"That girl is more beautiful than all"

4. Conclusion

As discussed earlier that morphology is the investigation of core structure of words. In Mewati language inflectional and derivation processes are found in the establishment of novel words like Urdu language. For example, inflection and derivation processes in Mewati language follow the rules and sequences steps like item-and process model and item-and – arrangement model. The inflection follows the IA model in which morphemes arranged sequentially in the formation of the new words e.g. bag~ bags, chair~ chairs etc.

While, IP model exhibits the features of derivational morphology in which rules are followed for the making of new words. For example, in Mewati language naam (*name*) a noun changes into naami (*famous*) adjectives like Urdu. Expression of case marking in Mewati language is also like the Urdu language in which nine cases are found.

Nominative case is unmarked in both the language and rest of the cases show agreement of nouns with the verbs and vice versa.

The adjectives and verbs follow various rules for the development of new verbs and adjectives. Such case also found in Urdu in which a change in morpheme form verb with different meaning. Hence, it is concluded that in Mewati language inflection derivation processes play significant role in the construction of novel words. In fect, infixation is absent in this language which is among the major word formation processes in Arabic language.

5. Discussion and Recommendation

The present study revealed the morphological structure of Mewati language which has no official script, which is among the eight dialects of Rajasthani language.

Mewati language is not restricted to only Mewat, it is spoken as vernacular wherever the Meo people reside. It shares many commonalities and differences with the Urdu and Hindi language as well. Although, the accent is quite rough than both the languages. It is also recommended that being a first research on the grammatical structure of this language other linguistic aspects are present for computational analysis e.g. phonology, syntax, semantics etc.

6. REFERENCES

[1], [7],[9] Iqbal, J. *Sikandar Sohrab Batoor Shair Tehki-ko-Tankeedi Mutala*. (Unpublished MPhil's thesis).Minhaj university. (2014)

[2]. Census of India. Mewati: Retrieved from http://www.censusindia.gov.in/2011-documents/lsi/lsi_Rajasthan/8_Mewati.pdf.2011, accessed in June, 2, 2016.

[3]. Bakshi,P.)*Mewat:Comparative Case Studies Mewati in two School Types*.(unpublished masters 'thesis).University of Sydney.(2013)

[4]. Nayyer, N.H...*Noor-ul-Lugat.* (4th Ed.).Islamabad. National Book Foundation. (p.744). (1985)

[5]. Meo, Q.C.*Tamadne Mewat*. Lahore:Fiction House Publishers.(p.510)(2012)

[6].Fareed, N. (2016). Morphemic structure of Lahori Mewati. (Unpublished MPhil's thesis). University of Management and technology Lahore.

[7], [8]. Khan, A.H.*Mewati Adab*.Molana Azad Mewat Academy, New Delhi, (p.17). (2011)

[10]. Maxwell, M. (1998, feb). Two theories of morphology,one implementation. SIL Electronicworking papers 1998-001.
Retrieved June 2, 2016, from http://www.doc88.com/p-1324509071887.html

[11], [14]. Islam, A.R. (2011). *The Morphology of Loanwords in Urdu: the Persian, Urdu and English Strands*. (PhD's published thesis).Newcastle University.

[12]. Sayal, P. & Jindal, D.V..*An Introduction to Linguistics*. New Delhi: Prentice hall of India. (2002)

[13]. Aronoff, M. & Fudeman, K. (2nd Ed.).*What is Morphology*? Blackwell Publishers Ltd. (p.1). (2011)

[15]. Srivastava. S.P. (2011). Mewati_ Census of India. Retrieved from *www.**censusindia**.gov.in/2011documents/lsi/lsi.../8_**Mewati**.pdf*

[16]. Sibarani, R. *Introduction to Morphology*. Medan: USN. (p.4). (2002)

[17]. Katamba, F. *Modern Linguistics – Morphology*. London: Macmillan. (Pp.61-65). (1994)

[18]. Blake, Barry. 2001. Case. Cambridge: Cambridge University Press. Second Edition.

Identification of Diphthongs in Urdu and the Acoustic Properties

Rashida Bhatti and Benazir Mumtaz
Center for Language Engineering,
Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore

Abstract

The present research is carried out to finalize the list of diphthongs for the development of Urdu Phonetic Inventory. The corpus is specifically designed in carrier sentences to attest the existence of diphthongs in Urdu. For the identification of diphthongs in Urdu, two approaches are used in this paper i.e. perceptual approach and acoustic approach respectively. In perceptual approach, diphthongs are identified by ten native speakers using syllable identification technique. Diphthongs which passed the perceptual test were sent forward for acoustic testing. In acoustic approach, speech of six native speakers is analyzed using durational and formant cues both at stressed and unstressed forms on PRAAT. The combined analysis of perceptual and acoustic approaches indicates that Urdu has fifteen diphthongs.

1. Introduction

Urdu language has total 67 sounds in its phonetic inventory, i.e. 36 consonants, 15 aspirate consonants, 7 long vowels, 3 short vowels and 3 medial (majhul) vowels [1]. There are also nasalized forms of 7 long and 3 short vowels [1]. Having a large count of vowels in Urdu language is the triggering point to identify the diphthongs in Urdu and define a final list in phonetic inventory of Urdu.

To date, four studies have been carried out to identify the diphthongs in Urdu language and in results there are four different lists of diphthongs. Moreover, there are also 2 diphthongs which are claimed in present research after analyzing the ten hours of speech developed for Urdu TTS [1]. Due to the lack of consistency among previous researches, this study is carried out.

This work is based on a perceptual and an acoustic analysis of Urdu diphthongs hence a defined list of diphthongs can be added in Urdu phonetic inventory.

The results of this study are very important as identification of diphthongs is very essential for smooth annotation process of Urdu speech corpus hence the development of speech database of Urdu language. Moreover, identification of diphthongs will help to develop a robust pronunciation lexicon of Urdu language. Annotated speech corpus and pronunciation lexicon play important role in the development of Natural Language Processing (NLP) tools, i.e. Text to Speech Synthesizer (TTS) system, Automatic Speech Recognition (ASR) system [2] and Screen Readers. By adding these diphthongs in Urdu Phonetic Inventory, quality of the automatic speech would be improved. It would be more natural, audible and pleasant to hear.

The paper is organized in the following sections. The previous researches on the identification of diphthongs in the languages of world, Indo Aryan languages and specifically Urdu are presented in section 2. The methodology to study Urdu diphthongs is detailed in section 3. Data analysis description is presented in section 4, results and conclusions are discussed in section 5, while future work and recommendations are presented in section 6.

2. Literature review

A diphthong may be defined as a sequence of two perceptually different vowel sounds within one and the same syllable [3] or as a single vowel with continuously changing qualities [4]. There is no evidence of phonemic diphthongs in standard Urdu language [5] but present study shows that phonetic diphthongs are observed in the speech of native Urdu speakers.

The reported researches in other languages have used different approaches to investigate the diphthongs, i.e. phonological marking of diphthongs through minimal pairs [6], perceptual identification of diphthongs using syllabification [7] and acoustic study of diphthongs using spectrum (formant) and duration analysis [6][8]. Acoustically, diphthong has some important information at three points, i.e. starting vowel, transition period and final vowel [9]. Okati,

Helgason & Jahani [10] have used these points for acoustic analysis for the identification of diphthongs; they have analyzed formant values and duration of these three components. In another study on the learners of South African English, pitch contour is used for acoustic analysis along with formant analysis [11].

Urdu is a member of Indo Aryan Languages [12]. Existence of diphthongs in the phonetic inventories of these languages is reported in different researches but those are less in numbers. In Bengali, there are 21 diphthongs [8]; two in Sanskrit while no diphthong in Sinhalese [13]. Six diphthongs are reported in Kashmiri and Pahari language [6]. Punjabi Language has 6 diphthongs which are comprised of short vowel and long vowels [2]. Such possibilities are also seen in Urdu. Samare [14] posits six diphthongs for the Persian language varieties spoken in Iran, but points out that they are only diphthongs from a phonetic point of view and can also be described as sequences of vowel and glide. Ganjavi et al. [15], Yaesoubi [16] and Hakimi [17] have reported diphthongs in different dialects of Persian, but many of those are either phonetic diphthongs or the combination of vowel and glides (i.e. /j/ and /w/). As Persian has much influence on Urdu, here is the possibility that Urdu language may have this type of diphthong combinations.

To date, four researches have been carried out for the identification of diphthongs in Urdu [7], [18], [19] and by CLE (Center for Language Engineering) researchers (available on: (www.cle.org.pk). These studies present 4 lists and every list has different number of diphthongs. Almost, all the researches have used syllabification technique to identify the diphthongs and then acoustic cues are used to describe the characteristics of Urdu diphthongs.

Waqar and Waqar [7] have proposed total 13 diphthongs, i.e. /ɪũ:/, /əe:/, /əi:/, /ɑ:o:/, /ɑ:i:/, /ɑ:e:/, /e:a:/, /o:i:/, /ɑ:ẽ:/, /o:e:/, /əĩ:/, /ɪa:/ and /a:ũ:/. Obtained results show that existence of diphthongs is speaker dependent and diphthong combination is in the result of deletion of phonemes, i.e. "ʔ", "j" and "v". Moreover, their analysis of duration shows that diphthongs are combination of one short and one long vowel and the duration of diphthongs is below 300 while duration of consecutive two vowels is round about 350ms. Phonemically the existence of diphthongs (using minimal pairs) cannot be proved, since they are formed as a result of deletion of either the consonant, or the timing slot.

Khurshid, Usman and Butt [18] have finalized 18 diphthongs in Urdu using syllabification and acoustic analysis. Their list includes /oi/, /oe/, /ɪo/, /əi:/, /əe:/, /ua:/, /uə/,/a:ɪ/,/ao/, aũ:/, /ɪũ:/, /io/, /ea/, /eo/, /ʊa/, /ui/ and /ue/ diphthongs. Another research conducted by Sarwar, Ahmad and Tarar [19] depicts that Urdu has 17 diphthongs, i.e. /ai/, /ae/, /ao/, /ɪu˜/, /ɪa/, /au˜/, /oi/, /oe/, /oi˜/, /əi/, /aẽ/, /ea/, /əi˜/, /ʊa/, /ui/, /ue/ and /əe/. The obtained results show that diphthongs have had three parts, i.e. on glide, off glide and transition period. CLE has finalized a list of 7 diphthongs. This list reports unique combination of diphthong having medial (majhul) and long vowel qualities. The common thing in all these studies is that diphthongs are identified perceptually using native speaker intuition. Moreover, previous researches report that identification of syllabification and diphthong is speaker dependent. It is also noticed that identified diphthongs are mostly consisted of one short and one long or combination of two long vowels. [7] and [18] previous work claims that a diphthong will be considered a diphthong only, if 50% votes are in favor of a particular diphthong whereas [19] prefers 60% votes for the selection of diphthong.

There are two new diphthongs, which are not mentioned in the previous researches, i.e. /ea:/ and /a:e/. These diphthongs are studied during the annotation of 10 hours Urdu speech corpus; /ea:/ diphthong in words like اعادہ /ea:d̪a:/ repeat and پیار /pea:r/ love etc and /a:e/ diphthong in words like جائزہ /ʤa:eza:/ overview, کائنات /ka:ena:t̪/ universe etc.

Hence, due to lack of inconsistency among lists there is no specified list of diphthongs which may add in phonetic inventory of Urdu language. Thus, the current study is built on the previous research efforts to develop a unified list of diphthongs.

The following section presents the methodology followed during the research.

3. Research Methodology

This study is carried out by combining the four previous lists of diphthongs along with 2 possible diphthongs proposed in present research. In combined list, total 26 possible diphthongs are selected for study.

3.1. Corpus development and speech recordings

In order to study the proposed phenomenon, corpus was specifically designed for recording. Seventy eight (78) words containing diphthongs were selected and embedded in carrier phrase. For example:

1) میں نے آئینی کہا
/mẽ: ne: ɑ:ɪni: kəhɑ:/
I said constitutional.

It was made sure that there was a valid coverage of all 26 possible Urdu diphthongs in the corpus (3 words for each possible diphthong (3*26=78). Recordings

were obtained from six native speakers of Urdu (3 males and 3 females) in an anechoic chamber who also use Punjabi in their daily routine. PRAAT software was used for recordings and analysis. The speech samples were recorded in mono form at the sampling rate of 48 KHz, and stored in wav file format.

Six speakers were asked to read out the sentences in their natural style of speaking. Three instances of each speaker's voice samples were recorded (3*3*26=234) and stored in wav file format for subsequent offline processing. Afterwards segments were marked on phoneme, syllable, word and stress levels using process described in [1]. Moreover, ten sentences containing distracters were also recorded from the speakers. Distracters were the words imbedded in carrier phrases to attest the respondents' perception regarding syllabification, i.e. عادلانہ /ʔɑːd̪ɪlɑːnɑː/ uprightly, اسطوخودُوس /ʊsʈuːxʊɖɖuːs/ *lavender,* اِسلامیات /ɪslaːmɪjaːt̪/ *Islamic studies,* اِشارت /ɪʃaːrət̪/ *insinuation,* بلاغت /bəlaːɣət/ *eloquence,* دلیرانہ /d̪əleːraːnaː/ *courageous,* رُویت /ruːjət/ *visibility,* رؤف /rəuːf/ *rauf,* ہوائی /həvaːiː/ *hawaii,* لاجورد /laːʤvərd̪/ *Armenian stone and* لاعلم /laːʔɪlm *unaware.*

3.2. Perceptual Experiment Methodology

To verify the defined list of diphthongs, first of all, recorded sentences containing 26 possible diphthongs were listened and segmented using PRAAT. Mispronounced or having bad quality voice were not selected for perceptual analysis. Three utterances of each diphthong were selected at word initial, middle and final positions from the speech of six speakers for the perceptual experimentation (26*3*6=468). However, there are few diphthongs such as ɑːe, ɪãː, ɪõː and əĩː which do not exist at word initial and middle position in Urdu. For these diphthongs, three instances are taken only at word final position. Thus, there were three waves of one diphthong, which consisted of utterances of six speakers.

Moreover, to evaluate the native speaker perceptual understanding of diphthongs, utterances of six speakers containing one diphthong and two sentences containing distracters were combined in one wave file.

Later on, 10 native speakers (5 males and 5 females whose age vary from 20 to 30) were asked to listen to these 78 wave files one by one using headphones to identify the diphthongs in Urdu. Syllable count is a good cue to identify the diphthongs; therefore, respondents were asked to count the syllables in recorded words. Respondents listened to all three files against each diphthong and wrote the syllable count in a given questionnaire. On the basis of their syllable count log sheet, 16 diphthongs are finalized by the respondents. Among these 16, 5 are the nasalized diphthongs (See Section 4 Table 1).

3.3. Acoustic Experiment Methodology

To verify the proposed list of 16 diphthongs, durations of finalized diphthongs and formant frequencies are analyzed manually. Duration of diphthongs in both stressed and unstressed forms are calculated separately (3 unstressed+3 stressed*16 diphthongs*6 speakers=576). Average values of males and females are calculated and enlisted in Appendix A. Moreover, minimum duration in unstressed form and maximum value at stressed form is also mentioned in Appendix A. Only one perceptually selected diphthong /ɑːĩː/ is rejected at this stage of experimentation.

Formant frequencies of finalized 15 diphthongs are measured from the recorded speech. To measure the formant frequency of first (F1), second (F2) and third formant (F3), diphthongs were divided into three components, i.e. on glide (1), transition (2) and off glide (3). F1, F2 and F3 are measured manually from the middle of component 1 and 3. Window of PRAAT was assured to be 20 ms to take formant values of each component. Three instances of every diphthong from the recorded speech of six speakers (3*15*6=270) are considered for formant values. Average formant frequencies of finalized diphthongs are reported in Appendix B.

4. Results

Selection of diphthongs was done on the basis of frequency of the responses. In this research, a diphthong is considered a diphthong only, if 70% votes are in favor of a particular diphthong. Only one diphthong /ɑːĩː/ having 60% votes was selected for further testing and highlighted in green color in table 1. The list of perceptually selected diphthongs is given in Table 1.

Table 1: Results of perceptual analysis

Sr No.	Diphthongs	Perceptual Agreed Diphthongs	Perceptual Disagreed Diphthongs
1	ɑːeː	90 %	10 %
2	ɑːe	80 %	20 %
3	ɑːẽː	70 %	30 %
4	ɑːɪ	90 %	10 %
5	ɑːiː	80 %	20 %
6	ɑːĩː	60 %	40 %
7	ɑːoː	80 %	20 %
8	ɑːuː	40 %	60 %
9	əe	100 %	0 %

10	æɑ:	90 %	10 %
11	əi	100 %	0 %
12	əĩ:	80 %	20 %
13	eɑ:[6]	70 %	30 %
14	e:o:	20 %	80 %
15	ɪɑ:	10 %	90 %
16	ɪã:	30 %	70 %
17	ɪo:	20 %	80 %
18	ɪõ:	80 %	20 %
19	ɪu:	0 %	100 %
20	ɪũ:	90 %	10 %
21	o:e:	50 %	50 %
22	o:i:	80 %	20 %
23	o:ĩ:	50 %	50 %
24	uɑ:	30 %	70 %
25	ue:	20 %	80 %
26	ui:	80 %	20 %

☐Grey highlighted are diphthongs which are rejected in perceptual analysis.

5. Discussion and Data Analysis

The obtained results from perceptual analysis show that there are 16 diphthongs in Urdu which are identified by ten native speakers of Urdu.

Respondents rejected ten proposed diphthongs out of 26, i.e. /ɪã:/, /u:e:/, /a:u:/, /e:o:/, /o:e:/, /o:ĩ:/, /ɪɑ:/, /ɪo:/, /ɪu:/and /ʊɑ:/ in perceptual testing. Sixty percent respondents accepted the diphthong /ɑ:ĩ:/ but we have had delimited the votes in favor to 70 %. However, we accepted this diphthong for further experimentation. Four speakers out of six have pronounced diphthong /iɑ:/, differently. Either the speakers pronounced it with 'j' or without 'j'. Respondents did not recognize it as diphthong (See Figure 1: J sound in /lijɑ:/). They counted it bi syllabic word. Similarly /ʊɑ:/ and /u:e:/ are also not pronounced as diphthongs by speakers. Only one speaker pronounced it as diphthong and only one listener recognized it as diphthong, but the rate of speech was comparatively speedy during this particular utterance.

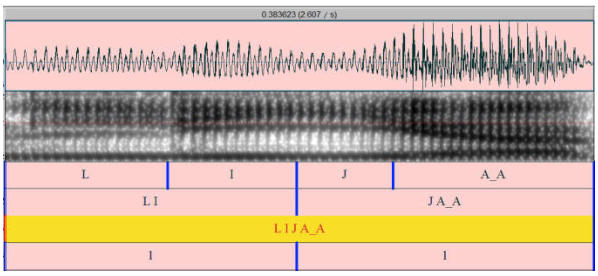


Figure 1: J sound in /lijɑ:/

Two speakers have pronounced /v/ in second syllable ہوئے /huve:/ happened. Either those are spoken with /v/ or without /v/; native speakers did not recognize those as diphthongs (See **Figure 2: V sound in /huv**/). /ɪu:/ is not accepted by the respondents as diphthong, although it is a controversially accepted diphthong in American English but Urdu speakers did not speak it as diphthong.

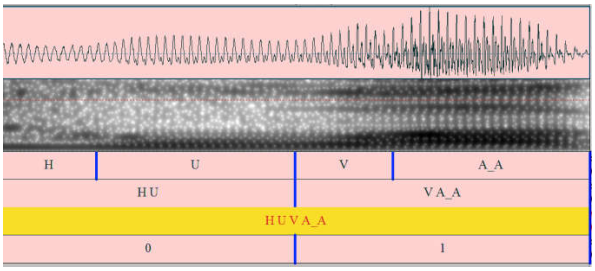


Figure 2: V sound in /huvɑ:/

Three speakers pronounced /jã:/ instead of proposed diphthong /ɪã:/ (As in ناکامیاں /nakamɪã:/ failures, لڑکیاں /laɽkɪã:/ girls and کھڑکیاں /khiɽkɪã:/ windows) while the other three pronounced it as diphthong. However, the 70% respondents did not recognize it as diphthong. Five speakers out of six could not speak the proposed diphthong /o:ĩ:/ rather they pronounced it as /o:i:/.

Data analysis of selected diphthongs suggests that diphthongs in Urdu can occur in 5 types of combinations.

- short and long oral vowels
  - /ɑ:ɪ/, /əe:/ and /əi:/
- long and long oral vowels
  - /ɑ:e:/, /ɑ:i:/, /ɑ:o:/, /o:i:/and /u:i:/
- medial and long oral vowels
  - /ɑ:e/, /æɑ:/ and /eɑ:/
- Short and long nasalized vowels
  - /əĩ:/,/ɪõ:/ and /ɪũ:/
- Long and long nasalized vowels
  - /ɑ:ẽ:/

[6] Red colored are proposed diphthongs in present research

Combination of long-long vowel and long-medial vowel is unique property of Urdu language. /æɑː/ is considered as diphthong, 5 speakers out of six pronounced it as diphthong while all respondents recognized it as one syllable hence diphthong.

Waqar and Waqar [7] had the conclusion that diphthong is made in the result of the deletion of any phoneme but the formant analysis shows that Urdu speaker alternate the schwa and J with medial vowel /æ/. The formants of medial vowel /æ/ in /æɑː/ diphthong and individually has almost similar values (See *Appendix B)*.

In the case of diphthong /eɑː/ (e.g. in زیادہ /zeɑːɖɑː/ excessive, تیاری /ʈeɑːriː/ preparation and فلکیات /falkeɑːʈ/ universe), Urdu speakers alternate the sound /ɪ/ and /j/ with medial /e/. This medial vowel blends with the following vowel /ɑː/ and makes the diphthong /eɑː/ as shown in Figure 3. All the speakers pronounced it as diphthong and 70% respondents recognized it as diphthong. Rest 30% respondents had the confusion to identify it as diphthong in word /falkeɑːʈ/. They counted it as tri syllabic word.

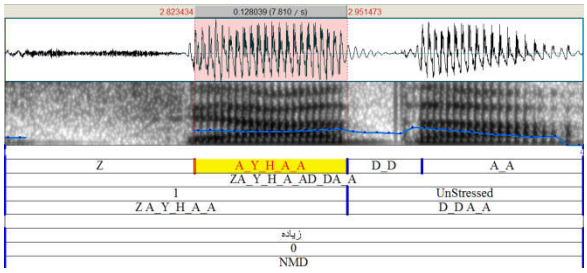


Figure 3: e sound in Diphthong /eɑː/

Duration analysis shows that diphthongs show the qualities as an entity like long vowels. The duration of diphthongs increase in the state of stressed syllable like long vowels.

On average the maximum durations of unstressed diphthongs is. 148 ms (See Appendix A). Therefore, on the basis of durations, /ɑːĩː/ diphthong was rejected at acoustic experiment stage. Obtained results show that there is no significance difference in average duration values of diphthongs on the bases of gender. Almost the duration is similar in the speech of males and females. During annotation and perceptual analysis of diphthongs, it is also observed that in diphthong both vowels blends in such a way that listeners cannot separate them as shown in Figure 4.

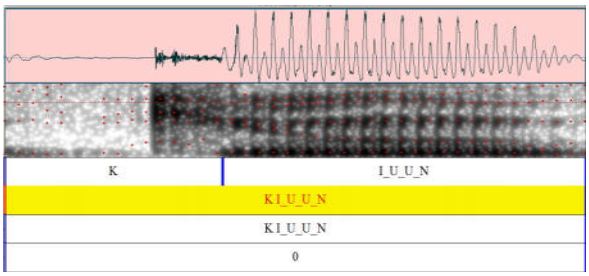


Figure 4: Diphthong I_U_U_N

Formant analysis of finalized 15 diphthongs shows that F1, F2 and F3 of vowel components in different diphthong combinations do not have much difference in values. One vowel shows almost similar values in different combinations of diphthongs. For example /ɑː/ has not much difference in different combinations of diphthongs like in /ɑːeː/, /ɑːe/, /ɑːẽː/, /ɑːiː/, /ɑːɪ/ and /ɑːoː/ (See Appendix B). Obtained results show that although vowel maintains their qualities and formant frequency but they blend with other component to become a diphthong. Moreover, analysis of Urdu diphthong highlights that similar to other languages, Urdu diphthongs have three components, i.e. first vowel, transition period and second vowel.

6. Future work and recommendations

Finalized list of diphthongs is presented in this research using perceptual and acoustic approaches. The selected diphthongs can be added in Urdu phonetic inventory. It will be helpful in maintaining the accuracy and consistency during the annotation of speech corpus. By marking diphthongs, syllables and stress tier annotation can also be done more smoothly and accurately. Moreover, pronunciation lexicon can become more robust using list of diphthongs. Hence, Urdu speech database would be more accurate and will represent the quality speech of native speakers. Moreover, this study reports that the sounds /ə/ and /j/ replace with /æ/ and the sounds /ɪ/ and /j/ replace with medial /e/ to form diphthongs. These alternation results are based on the speech of six native speakers. This phenomenon needs to be studied on a large sample to confirm the trend of vowel shifting or alternation among the native Urdu speakers.

7. Acknowledgement

This work has been conducted through the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers supported through a research grant from ICTRnD Fund, Pakistan.

8. References

[1] B. Mumtaz, A. Hussain, S. Hussain, A. Mehmood., R. Bhatti, M. Farooq, & S. Rauf. "Multitier Annotation of Urdu Speech Corpus", Conference on Language and Technology (CLT14), Karachi, Pakistan, 2014.

[2] L. Swaran, "Challenges for Design of Pronunciation Lexicon Specification (PLS) for Punjabi Language", Department of Information Technology, Govt of India, 2011.

[3] J. C. Catford, "Fundamental Problems in Phonetics", Edinburgh: Edinburgh University Press, 1977, pp 215.

[4] P. Ladefoged,. "A Course in Phonetics", Los Angeles: University of California, ed 2, 1982, pp 171.

[5] M. A. Khan, "Urdu Ka Sauti Nizam", National Language Authority, Pakistan, 1997.

[6] A. Q. Khan, "A Preliminary Study of Pahari and Its Sound System", The criterion An International Journal in English Vol IV, Issue IV, 2013, pp 1-20.

[7] A. Waqar, & S. Waqar, "Identification of Diphthongs in Urdu and Their Acoustic Properties", 2002.

[8] F. Alam, S. M. Habib & M. Khan, "Acoustic analysis of Bangla vowel inventory" BRAC University, 2008.

[9] R. D. Kent, C. Read, & R. D. Kent, "The acoustic analysis of speech", (Vol. 58). San Diego: Singular Publishing Group, 1992.

[10] F. Okati, P. Helgason, & C. Jahani, "Diphthongization in Five Iranian Balochi Dialects", Orientalia Suecana, Vol 61, 2013, pp 107-119.

[11] C. P. Prinsloo, "A comparative acoustic analysis of the long vowels and diphthongs of Afrikaans and South African English", 2006.

[12] B. Comrie, "Languages of the world. *The Handbook of Linguistics*", 2001, pp 522 (online book: cited on May 09,2016).

[13] C. P. Masica, "*The Indo-Aryan Languages*", Cambridge University Press, 1993.

[14] Samare, Yadollah (Samare, Yadollāh). "*Āvāšenāsi-ye zabān-e fārsi, āvāhā va sāxt-eāvāyi-*

[15] *ye heǰā*", Tehran: Markaz-e Našr-e Dānešgāhi, 1368 [1989–90].

[16] S. Ganjavi, P. G. Georgiou, & S. Narayanan, "ASCII based transcription systems for languages with the Arabic script: The case of Persian", 2003, In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on IEEE, pp 595-600.

[17] M. Yaesoubi, "*Machine Transcription Conversion between Perso-Arabic and Romanized Writing Systems*", Master thesis, Linköpings University, 2010. Online at: http://liu.divaportal.org/smash/record.jsf?pid=divas2:360341 [Retrieved May 30, 2016].

[18] A. Hakimi, "Comparative Phonetic Study of Frequently Used Words in Iranian Farsi versus Tajik Farsi", *Journal of American Science* 8(4), 2012, pp 6–16.

[19] K. Khurshid, S.A. Usman, & N. J. Butt, "Possibility of existence and identification of diphthong sand triphthongs in Urdu language". Akhbar-e-Urdu, 2003, pp 16-21.

[20] A. Sarwar, S. Ahmed, & A. A. Tarar, "Diphthongs in Urdu language and analysis of their acoustic properties", Annual Student Report, Center for Research in Urdu Language Processing (CRULP), 2003, pp 9-15.

Appendix A

Sr No.	Diphthong	Average Duration Males (ms)	Average Duration Females (ms)	Minimum Duration males in Unstressed (ms)	Minimum Duration females in Unstressed(ms)	maximum Duration males in stressed (ms)	maximum Duration females in stressed (ms)
1	ɑːeː	254	207	127	147	296	267
2	ɑːe	231	212	186	169	280	280
3	ɑːɛ̃ː	282	212	200	170	277	276
4	ɑː ɪ	212	175	182	134	288	228
5	ɑːiː	233	205	205	150	261	262
6	ɑːoː	205	221	127	179	268	286
7	əeː	236	186	177	153	272	226
8	æɑː	321	215	162	148	313	342
9	əiː	234	186	176	153	302	226
10	əĩː	235	206	201	147	289	258
11	eɑː	207	195	142	142	241	271
12	ɪõː	203	152	179	146	225	165
13	ɪũː	227	126	160	147	273	209
14	oːiː	244	192	229	96	267	293
15	uiː	201	210	167	161	231	327
16	ɑːĩː[7]	300	285	350	336	322	354

[7] Diphthong rejected on the basis of durations.

Appendix B

		First Component			Second Component		
Sr No.	**Diphthong**	**Average F1 Hz**	**Average F2 Hz**	**Average F3 Hz**	**Average F1 Hz**	**Average F2 Hz**	**Average F3 Hz**
1	ɑːeː	813	1707	2990	501	2407	2977
2	ɑːe	797	1376	2700	542	2035	2970
3	ɑːẽː	890	1534	3144	716	2221	3126
4	ɑː ɪ	807	1527	2832	531	2154	2928
5	ɑːiː	788	1694	2792	330	2379	3017
6	ɑːoː	712	1431	2884	612	1184	2956
7	əeː	454	1830	2631	422	2187	2752
8	æɑː	525	2048	2759	746	1541	2622
9	əiː	517	1835	2682	319	2477	3094
10	əɪ̃ː	418	2228	3039	332	2118	3046
11	eɑː	419	2103	2763	683	1492	2582
12	ɪõː	351	2104	2893	553	1412	2827
13	ɪũː	331	1637	2596	347	1146	2732
14	oːiː	431	1212	2782	332	2360	2889
15	uiː	313	1735	2708	295	2464	2974

Automatic Derivation of Nouns from Adjectives in Pashto

Tariq Naeem
naeemtarik@gmail.com
Mohammad Abid Khan
mabid@upesh.edu.pk
Department of Computer Science, University of Peshawar, Peshawar

Abstract

The paper explains the development of class changing derivational morphological analyzer of Pashto language. This can examine various derivations of Pashto nouns from adjectives using a corpus. The lexemes (Pashto words) are tested through finite state machines and are able to accept input in the form of Pashto adjectival derivation. The input is given in Arabic-scripted Pashto equivalent form. The derivational morphological analyzer displays all the occurrences of lexemes (words) and sentences by a simple search in the corpus.

1. Introduction

There are quite a lot of Natural Language Processing (NLP) applications. Morphological analysis is a pre-imperative in different areas of NLP e.g. are lemmatizers, stemmers and spell checkers in dictionaries/corpora. This paper addresses the development of an application that can perform the derivational analysis of the inputted word showing its root along its morphological units and features (also called morpho-analysis) by retrieving all the sentences of the use of the word from Pashto corpus.

A significant usage of the system, prepared in this paper, can be in the improvement of Part of Speech (POS) tagging program and stemmers in Pashto language.

Derivations, in computational linguistics, are also called lexeme formation and word formation. According to [1] derivational affixes change completely the grammatical category of the root words to which they are connected. Derivation is divided into two parts i.e. class-maintaining derivation and class-changing derivation. This paper explains the class changing aspect of derivation in Pashto.

The rest of the work is devised in following manner. Section 2 covers the related work of class changing derivations. Section 3 explains the existing of class changing derivation of nouns from adjectives in Pashto language. Section 4 describes the transformation of those linguistics patterns, discussed in section 3, in FSTs. Discussions and results are drawn in section 5. Section 6 concludes the complete study. Limitations and future work of the proposed system is discussed in section 7.

2. Related Work

[2] Investigated Paumari verbs as derivational additions which assisted in different ways. The author depicted the particular affixes utilized as a part of Paumari language with eminent relevance.

[3] Inspected the association between morphologically referred lexemes in Malay which can adapt to derivational morphology in pairs. Their tests demonstrate translation from Malay into English. These authors determined that new lexemes can be shaped by three morphological procedures i.e. Affixation, Compounding (the shaping of new lexemes by placing two or more lexemes together) and Reduplication (also called word repetition) [1]. Inflection and derivation are because of the sub-classifications of affixations yet they are in contrast to each other [2].

As per [4], morphological generation is a critical matter for producing derivative frames of word from semantic representation. Arabic language, of Semite peoples, has frequent derivational appearances. The research work of [4] demonstrates that great part of the Arabic vocabulary is comprised of words which are deduced from stems by insertion of prefixes, infixes and suffixes. In initial stage, these authors perceive prefixes and suffixes. Then at that point utilizing affixes limitation, they filter the mixed up relationship of affixes to perceive the radical word (root word). Their outcomes show whether this

'radical word' (new lexeme) has a place in Arabic language [4]. The work of these authors influences the derivation of words from the root words thereby changing its class. These authors demonstrated the linguistic constructs of word categorization and morpho-syntactic enactment on a written work. As a result, the new word gives a whole different meaning and changes its grammatical family [4] [3].

The work of [5] examines the role of sub-lexical units as a response for dealing with productive derivational strategies, in the building of a lexical utilitarian grammar for Turkish. Such sub-lexical units make it possible to reveal the internal structure of words with various derivations to the grammar rules consistently [5]. In conclusion, this reminds more minimal and sensible guidelines. Further, the semantics of the findings can similarly be purposely seen in grouping.

The procedures of framing new words from existing words through the affixation of morphemes or by removal of affixes where the resulted new lexeme is different from its previous original root goes to a different grammatical category. [6] Trained a derivational analyzer for Hindi over a past existing inflectional analyzer. In the proposed methodology, the authors determined words in Hindi language which were examined to get the derivational affixes. Then the patterns were arranged by understanding the properties of the affixes.

[7] Proposes morphological analysis using word and paradigm (Prefix-Suffix) model using a corpus. These authors evidenced the parser precision and effective use of memory and the corpus.

[8] Give a clear engineering model of a basic and precise framework for developing a morphological analyzer with finite state transducers (FST) using Telugu noun forms. These authors have indicated the regular expressions and unicode for their data format. Rule based methodology was utilized for constructing the morphological analyzer for Tamil language. The machine learning techniques were utilized for carrying out morphological analysis for Tamil

[9]. These researchers have trained separate modules for verb and noun. These authors have segregated their morphological analyzer in three levels. The first module, to begin with, is the pre-processing level. The input is changed over into a section of a sequence of units for handling by the morphological analyzer. Second module segments the linguistic units called morphemes. As per the morpheme boundary, preprocessed words are fragmented into smaller chunks. In third and last module is distinguishing morphemes. The segmented morphemes are given to the developed analyzer; it then predicts the linguistic classification to the segmented morphemes.

A blended methodology (of previous methodologies) for building a morphological analyzer for Malayalam language is used by [10]. These authors combined the methodologies of root (word) recognizer and addition suffix stripping approach. They enforced a lexical dictionary for better outcome and fast processing.

[11] Have done a relative study on Malayalam language using distinctive methodologies e.g. Brute force technique, Root driven strategy and Suffix stripping. By this hybrid approach, these authors took the advantage of both paradigms and had separate groups of classes whose morphophonemic patterns are the same. Their Malayalam morph-analyzer assists in automatic spelling and sentence structure checking, NLU (Natural Language Understanding), Speech synthesis, POS (Part of speech) taggers and parsers.

Before commencing this computational study, the work done by Pashto linguists was studied. They are Roberts [12], Tegey [13], Ziyar [14], Tegey and Robson [15], Babrakzai [16] and Rishteen [17]. The work of these language specialists frame the premise for the exploration work exhibited in this paper.

1. Class Changing Derivatives from Adjectives to Nouns in Pashto

The work of [1] and [2] can also be implemented for Pashto language. The affixes (prefixes and suffixes) can also be found in Pashto. According to [18], derivation can take place from adjectives to abstract nouns as follows:

Example 3.1

وږي [wagey]'hungry' or لوږَ [lawaga]'hunger'.
لوړه تنده پر غالبه شوه یک باره
[Lowagah tandah prey ghaliba showa yak barah]
په صورت ور پاتِ نه شه طاقت توان
[pah sorat wer paeti nah sha taqat twan]

"Hunger and thirst all at once overpowered him.
In his body no power or strength remained." [18]

In the above example, it is clear that the suffix ي [ye] of **وږي** is removed and prefix**ل** [laam] is added to the root instead to change the word from adjective to noun to make it a new lexeme.
The word **لوږَ** or **لوړه** is same, the word is sometimes written with ◌َ [zabar] and sometimes **ه**. This is a poetic stanza and we know that poets molds words the way they want and that's why it is sometime very hard to understand what they want to say.

Example 3.2

تږي [tagey] ‘thirsty’, تنده [tandah] or تندَ [tandah] ‘thirst’.

لوړه تنده نه شته دَ قانع په قناعت کښې

[lowagah tandah nah shtah da qanae pa qina’at ke]

دا کیمیا چه زده کا په خرقه کښې اُمرا وي

[da chemia che zda ka pah kharqah kae omurah we]

“In the contentment of the contented man, there is neither hunger nor Thirst;

And they who acquire this alchemy will be noble, tho’ clad in rags.” [18]

The تنده [tandah] ‘thirst’ is formed by dropping two letters in the تږي [tagey] ‘thirsty’ i.e. (ږ and ي)and three other letters i.e. (د , ن and ه) are affixed with ت [tey] to make it a noun from adjective.

Example 3.3

رُونړ [ronrr] or رُونڼ [ronrr] ‘bright’, رَنړا [rarra] or رَنڼا [rarra] ‘brightness’.

په رنړا ئې دَ چا کار نه پوره کیږي

[pah rarrae ked a cha kaar nah porah kege]

دَ آسمان برق و بریښنا ده دا دنیا

[da asman barq o braekhna dah da dunya]

“By the Light of it the business of this life cannot be perfected;

For this is as the lightening and the light of the sky.”[18]

Sometimes the lexeme takes ئِي **[aey]**, as in the following example:

لکه نمره په جهان وُ خیژي رنړائي شي

[lakah namra pah jahan aokhejey rarrae she]

دم قدم هسي زنده کاندِ اخلاص

[dam qadam hasse zinda kande ikhlaas]

“As when the sun riseth on the world, *Light* and *Brightness* cometh,

So doth friendship and affection give life to both breath and footstep?” [18]

In both examples discussed above و of the word رُونړ **[ronrr]** is bumped off and ا **[alif]** is appended with the root word in order to change the lexeme from adjective to noun.

Example 3.4

تور [tor] ‘dark’ or ‘black’ تّیاره [tiyarah] or تیارَ **[tiyara]** ‘darkness’ or ‘blackness’.

کل جهان توره تیاره شه له هغه ګرد و غبار

[kul jahan tora tiyarah sha lah hagha garrd o ghubaar]

آسمان رعد بریښیده تکه شمسیران

[asmaan ra’ad brikheda takah shamsheraan]

The whole world filled with *Darkness* from this dust and vapour;

In the heavens thunder and lightning flushed as from swords.”[18]

The infix و in the lexeme تور **[tor]** is removed and یا is infixed to change it to a new lexeme تیارَ **[tiyara]** i.e. from adjective to noun.

Example 3.5

یون ګران په لار دي بوالهوس ته

[yoon Gran pah laar de boalhoos tah]

مرد هغه ګنړه چه بنیګړي که بنا

[mard hagha garrah che khaegarre keh bina]

“Journey on this road is difficult to the fickle and capricious:

Consider him a man who layeth the foundation of *Goodness*”. [18]

In the above illustration, ګړ and ي is infixed with the lexeme بنه to make it noun from adjective.

The whole of the nouns of the above classes mentioned from 1 to 5 are feminine. The following i.e. 6 to 8 are all masculine, with the exception of those words formed by affixing تیا **[tiya]**, ستیا **[stia]**, ګلوي [galwe]and ولي **[wali]**, are feminine.

Example 3.6

مخ دِ سپین لکه آفتاب وُه

[makh de speen lakah aftab woh]

تر آفتابَ ئې لا تاب وُه

[ter aftaba ye la taaba woh]

ولي اوس دا هسي تور شه

[wali aos da hassi toor sha]

په تورواليِ لکه سکور شه

[pah toorwaali lakah skoor sha]

“Thy countenance was white like unto the sun- yea!
It was brighter than the orb of the day:
But now, alas! It is become so black,
That it’s *Blackness* is like unto charcoal”. [18]

It is justified from above examples that the lexeme تور **[toor]** is an adjective by suffixing والي **[waali]** to تور **[toor]** makes تور والي **[toorwaali]** which is a noun.

Example 3.7

کله ما وته امید دَ خپل ژوندوُن شي

[kalah ma wata umeed da khpal zowandoon she]

په هجران به ئي ژوندوُن راته زبون شي
[pah hijraan ba ye zowandoon rata zaboon she]
"When shall I entertain hope for my own *E*xistence?
Since separated from her, *Life* itself to me is infamous". [18]

The ئ **[ye]** of the word ژوندئ **[zowande]** is removed and ون **[woon]** is affixed instead to change the word into a new lexeme.

Example 3.8

ناګاه ويښه شوه له خوبَ
[na gah wekha showa lah khoob]
ز ړه ئی ډک له مَيَن توبَ
[zrra ye ddak la maentoob]
کښيناسته نګاه ئی وُ کړ
[kaenaasta nigah ye okarr]
يار ئی نه ليد آه ئی وُ کړ
[yaar ye na lid aah ye okarr]
"Suddenly she awoke from her slumbers,
Her heart filled with *Love* and *Affection*.
She sat up and gazed around, but signed
For she beheld not her beloved one." [18]

In this example, second stanza, the word توبَ **[toob]** is suffixed with مَيَن **[maen]** which is an adjective to make a new lexeme مَيَن توبَ **[maentoob]**, a noun.

خداي دِ نه کا ندِ بيلتون دَ دوهٴ يارانو
[khudae de nah kande baeltoon da dowao yaraano]
په بيلتون عاشق په روغ صورت بيمار دي
[pah baeltoon aashiq pah rough sorat bemaar de]
"God forbid that *Separation* should be caused between two lovers;
For in *Separation* the lover, though healthy in body, is sick at heart." [18]
In the above example, ټون [toon] is suffixed with the word بيل [bael] to make it a noun i.e. بيلتون [baeltoon]
چه په ديدن مِو ړتيا نه شوه
[che pah dedan de morrtiya na showah]
اوس د يار غمو کړي مو ړ
[os de yaar ghamo karre moorr]
"Whereas from her presence thou didst not acquire *Satiety*,
Grief on her account has now *Satiated* thee" [18]

In the second stanza of the above example ړ مو **[مو ړ]** is an adjective but by putting the suffix تيا with مو ړ makes ړتيا مو which is a noun.

4. Modeling and Finite State Transducers

Finite state machines (FSM) also known as finite state transducers (FST) efficiently compute many useful Natural Language Processing (NLP) functions and weighted transitions on strings. In many fields of NLP, FSTs are applied as a core technology in developing spell checkers, parsers, POS taggers, speech recognition systems, morphological analyzers, information retrieval systems and lexical analyzers [19]. This work encouraged the modeling and design of FSTs. Using FSTs, the natural language modeling is efficient and more effective because they are mathematically derived models, by Chomsky, and are well-understood [20].

a. وړي **[wagey]** 'hungry' or لوړَ **[lawaga]** 'hunger'

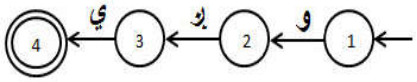


Figure 1: The modeling of adjective وړي 'hungry'

The strings recognized in this finite automata is لوړَ **[lawaga]**'hunger' with zabar and لوړه **[lawagah]** 'hunger' with ه **[hy]** as it ends at state 5. Both are nouns.

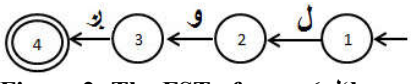


Figure 2: The FST of noun لوړَ'hunger'

The above automaton, its derivative is وړي [wagey] 'hungry' is form when the finite automaton starts from the initial state and ends straight at the final state 4 is an adjective. In this finite automaton we have 6 states.

b. تږي **[tagey]** 'thirsty',تنده **[tandah]** or تندَ **[tandah]** 'thirst'.

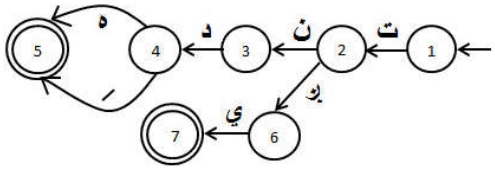


Figure 3: Automaton of adjective changed to noun

In this finite automaton, we have 7 states. Starting from state 1, marked as initial state and rests at state 5 and 7 as they are the final states. This automaton forms three strings. First, if we start with state 1 and move along with transition to state 5, two strings are identified which has the same semantics. The strings are تندَ**[tandah]** 'thirst' with zabar and تنده **[tandah]**

‘thirst’ with ه **[hy]** are nouns. Its derivatives are formed by following the path from state 1, 2, 6, and 7 i.e. **تږي [tagey]** ‘thirsty’ is an adjective.

c. **رُونړ [ronrr]** or **رُونڼ [ronrr]** ‘bright’, **رَنړا [rarra]** or **رَنڼا [rarra]** ‘brightness’.

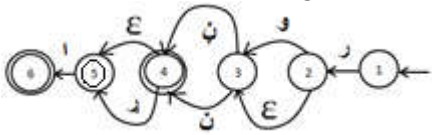


Figure 4: Automaton of both adjective رُونڼ ‘bright’ and noun رَنړا [rarra] ‘brightness’

The above automaton accepts variety of words. It has total 6 states with 3 marked as final and 1 as initial.

The first word which is recognized by this automaton is **رُونڼ [ronrr]** ‘bright’ if we follow state 1, 2 and 3. You can get **رُونړ [ronrr]** ‘bright’ by following the path from 1, 2, 3, 5 and 6. The first derivative of the two nouns are **رَنړا [rarra]** ‘brightness’ is recognized by state 1, 2, 3, 4, 5 and 6. Furthermore, if you go from the transition marking state 1, 2, 3, 4 and 5 will get the string **رَنڼا [rarra]** ‘brightness’.

d. **تور** [tor] ‘dark’ or ‘black’ **تیاره** [tiyarah] or **تیارَ [tiyara]** ‘darkness’ or ‘blackness’.

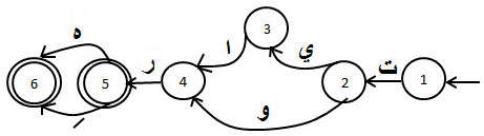


Figure 5: The FST of adjective تور ‘dark’ and noun تیاره ‘darkness’

The next string and its derivative can be recognized by visiting the state 1, 2, 4 and 5 and its derivative, Princeton, New Jersey.

can be extracted by the following two paths from state 1 to state 6 taking the alternative path **ې** to state 3 will lead you to **تیارَ [tiyara]** ‘darkness’ or ‘blackness’ and the string **تیاره** [tiyarah] ‘darkness’ or ‘blackness’.

e. **ښه** [khah] ‘good’, **ښیګړه** [khaegarrah] ‘goodness’.

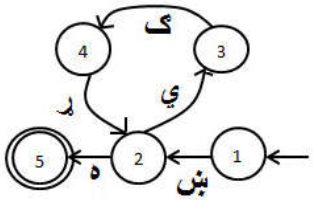


Figure 6: FST of adjective ښه ‘good’ and noun ښیګړه ‘goodness’

The strings **ښه** [khah] ‘good’ and **ښیګړه** [khaegarrah] ‘goodness’ is recognized in a similar and simple manner. The word **ښه** [khah] ‘good’ is recognized by visiting state 1, 2 and 3 only follow the alternative path from the initial state.

f. **تور [toor] ‘black’, تورواليي [toorwaali] ‘blackness’; کلک [klak] ‘hard’, کلک والي [klakwaali] ‘hardness’**

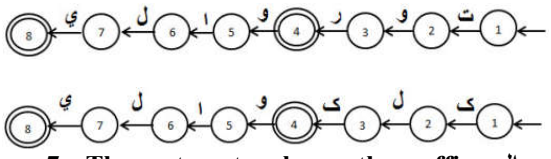


Figure 7: The automata shows the suffix والي [waali] with adjectives to make it a noun

The automata of **تورواليي [toorwaali] ‘blackness’** and **کلک والي [klakwaali] ‘hardness’** is more or less the same. The have the same number of states with the same states as final states. The string **تورواليي [toorwaali] ‘blackness’** is recognized as follow. The string **تور [toor] ‘black’** is recognized by starting at the initial state and ends at state 4. And its derivative **تورواليي [toorwaali] ‘blackness’** take all the way long from initial to the latter final state i.e. state 8.

Starting at state 1 and stops at state 4 give us the string **کلک [klak] ‘hard’**. If we go straight from state 1 to state 8 we get the adjective **کلک والي [klakwaali] ‘hardness’**.

g. **ژوندێ [zowande] ‘alive’ or ‘existing’, ژوندون [zwande] ‘life’, ‘existence’; نښتي [nkhate] ‘captive’, ‘prison’, نښتون [nakhtoon] ‘captivity’, ‘imprisonment’.**

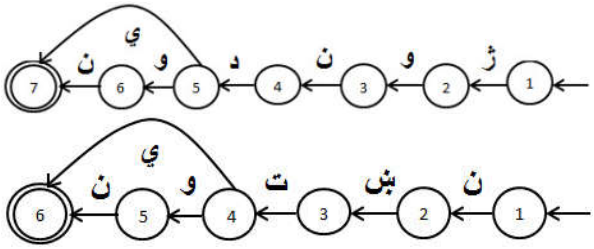


Figure 8: Automata depicts adjectives with suffix ون [oon] converting it to nouns

The finite automata of **ژوندون [zwandon] ‘life’** and **نښتون [nakhatoon] ‘captivity’, ‘imprisonment’** have the same number of states. The string **ژوندون [zwandon]** ‘life’ is produced by starting at the initial state i.e. state 1 and to states 2, 3, 4, 5, 6 and finally stopping in state 7, marked as the final state. Its adjective **ژوندێ [zowande]** ‘alive’ or ‘existing’ takes all the way long from initial state 1 to states 2,

3, 4, 5 and then jumping straight to state 7 on the input symbol يّ.

Similarly, the other word نښتون **[nakhatoon]** starting at state 1 and to states 2, 3, 4, 5 and finally stopping at the final state 6, reading all the input symbols generates a noun. The string نښتي **[nkhate] 'captive', 'prison'** is an adjective and can be produced if we start crawling from state 1 and to states 2, 3, 4 and finally to state 6.

h. بيل [bael] 'separate', بيلتون [baeltoon] 'separation'; خاي [zaey] 'a place', خايتون [zaeytoon] 'a dwelling place', 'a home', 'a birthplace'; مَين [maen] 'affectionate' مَينتوب [maentoob] 'affection', 'love', ليوني [lewane] 'mad', ليونتوب [lewantoob] 'madness'; مورتيا [morr] 'satiated', مورتيا [morrtiya] 'satiety'; خمسور [khamsoor] 'impudent', خمسور تيا [khamsoortiya] 'impudence', 'familiarity'.

The finite automata with the suffix تون and توب are drawn. The strings can be recognized in the following manner.

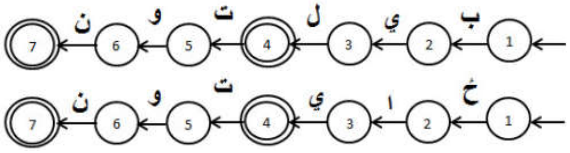


Figure 9: FSTs indicates the suffix تون [toon] with adjectives converting it to nouns

Starting at initial state and visiting only 2, 3, and 4 will give you the string بيل **[bael] 'separate'** but if you move along till state 7, you'll get بيلتون **[baeltoon] 'separation'** which is the derivative of the first.

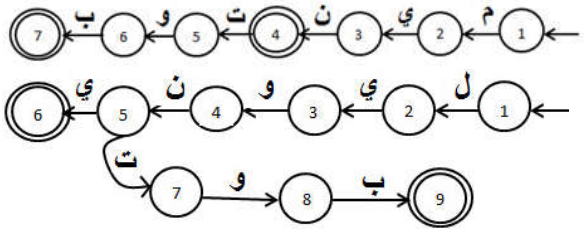


Figure 10: Automaton drawn shows adjectives changed into noun by the suffix توب [toob]

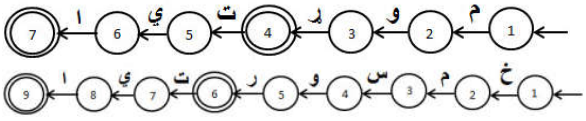


Figure 11: FSTs depicts the suffix تيا [tiya] with adjectives to convert it into nouns

5. Results and Discussions

In this section, the implementation of derivational morphological analyzer is discussed. The FSTs, created during the modeling stage, are executed and ensured.

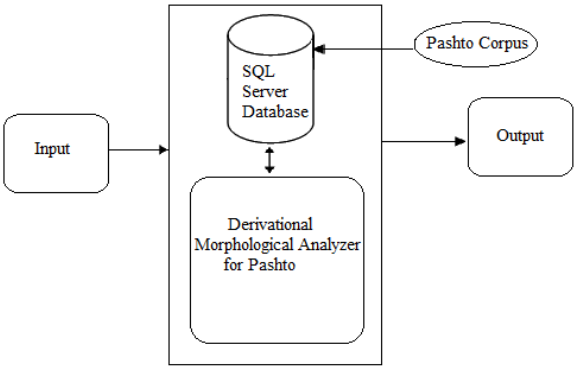


Figure 12: Entire framework overview

5.1 Pashto Corpus and its use in proposed work

Developing a corpus of Pashto language was the first and foremost activity to be used in proposed work. A Pashto corpus was created utilizing the corpus improvement tool XML Aware Indexing and Retrieval Architecture (XAIRA). Tagging of Pashto text was done in Extensible Markup Language (XML). The corpus is composed of Pashto textual material. Data was collected from different areas, like, news, memos, letters, research articles, books, fiction, sports and magazines, making it a representative corpus. Then, these XAIRA tagged files were used in Microsoft Structured Query Language (SQL) Server Management Studio 2012 tables. Each entry in database tables has sorting rules based on alphabet or language and comparison styles using collation feature of Microsoft SQL Server. The 'nvarchar data type' was used to insert Pasto text in SQL tables. The window locale collation to insert Pashto text is "Pashto_100_BIN". Collation determined for unicode data specify rules for text entered in columns. The corpus as of now contains 0.5 million words containing lexemes with its grammatical class. The corpus is utilized for considering the derivational morphological arrangement of Pashto. The derivations were broke down into stems and affixes. An example is displayed given in table 1.

Table 1 Derivational Stems and Suffixes

DESKTOP-HU8H0NP....bo.ExtractedData

	Root	GramClass1	Affix	RootAffix	GramClass2
	تور	Adjective	والي	توروالي	Noun
	کلک	Adjective	والي	کلکوالي	Noun
	ژوندۍ	Adjective	ون	ژوندون	Noun
	نښتي	Adjective	ون	نښتون	Noun
	بیل	Adjective	تون	بیلتون	Noun
	ځای	Adjective	تون	ځایتون	Noun
	هټن	Adjective	توب	هټنتوب	Noun
	مور	Adjective	تیا	مورتیا	Noun

The assessment of derivations demonstrates that the nouns in Pashto language have several types based on suffixes. Each of these examples illustrated in Example 3.1 to example 3.8 has an exceptional type of morpheme unlike from the other families for depicting the equivalent aspect.

5.2 Implementation of derivational morphological analyzer

The automata were implemented which were formulated during the modeling and design stage. Four programming languages and tools were used for implementation. First, the XEROX LEXC tool was used. LEXC (Lexicon Compiler), is a writing apparatus for making vocabularies and lexical transducers. The LEXC tool was used to draw finite state diagrams from the lexicon of adjectives and nouns in Pashto.

A universal application XEROX, (XEROX Finite State Transducer) XFST is very useful for processing FSTs. Compilation from .txt files are done efficiently by XFST reading from binary files. XFST gives numerous approaches to get data around a system and to review and adjust its structure. The input of XFST, in our proposed study, was the output generated by LEXC compiler.

```
LEXICON Root
  مور        Adjective;
  خمسور      Adjective;

LEXICON Adjective
NounSuffix;
#;

LEXICON NounSuffix
  تیا #;
    #;
```

Figure 13: Sample lexicon processed by XEROX's LEXC

The sample of lexicon in figure 13 is the representation of lexicon developed and stored in notepad with utf8-mode for reading Pashto text. The LEXC compiler starts evaluating the Root lexicon first when compiled by XEROX's LEXC compiler. Then, it trails through other lexicons by suffixation rules.

```
Starting in utf8-mode.
lexc> compile-source lexc_CCD.txt
opening "lexc_CCD.txt"
Opening 'lexc_CCD.txt'...
Root...2, Adjective...2, NounSuffix...2
Building lexicon...Minimizing...Done!
SOURCE: 2.5 Kb. 11 states, 11 arcs, 4 paths.

lexc>
```

Figure 14: LEXC interface

Figure 14 shows the compilation of lexicon displayed in figure 12. The source in LEXC compiler depicts a total of 11 states, 11 arcs and 4 paths transducer. The lexicon of the finite state transducers, thus developed, contains 621 nouns and 953 adjectives. This information is carried and processed by XEROX XFST tool. The XFST generated output file which was forwarded and stored in Microsoft SQL Server. Finally, the Microsoft Visual Studio CSharp dot net framework was used as a front end to work on the corpus stored in Microsoft SQL Server.

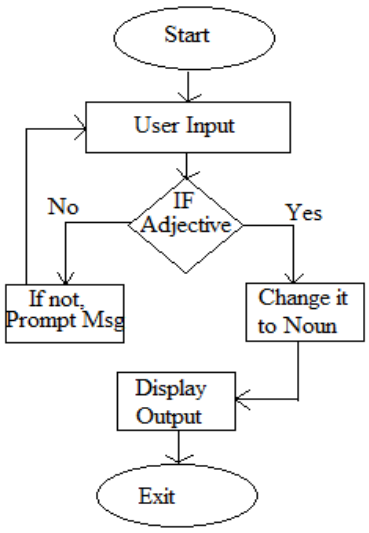


Figure 15: Flowchart of the system

The flowchart of figure 15 depicts the complete working of derivational morphological analyzer converting adjectives to nouns. The screen shot of sample interaction with the proposed system is given in figure 16:

Figure 16: The developed derivational morphological analyzer

The above morphological analyzer analyzes derivation from adjectives to nouns. In corpus, the sentence containing the query word retrieves all the sentences and displays them in the left pane. Similarly, the morphological analysis is done in the search query.

5.3 Error Analysis

The accuracy of the system was measured by error analysis was performed. The output was recorded and matched manually with the system generated output. A sample of 174 nouns and 169 adjectives were collected from the written Pashto data and given to the system as input. Out of this input 147 nouns and 142 adjective words were accurately examined. Consequently, the complete accuracy of the overall system is:

((147+142) / (174+169)) * 100 = 84.25%

Due to limited lexicon size, the system could not search and analyze the root word with its derivational affixes in Pashto corpus.

15.75% of the errors were because of typographic variation and borrowed words from other languages in Pashto.

6. Conclusion

In this paper, the derivational properties of Pashto language are discussed. There are two kinds of derivations that exist in Pashto i.e. Class changing and class maintain derivation.

Being the first derivational morphological analyzer created for Pashto, it achieved 84% overall accuracy and it can be further improved by using big corpus. A lot of work has been done to build a representative corpus in Microsoft SQL Server 2012. The collation feature, not available in older versions, was used to store the corpus in its original form.

Due to complexities involved in derivation, this area is very sensitive for making new words to enrich Pashto because a slight change of ◌ً **[zabbar],** ◌ٍ **[zeir]** and ◌ٔ **[paish]** can be disastrous.

7. Limitations and Future Work

The current state-of-the-art derivational morphological analyzer changes adjectives to nouns. This paper discussed class changing derivation of Pashto. Further work can be done on class maintaining derivatives in Pashto language. Also, the corpus size can be increase to achieve higher accuracy for derivational rules.

References

[1] M. Aronoff and K. Fudeman, "What is morphology?", Fundamentals of linguistics 1, Blackwell Publishing, Malden, MA, **2005**, 1-3

[2] Chapman S., Paumari Derivational Affixes *Associacao Internacional de Linguistica SIL Brasil* **2008**, 9, 10.

[3] M.A Malik, "An approach to the study of linguistics" New Kitab Mahal, Lahore, **2010**, 116-118

[4] P. Nakov; H.T. Ng. Translating from morphologically complex languages: a paraphrase-based approach *HLT' 11: Proceedings of the 49th Annual Meeting of the ACL* **2011**, 1299.

[5] H.G. Raverty, "A grammar of the Pashto or Afghan language" Bombay, Calcutta, **2007**, 9-12

[6] Buckwalter T.; Arabic Morphological Analyzer version 2.0 LDC **2004**, 11.

[7] K.G. Mkanganwi, Shona (derivational) Morphology: Observation in Search of a Theory. *Zambezia* **2002**, 175,176

[8] Cetinoglu Ozlem; Oflazer Kemal, Morphology-syntax interface for Turkish ACL **2006**, 153,160, 10.3115/1220175.1220195

[9] Kanuparthi N.; Inumella A.; Sharma D.M., Hindi derivational morphology analyzer, ACL **2012**, 12

[10] Uma Maheshwar Rao G; Ambar Kulkarni P.; Christopher Mala, The study effect of length in morphological segmentation of agglutinative languages *ACL*, **2012**, 18, 19.

[11] D.L Sneha and K. Bharadwaja, "A novel approach for morphing Telugu Noun forms using finite state transducers" IJERT, **2013**, 2(7), 550.

[12] V.P Abeera, S. Aparna, R.U Rekha, K. Anand, Dhanalakshmi, Soman and Rajendran, "Morphological Analyzer for Malayalam using Machine Learning" ICDEM'10, **2010**, 253

[13] Vinod P M; Jayan V; Bhadran V K, Implementation of Malayalam morphological analyzer based on hybrid approach *ACL* **2012**, 310

[14] Jisha P.; Jayan; Rajeev; Rajendran, Morphological Analyzer for Malayalam – A comparison of different approaches *IJCSIT* **2009**, *2*. 156,157

[15] Roberts Taylor, Clitics and Agreement **2000**, 17-23, Massachusetts Institute of Technology (MIT), Linguistics

[16] Tegey Habibullah, The grammar of clitics: Evidence from Pashto (Afghani) and other languages **1977**, University of Illinois
[17] Ziyar Mujawar Ahmad, Pashto grammar, **2005**, 83-99, vol-3, Danish Publishers branch association
[18] Tegey Habibullah; Robson Barbara, A reference Grammar of Pashto **1996**, 46-88, Center for Applied Linguistics, Washington D.C
[19] Babrakzai F., Topics in Pashto syntax, PhD thesis, Linguistics department, University of Hawaii, Hawaii, **1999**
[20] Rishteen SaqeedUllah, Pashto grammar, **2001**, 389-392, University Book Agency publisher
[21] C.E. Biddulph, Afghan Poetry of the 17th Century: being selected from the poem of Khushal Khan Khattak Denzil lbbetson Atlantic publishers, London **2006**, 31.
[22] Beesley Kenneth R.; Karttunen Lauri, Finite State Morphology, **2003**, 501-505 Stanford, CA: CSLI Publications
[23] Chomsky Noam, On certain formal properties of grammars, **1959**, 141-166, Massachusetts Institute of Technology, Massachusetts and The institute for advanced study

Named Entity Dataset for Urdu Named Entity Recognition Task

Wahab Khan[a]*, Ali Daud[b,a], Jamal A. Nasir[a], Tehmina Amjad[a]
[a]*Department of Computer Science and Software Engineering, IIU, Islamabad 44000, Pakistan*
[b]*Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
{wahab. phdcs72, ali. daud, jamal. nasir, tehminaamjad} @iiu.edu.pk

Abstract

Named entity recognition (NER) and classification is a very crucial task in Urdu. One challenge among the others which makes Urdu NER task complex is the non-availability of enough linguistic resources. The NER research for English and other Western languages has a long tradition and significant amount of work has been done to solve NER problems in these languages. From resource availability aspect Western languages are counted resource plentiful languages. On the other hand, Urdu lags far behind in terms of resources. In this paper we reported the development of NE tagged dataset for automated NER research in Urdu, especially with machine learning (ML) perspectives. The new developed Urdu NER dataset contains about 48000 words, comprising of 4621 named entities of seven named entity classes. The contents source of this new dataset is BBC Urdu and initially contains data from sport, national and international news domain. This new dataset can be used for training and testing purpose of various statistical and machine learning models such as e.g. hidden Markov model (HMM), maximum entropy (ME), Conditional random field (CRF), recurrent neural network (RNN) and so forth for conducting computational NER research in Urdu. Our goal is to make this dataset freely and widely acquirable, and to promote other researchers to exercise it as a criterial testbed for experimentations in Urdu NER research. In rest of the paper the new NER dataset will be referred as UNER dataset.

Key Words: Urdu, Named Entity Recognition (NER), Resources, Machine Learning (ML)

1. Introduction

Named Entity Recognition (also known as entity identification, entity chunking and entity extraction) is a task to identify and classify all proper nouns in texts into predefined categories, such as persons, locations, organizations, expressions of times, quantities, monetary values, etc. NER is an important subtask of many natural language processing tasks, such as information extraction, co-reference resolution, relation extraction, question answering, machine translation, etc [1-3]. NER system came in focus during the Message Understanding Conferences (MUCs) [3, 4]. After that plethora of NER techniques and systems are available [1, 5]. Most of these systems are developed for European languages [6], especially English and are fairly accurate. Many automatic frameworks for NER have been proposed for other non-European languages like Arabic, Persian and South Asian languages [7, 8]. For Urdu language, NER systems are yet in developing phase [9]. As most of the existing NER systems rely on the use of rich external linguistic resources e.g. annotated corpora, human-made dictionaries, gazetteers etc., to improve the system accuracy [10]. Urdu language lacks in having abundant linguistic resources. Characteristics of Urdu language make the NER task more difficult. For example, Capitalization is a bloom characteristic employed by NER systems for European languages Urdu language does not have Capitalization. Thus, Urdu NER task requires detailed analysis and adaptation.

2. The Urdu Language

In Pakistan more than sixty languages are verbalized, including a numeral of provincial languages. Urdu, also known as lingua franca, is the national language of Pakistan. Recently, Supreme Court of Pakistan also ordered the central Government of Pakistan to adopt Urdu as an official language throughout the country. Urdu is the main source of communication nationwide and is easily understood by about 75% people in Pakistan. But

statistics shows that Urdu is mother tongue of about 8% Pakistanis. Urdu is also a most popular language in India. In India Urdu is official language of six states and also the constitution of India recognizes Urdu as one among the 22 recognized languages. As per Wikipedia statistics, there are about 65 million native speakers of Urdu in both India and Pakistan. In Pakistan there are around eleven million Urdu speakers, 52 million native speakers are in India and world widely there are more than 300 million Urdu speakers [11, 12]. Pakistan, India, USA, U.K, Gulf countries, and Canada own Urdu speakers in copiousness. In the last few years NLP research community observed incredible fast growth of multilingual content on the web. As a result, NLP research community is attracted to explore monolingual and cross-lingual Information Retrieval (IR) tasks [3]. Initially, the web was designed to present information to users in English, but gradually with the passage of time, with the development of standard technologies and with opportunity of accessing the web resources uniformly around the globe, the web became multilingual source of information. Monolingual IR is centred on the queries and information present in same language, while cross-lingual IR is centred on the queries and information provided in numerous varied languages [13].

In the recent years, South Asian languages have gained intense research interest. Particularly, Urdu language is major research stake holder in Asian language processing research community [9].

NLP tasks such as part of speech (POS) tagging and named entity tagging etc. have paramount importance in all NLP systems [14]. NLP systems developed for English and other Western languages have criterial accuracies compared to Urdu [15-17].

3. Available Dataset

For automated Urdu language processing existence of bench mark datasets is mandatory. To train machine learning models, it is necessary to provide relevant pre-labeled training data and in large amount [3, 14]. As far as Urdu is concerned, it lacks in having abundant linguistic resources for development of Urdu NER tools and conducting experiments. The Center for Language Engineering (CLE)[8] in Pakistan has taken initiatives efforts in corpus-building activities as well in promoting and conducting research in Pakistani and many other Asian languages. For promoting research in Urdu, CLE has lunched many linguistic resources for Urdu language processing. Details of the available CLE linguistic resources can be found one the home page of CLE store http://www.cle.org.pk/clestore/index.htm. CLE provides all the linguistic resources with little amount of processing fee.

Although CLE provides a variety of linguistic resources but, so far, Urdu named entity tagged dataset is still missing from their linguistic resources list. CLE store have POS tagged dataset which is about 100K size. This POS tagged dataset can be used NER task to assign POS tags to words and then to use these POS tags as feature in training phase of ML models.

4. Related Work

The research work about Urdu NER task using machine learning techniques is still in initial stages. The core reason is the non-availability of standard NER linguistic resources. To accomplish supervised machine learning based Urdu NER task, a large pre-labeled NER dataset is mandatory [18], which is not available in case of Urdu. The research community from ULP domain is limited to use only two available NE tagged dataset for machine learning based research in Urdu NER task. The first one is the IJCNLP-2008 NE tagged dataset.

The IJCNLP-2008 dataset comprises of about 40000 words and in its annotation, twelve named entity classes are used.

The research group namely Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan and IIIT Hyderabad, India have jointly make efforts to create the IJCNLP-2008 dataset, after creation they donated it to the NER workshop [3, 12]. This dataset can be downloaded freely from its source UR: http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5. The second NE tagged dataset which we refer as Jahangir et al., [19], is a dataset of about 31860 words with total 1526 named entities. The dataset is annotated with four named entity classes. Details of these two dataset can be found in

[8] http://www.cle.org.pk/

Table 1. In Table 2 entity wise statistics are provided of Jahangir et al., and IJCNLP-2008 NE Tagged Dataset. The statistics are extracted from the available version of both the datasets with authors of this paper.

Table 1 Details of Two NE tagged datasets

Dataset	No. of Words	No. of Sentences	No. of NEs
Jahangir et al.,	31,860	1,315	1526
IJCNLP-2008	40408	1097	1115

Table 2 Entity wise Statistics

Entity	IJCNLP-2008	Jahangir et al.,
Person	277	380
Location	490	756
Organization	48	282
Date	123	101
Number	108	---
Designation	69	---

5. Development

Linguistic resources for most southeastern languages are not radially available due to which these languages are termed scared resource languages. Urdu is a southeastern language, and is spoken in a vast area of sub-continent. It is a low resource language. Due to resources scarceness not plenty of research work has been actioned for Urdu language [3].

The dataset we present contains all text from BBC Urdu cyber space. The current version UNER NE tagged dataset contains text from three news domain e.g. 1) national news 2) international news 3) sports news. In future, we will include text from other popular news domains e.g. entertainment, science, business, health not only form BBC Urdu but from other sources such as Express news, Dunia news etc. The size of our current NER dataset is about 0.48k words with total 4621 named entities. Entities in the text are manually tagged in the guide line of IJCNLP-2008 and Jahangir et al., dataset. Initially only seven named entity classes are used in tagging. The seven named entity classes used in tagging includes: PERSON, LOCATION, ORGANIZATION, DESIGNATION, NUMBER, DATE, TIME. After manual tagging the samples from the three domains are reviewed through Urdu linguistic experts from two different organizations, and changes mentioned by them are incorporated accordingly in the whole dataset. During the development process all the entities are tagged from right to left and also text is stored sentence level. The symbols "-" dash and "?" are used as sentence separators. During tagging entity are enclosed in start and end tags. E.g. The entity (پاکستان, Pakistan) where it occurs in text it is tagged with stat and end tags of LOCATION such as <LOCATION>پاکستان</LOCATION>. For all the seven entity class labels the same approach is adopted. For storage purpose we used notepad with UTF-8 encoding system. Text I n files are organized sentences wise because most of machine learning models takes inputs sentences wise. E.g. CRF, RNN, DRNN and so on.

The UNER NE tagged dataset we reported is freely available for research purposes and can be requested by sending an email to any of the authors. Although UNER NE tagged dataset can be used for rule based work, but its structure and organization is more feasible for machine learning based approaches. We hope that our NER dataset will help in promoting ML based research in Urdu particularly in NER task. Below Table 7,Table 8 and Table 9 shows single sentence from each news domain in which named entities are labeled with its corresponding class label.

Table 7 Structure of National News Domain

<LOCATION>پاکستان</LOCATION> کے صوبہ
<LOCATION>بلوچستان</LOCATION> کے دارالحکومت
<LOCATION>کوئٹہ</LOCATION> میں فائرنگ کے واقعے میں
<NUMBER>ایک</NUMBER> پولیس اہلکار سمیت
<NUMBER>تین</NUMBER> افراد ہلاک ہوگئے ہیں۔

Table 8 Structure of International News Domain

<LOCATION>پیرس</LOCATION> میں شدت پسند حملوں سے
<PERSON> عبدالقدیر حکیم منسلک ایک اور شدت پسند
<TIME>دو روز</TIME> قبل بھی <PERSON>مدنی
<LOCATION>عراق</LOCATION> کے شہر
<LOCATION>موصل</LOCATION> میں مارا گیا ہے۔

Table 9 Structure of Sports News Domain

<LOCATION>آسٹریلیا</LOCATION> کی جانب سے پہلی اننگز میں
<PERSON>مچل</PERSON> سٹارک
<NUMBER>5</NUMBER> وکٹوں کے ساتھ سب سے کامیاب بولر
<NUMBER> نے <PERSON>ہیزل وڈ</PERSON> ہے جبکہ

تین</NUMBER> اور سپنر <PERSON>لیون</PERSON> نے
<NUMBER>دو</NUMBER> وکٹیں لیں۔

As most of the ML models require the training data in the format of IOB2 (Inside, Outside and Begin), IOE2 (Inside, Outside, and End) and SBME (Single, Begin, Middle and End) format [2]. The UNER dataset is fully compatible with the above mentioned format and one can easily convert into any format easily. Table provides an overview of SBME format of UNER dataset.

Table 6 Example of SBME format

Token	Type
پیرس	S_Location
عبدالقدیر	B_Person
حکیم	M_Person
مدنی	E_Person
دو	B_Time
روز	E_Time
عراق	S_Location
موصل	S_Location

While S means that this entity is single, B represents the beginning of an entity, M represents middle portion of an entity while the letter E represents the ending of an entity.

Details of our presented UNER dataset can be found in Table and

Table . Consolidated statistics such as total number of words, total number of comprising named entities and total number of sentences are provided in Table . Domain wise consolidated statistics of each named entity class are provided in

Table . Table provides lists of typical named entity types that are considered during construction process of our NE tagged dataset along with its description.

We intent to balance the dataset with consideration of genre, and proportion of each entity class.

Table clearly reflects that the proportion of DATE and TIME entity classes are quite small compared to others because the occurrence and mentioning these two entities in national, international and sorts news are not customary. After annotation process the whole dataset is stored in 150 notepad documents using UTF-8 encoding scheme. Table provides domain wise document detail of UNER dataset.

We believe that UNER dataset is a rich dataset and has ability to compensate the indigence's of NER and NLP research, utilizing the present-day Urdu.

Table 7 Consolidated Statistics of UNER Dataset

Total of No. of Words	48673
Total No. of sentences	1744
Total No. of Named Entities	4621

Table 8 Domain wise consolidated statistics of each entity class

Entity\Domain	Nat:	Inter:	Sport	Total
Person	401	201	605	1207
Location	390	360	455	1205
Organization	400	210	53	663
Designation	167	70	42	279
Number	270	132	589	991
Date	81	74	48	203
Time	40	23	10	73
Total	1749	1088	1809	4621

Table 9 Domain wise No. of Documents

Domain	File No.	No. of Document
National	1-60	60
Sports	61- 110	50
International	111- 150	40
Total		150

Table 10 List of Generic Urdu Named Entity Types with the kind of Entities they refer.

Type	**Tag**	**Sample Category**
Person	<PERSON>	Individuals, small groups
Location	<LOCATION>	Territory, land, kingdom, mountains, site, locality etc
Organization	<ORGANIZATION>	firms, group of players, Political parties, bureau etc
Designation	<DESIGNATION>	Various designations e.g. Professor, Dean, Mufti, Captain etc.
Number	<NUMBER>	Counts e.g.

		Hundred, Ten Thousand One, 10 million etc.
Date	<DATE>	Date stamps
Time	<TIME>	Clock time stamps

6. Conclusion

These days the state of the art approaches that are widely adopted around the globe for the development of NER tools in almost all languages, including Urdu are machine learning approaches. The core reason behind its wide usage is based on four features: a) the capability of automatic learning b) the degree of accuracy c) the speed of processing and d) generic nature. The basic need for ML approaches for training and testing is the availability of pre NE tagged dataset. As far as Urdu is concerned, it is termed as resource poor language. Therefore, in this work we tried to contribute in Urdu language resource with a large enough newly created NE tagged dataset. Significant efforts were made to build this huge NE tagged dataset compared to existing NE dataset with text from multi news domains. The fascination aspect of the UNER dataset is its size as well as its very rich NE contents. These two aspects make UNER dataset more feasible for ML techniques. We hope that this new dataset will spark light in ULP research community and will attract researcher in future to promote research in ULP.

7. References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, pp. 3-26, 2007.

[2] M. K. Malik and S. M. Sarwar, "urdu named entity recognition and classification system using conditional random field," *Science International* vol. 5, pp. 4473-4477, 2015 2015.

[3] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review,* pp. 1-33, 2016.

[4] B. M. Sundheim, "Overview of results of the MUC-6 evaluation," in *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996*, 1996, pp. 423-442.

[5] A. Roberts, R. J. Gaizauskas, M. Hepple, and Y. Guo, "Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation," in *LREC*, 2008.

[6] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 2003, pp. 142-147.

[7] K. Shaalan and H. Raza, "NERA: Named entity recognition for Arabic," *Journal of the American Society for Information Science and Technology,* vol. 60, pp. 1652-1663, 2009.

[8] U. Singh, V. Goyal, and G. S. Lehal, "Named Entity Recognition System for Urdu," in *COLING*, 2012, pp. 2507-2518.

[9] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 9, pp. 1-43, 2010.

[10] J. i. Kazama and K. Torisawa, "Exploiting Wikipedia as external knowledge for named entity recognition," in *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 698-707.

[11] K. Riaz, "Baseline for Urdu IR evaluation," in *Proceedings of the 2nd ACM workshop on Improving non english web searching*, 2008, pp. 97-100.

[12] S. Hussain, "Resources for Urdu Language Processing," in *IJCNLP*, 2008, pp. 99-100.

[13] J. Capstick, A. K. Diagne, G. Erbach, H. Uszkoreit, A. Leisenberg, and M. Leisenberg, "A system for supporting cross-lingual information retrieval," *Information processing & management,* vol. 36, pp. 275-289, 2000.

[14] K. Riaz, "Rule-based named entity recognition in Urdu," in *the 2010 Named Entities Workshop*, 2010, pp. 126-135.

[15] F. Adeeba and S. Hussain, "Experiences in building the Urdu WordNet," *In proceesings of the 9th Workshop on Asian Language Resources collocated with IJCNLP, Chiang Mai, Thailand* pp. pp. 31-35, 2011.

[16] E. T. Al-Shammari, "Towards an Error-Free Stemming," in *IADIS European Conf. Data Mining*, 2008, pp. 160-163.

[17] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration,"

in *Conference on Language and Technology*, 2009.

[18] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait journal of Science,* vol. 43, pp. 66-84, 2016.

[19] F. Jahangir, W. Anwar, U. I. Bajwa, and X. Wang, "N-gram and gazetteer list based named entity recognition for urdu: A scarce resourced language," in *10th Workshop on Asian Language Resources*, 2012, pp. 95-104.

Acoustic Investigation of /l, j, v/ as Approximants in Urdu

Sahar Rauf, Saadia Ambreen, Sarmad Hussain, Benazir Mumtaz
Center for Language Engineering, Al-Khwarizmi Institute of Computer Sciences
University of Engineering and Technology
Lahore, Pakistan

Abstract

This presented research work aims to investigate the acoustic properties of /l, j, and v/ in Urdu as approximants. For the acoustic analysis of /l, j, and v/, data has been recorded from 4 native speakers of Urdu (2 males and 2 females) and total 280 utterances of approximants have been recorded at three positions of word i.e word initial, middle and final. Two experiments have been conducted using PRAAT to investigate the acoustic properties of approximants in Urdu; first experiment is based on the spectrogram analysis of approximants and second experiment analyzes the periodicity level of these approximants. The second experiment is conducted by calculating the median of Harmonicity to Noise Ratio (HNR) values of these sounds. The analysis indicates that approximants in Urdu also behave like fricatives. Moreover, this research also explores the controversial issue about the existence of aspirated approximants i.e. / l^h, j^h, v^h /. Results indicate that /j^h/ is no more used by Urdu speakers. Only two aspirated approximants exists in Urdu i.e. / l^h, v^h /.

Keywords: *Urdu approximants, fricatives, acoustic measures, median, HNR*

1. Introduction

The term "approximant" was first coined by Ladefoged and he defined it as approximants belong to the two phonetic classes; one is resonant oral and second is consonant
[1]. He also claims that approximant is an "approach of one articulator towards another but without the vocal tract being narrowed to such an extent that a turbulent airstream is produced"
[2].The second definition is also used by IPA as standard. Approximants have vocal tract which is not as much closed as it is for the consonants and not as much wide as it is for the vowel but the vocal tract of approximant is in the middle of vowel and consonants. That's why approximants have both the properties of vowels and consonants.

Trask [3] on the other hand placed approximants in between vowels and fricatives because of the constriction of the airflow and he claims that approximants produce frication noise. However, researchers sometimes disagree and try to find out which segments come under the approximant category.
[3].The aim of this research is also to investigate the acoustic properties of /l, j, v, l^h and v^h / as approximants in Urdu language. These sounds are selected in this research because approximants in Urdu do not always behave like semi-vowels. They also appear as fricative in some cases. The motivation behind this research is to find out the contexts where approximants behave like semi-vowels and where approximants turn themselves into fricatives.

The paper is organized as follows: Section 2 overviews different features of approximants that have been researched in different languages, the procedure for carrying out the research is discussed in section 3, results and data analysis of Urdu approximants is given in Section 4. Section 5 illustrates the research findings and finally, conclusion and dimensions for the future work are presented in Section 6.

2. Literature review

IPA [4] categorizes approximants as laterals [l, ɭ, ʎ ɻ], non-laterals (or centrals) [ʋ ɹ ɻ, β̞] and semi-vowels [j ɥ w ɰ]. [ʋ β̞] are "Spirant approximants" or approximant-like versions of voiced fricatives. A special openness diacritic [˕] is used to indicate the approximant like version of fricative. Spirant approximant [ʋ] is found in Dutch [3] and it is assumed that the approximant behave as a fricative when it comes at onset and coda position. /j/ with [˕] is also used in Spanish to show the noise or turbulence in the /j/ in emphatic speech [2].

Acoustic measures have been used to analyze the linguistic properties of American English semivowels [5]. Different features such as sonority, consonantal, high, front and retroflex were analyzed and used to distinguish the /l, w, j, r/ from one another. The corpus was consisted of 233 polysyllabic words and collected from 4 speakers (2 males and 2 females). The sonorant property differentiates semivowels from other consonants and retroflex property separates /r/ from /l, j, w/. Formants values differs for each approximant as F3 is weaker for /w, l, r/ and stronger for /j/. F1 differentiates glides /w, j/ from the liquids /l, r/ and F2 value separates /w/ from other semi-vowels as it falls below 1000Hz than others.

Korean language has three approximants [w, j, l] [6]. At word initial position, [l] is deleted when it is followed by [i] or change to [n]. [l] is produced as [r] at the word initial position of loan words and some Korean names. [w] and [j] are glides and mostly come at prevocalic position in Korean language. [w] phoneme shows same vocal tract like [u] vowel when it comes before [ɪ,e,a,ə] and [j] shows the same vocal tract as of [i] when it comes before the vowels [e,a,u,ə,ɛ,o]. The basic difference among Korean stops, glides and vowels transitory pattern is the duration. The transitory duration of stops is shorter, vowels have longest transitory pattern whereas glides have intermediate transitory duration. F1 and F2 frequencies of [w] are lower than [u], F1 of [j] is lower and F2 is higher as compared to [i]. Korean [r] has very short duration but in this duration, closure and release time period is present. In Korean language, the formant values of approximants are higher in female speakers as compared to the male speakers. Different /w/ variants of Korean have been investigated in [7]. Two variants of /w/ have been explored due to F2 transitions; high back glide [w] with lower F2 and high front glide [ч] with higher F2. /w/ is usually fully realized as [w] before [a] and realized as the [ч] before [e] but not before [a] for both males and females.

Glides or semi-vowels in Sindhi language have some acoustic vowel characteristics like; formant structure and periodic wave forms [8]. In Sindhi language, Glides [w] and [y] are voiced phonemes and voiced region can also be seen acoustically. Glides show sharp transitory segment. In Sindhi language, /w/ shows periodic signal activity and voicing at the low region of the glide. It is bilabial, F1 moves downward when it approaches to the glide and moves upward when it moves away from the glide. On the other hand, F2 remains unchanged. /y/ is palato-alveolar glide in Sindhi language. It shows energy at lower region of the glide and periodic signals in the spectrogram. It also shows sharp transitory formant transition. F2 goes upwards when it approaches to the glide and moves downward when it goes away from the glide. On the other hand, F1 motion is vice versa to the F2 motion.

Allophonic variation of /v/ and /w/ has been studied for Hindi in [9]. There is only one grapheme in Devanagari for these two phones. To find out the answers to the queries, 154 words were selected with possible prosodic positions. The speech thus analyzed through the use of speech form editor and lip movement has also been noted for some cases. Through the spectrogram analysis, it has found that /w/ shows lower second formant with lack of friction. Moreover, moraic structure is also considered to describe the syllable structure in these allophones.

Different parameters have been used to analyze the approximants for different languages. However, in case of Urdu approximants very little or unpublished work has been found. Some Urdu approximants have been studied through acoustic analysis but the major challenge was to finalize a proper method that can justify the presence of these approximants in Urdu. So the aim of this research is to explore the existence of /l, j, v, l^h/ and/v^h/ as approximants in Urdu.

3. Methodology

Three Urdu consonant /l, j, v/ have been studied acoustically. These consonants have been selected because of their vowel like formant patterns. Aspirated sounds /l^h/ and /v^h/ have also been studied to find out their existence in Urdu language. In order to analyze the acoustic properties of these consonants, speech data is recorded from 4 speakers (2 males and 2 females). For this study, the data has been collected from only 4 speakers as the time was required to find out the appropriate method to study the Urdu approximants and also less work was found on it.

Data is recorded in an anechoic chamber at the sampling rate of 8 kHz. PRAAT software is used to record and analyze the data. These sounds have been studied at three word positions; initial, middle and final. The selected numbers of utterances for /l, j, v/ are given in Table 10.

Table 10 No of utterances for /l, j, v/ at three positions of word

Phonemes	No. of Utterances		
	Initial	Middle	Final
/l/	10	10	10
/j/	10	10	-
/v/	10	10	-

The utterances have been embedded into the carrier phrase e.g.

1 - میں نے ----کہا
2. /mæ̃: ne: ….. kəha:/
3. I have said …..

The above mentioned 70 utterances have been recorded by each speaker so total of 280 utterances have been recorded and analyzed to find out the results. Some instances of /v/ at final position have also been recorded and analyzed. The words were /عضو/ / /uzv/ /body part / and /جزو/ /ʤuzv/ /portion or part/. However, analysis of these utterances indicates that speakers do not pronounce /v/ sound at the end of words.

In Urdu [10] there is direct relationship of word orthography and its pronunciation but still there are some words in Urdu which are not pronounced as they were written specifically in case of /l^h and v^h/. To study the acoustic property of the aspirated /l^h and v^h /, words have been selected from Urdu Lughat [11] and Oxford dictionary [12]. In both dictionaries, there was no instance of /j^h/ and only one instance of /v^h/ has been found. So the selected words for /l^h/ was /دولها/ /ɖu:l^{h}a:/ /Groom/ and / چولها/ /ʧu:l^{h}a:/ /Stove/. For /v^h/ sound, the selected word was /وهیل/ /v^he:l/ /Whale/. Alternative versions of these words were also found in dictionary; /دولہا/ /ɖulha:/ /Groom/, /چولہا/ /ʧulha:/ /Stove/ and /وہیل/ /vhe:l/ /Whale/ . However, for the recording purposes standard words with /ہ/ were given to the speakers to investigate whether speakers can articulate aspirated versions of approximants or not.

3.1. Acoustic analysis: experiment 1

Five Urdu consonants /l, j, v, l^h and v^h/ are investigated acoustically in this paper to find out their occurrences as approximants or fricatives. All of the selected consonants are analyzed at three word positions; initial, middle and final. Along with the study of formant like patterns or frication noise in sounds, other acoustic cues durations and formant transitions are also used to differentiate among these sounds. See Section 4 Table 2 and 3 for the results of acoustic analysis.

3.2. HNR analysis: experiment 2

The selected sounds are also tested through calculating the median of HNR values of /l, j and v/ utterances. HNR measures the replacement of harmonic structure in spectrogram by the frication noise [13]. The method for harmonicity median is described in [14] that measures the acoustic periodicity. In this research, HNR values have been calculated using PRAAT. For calculating the HNR median, the number of frames of each sound utterance and their values were extracted from PRAAT query after analyzing the periodicity. Afterward, the values are calculated to find out median of the sound. See Appendix A for the median values.

4. Results and data analysis

This session is subdivided into two sub-sections i.e. experiment 1 results and experiment 2 results along with their analysis.

4.1. Experiment 1 results and data analysis

Table 2 Duration values of /l, j, v/

Urdu phonemes	Duration at word initially with pause	Duration at word medially	Duration at word finally with pause
/l/	105ms	78ms	118ms
/j/	93ms	65ms	-
/v/	70ms	56ms	-

Table 11 Formant values of /l, j, v/

Urdu phonemes	Formant values word initial		Formant values word medial		Formant values word final	
	F1	F2	F1	F2	F1	F2
/l/	292	1584	325	1584	295	1592
/j/	311	1830	306	1867	-	-
/v/	290	1212	324	1289	-	-

Acoustically, /l/ sound takes formants and exists at all three word positions; initial, middle and final. When /l/ occurs at word initial position with pause, its duration increases up to 27ms than its word medial position version. No transition is found in succeeded vowel and formants remain stable but lighter than the succeeded vowel. Formant values for /l/ at three positions are given in Table 11. At word medial position, the duration of /l/ decreases that is described in Table . Formants become lighter at word medial position. At word final level with pause, /l/ duration increases up to 40ms than its word middle position. /l/ forms light formants but sometimes takes light frication with formants and voicing at the base level. Figure presents /l/ at middle position.

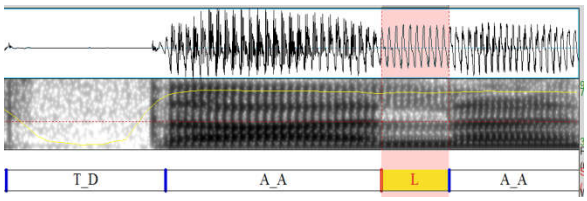


Figure 1: /l/ at middle position

In Urdu, /j/ occurs at only two positions; word initially and medially. /j/ behaves differently in different contexts. At word initial with pause position, /j/ behaves in three ways (i) vowel like (ii) fricative like and (iii) both vowel and fricative like. The average duration of /j/ increases at word initial level up to 28ms. After acoustic analysis, it has been observed that the dual characteristic of /j/ is only the part of speaker (Sp) utterances. It is also found out that its duration value is more than the other values and reaches up to 132ms. Figure 2 presents the dual property of /j/ at initial position with preceded pause.

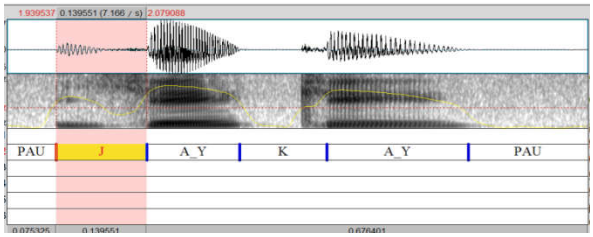


Figure 2 Dual property of /j/

It is difficult to analyze when /j/ comes medially, it sometimes takes frication and sometimes takes formants. This phenomenon is tested by experiment 2 and described in detail in section 4.2. When /j/ comes between two vowels, it gets merged with the vowel. However, in this context intensity helps a lot as a slight decrease in intensity indicates the presence of /j/ e.g. in /کیا/ /kija:/ /Did/ or / آیا / /a:ja:/ /Come/. Table 11 indicates that /j/ has higher F2 values than other sounds.

When /v/ comes at initial position, its duration increases up to 14 ms than its middle position and it can take the property of a fricative or semi-vowel. /v/ forms light formants and sometimes it has only frication. At middle position, F2 of the preceding vowel falls whereas F3 remains same. This behavior of medial /v/ has also found across the speakers which might be an indication of occurrence of /v/ as retroflex in Urdu. /v/ at final position has been analyzed and found out that speakers have not pronounced /v/ at this position. The F2 value of /v/ is also significant that is lower than all other sounds and differentiates its property from others.

The data from the speakers has been analyzed and found out that /l^h/ is not pronounced in word /دولها/ /ɖu:l^{h}a:/ /Groom/. Instead of pronouncing /l^h/, speakers have pronounced /l/ and /h/ separately as /ɖulha: /دولہا/. Other selected aspirated word was /v^he:l/ /وہیل/ /Whale/ and it is observed that speakers has pronounced it as /ve:l/ /ویل/ /Whale/ without the aspirated sound. The speakers have changed the aspirated /v^h/ into un-aspirated /v/. In Urdu, /ہ/ /do chasmi hey/ is used to represent the aspiration of a sound [10] . However, it is observed that in Urdu Lughat [11] and Oxford dictionary [12] words like /وہیل، دولہا، چولہا/ are also written as /ویل، دولہا، چولہا/. The alternative orthographies of such words shows that /l^h / is gradually replacing with /l+h/ sounds and /v^h/ with its un-aspirated version /v/.

4.2. Experiment 2 results and data analysis

The findings from acoustic analysis are then tested by experiment 2 and try to answer the expected queries. Different behaviors of /l, j, and v/ have been tested in all selected positions and generate a table to show the occurrence of these sounds as approximant or fricative in percentages.

Table 4 Acoustic behavior of /l, j, v/ across speakers

No. of Sp	Positions	Approximant (%)	Fricative (%)	Mixed (%)
Sp1 (F)	/l/ initial	70	30	-
	/l/ medial	90	10	-
	/l/ final	100	-	-
	/j/ initial	20	20	60
	/j/medial	50	50	-
	/v/ initial	60	40	-
	/v/ medial	70	30	-
Sp2 (F)	/l/ initial	90	10	-
	/l/ medial	100	-	-
	/l/ final	100	-	-
	/j/ initial	70	30	-
	/j/medial	100	-	-
	/v/ initial	60	40	-
	/v/ medial	90	10	-
Sp3 (M)	/l/ initial	100		
	/l/ medial	80	20	-
	/l/ final	100	-	-
	/j/ initial	100	-	-
	/j/medial	100	-	-
	/v/ initial	80	-	20
	/v/ medial	90	10	-
Sp4 (M)	/l/ initial	90	10	-
	/l/ medial	100	-	-
	/l/ final	80	20	-
	/j/ initial	60	40	-
	/j/medial	100	-	-
	/v/ initial	50	50	-
	/v/ medial	100	-	-

Table 4 describes the percentages for approximant, fricative and mix property that have been discussed for /j/ sound. The results have been

given for each speaker in which the first 2 are females and last 2 are males.

/l/ at initial position is approximant in all speakers, only in speaker1 it is 30% fricative as it sometimes takes frication at start which causes less periodicity in /l/. At medial position, there were 2 utterances of speaker 3 and one utterance of speaker1 in which /l/ was found fricative. At final position, only 2 utterances of speaker 4 were found in which the /l/ was fricative.

In speaker 1, different behavior of /j/ has been observed at initial level. At this level, /j/ had dual behavior of approximant and fricative which is tested by the experiment and found out that there were 6 utterances of Sp1 in which /j/ had the dual property. However, it was observed that this feature was only the part of Sp1 utterances and Sp 1 was taking this feature with preceded pause in stress syllable. In Sp3 /j/ is 100% approximant. At medial position, there is equal tendency of approximant and fricative in Sp1 but in other speakers /j/ is 100% approximant. So, only sp1 is creating a slight disagreement with other speakers as it is showing very independent behavior.

There is tendency of /v/ to occur as either fricative or approximant at word initial position. There were 2 utterances in Sp3 that shown dual property (approximant and fricative) in /v/ at word initial position. At word medial position, /v/ is 30% fricative in Sp1 only and for other speakers' utterances it is more approximant like.

Through the experiment 2, median of HNR values have been calculated for each utterance. After the experiment conducted for all 4, it was observed that the threshold for the median values of HNR were different for males and females. For females, the value for voiceless fricatives was less than 3dB and the value for voiced fricative was reached up to 17dB. So the value more than 17dB was indicating that the consonant is more vowel like as the vowel values were starting from this range. However, the threshold for males was totally different. The value for voiced fricative was starting from 3dB and the vowels values were starting from 10dB. So, the value for a consonant at 10dB or above was indicating the occurrence of consonant as approximant.

Figure describes the /l/ initial median values of 10 utterances for each speaker. The different line colors show different speakers. The graph shows a gradual increase in Sp3 behavior that is above 10dB indicating 100% approximant behavior but the behavior in Sp4 is very abrupt going from voiced fricative to approximant and reaches up to highest range of 20dB. The graph lines for Sp1 and 2 show some utterances below 20dB as fricatives.

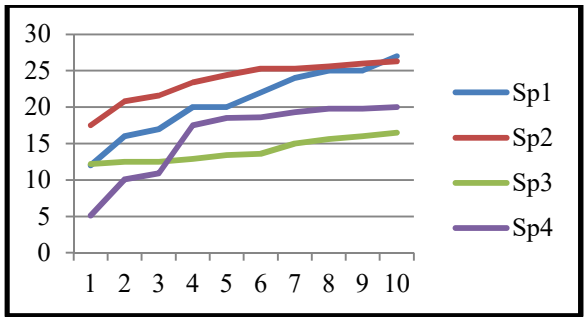


Figure 3 Median graph for /l/ initial position

Figure describes median values of /l/ at medial position. The graph describes 10 utterances for each speaker. It represents some utterances of Sp1 and Sp3 which have fricative property. It is observed through the analysis that /l/ takes fricative property in such cases due to the presence of other fricatives in its surrounding.

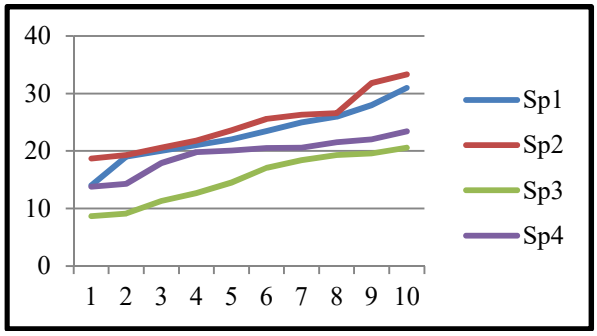


Figure 4 Median graph for /l/ medial position

Figure 5 describes the median values for /l/ at word final position. Graph presents that the lines of Sp3 and 4 are merging at the end as they are sharing same values. The figure indicates that only in Sp4, /l/ has fricative part and in all others the /l/ is 100% approximant. Only two utterances of Sp4 are fricative like because they occurred with following pause in the data which is actually the main cause of less periodicity in these two utterances. The lines show very smooth transition in the graph as the reason might be the values are less fricative.

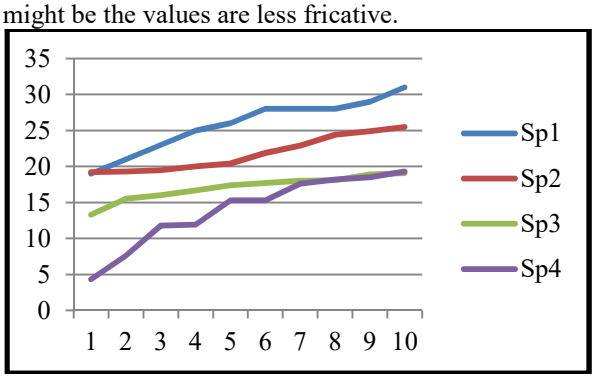


Figure 5 Median graph for /l/ final position

Figure represents the median values for /j/ at initial position. The figure represents only 3 speakers excluding Sp1. The reason is clear from the above described table that /j/ at initial position in Sp1 was showing dual behavior of approximant and a fricative at the same time. It is significant that in Sp3 the graph line is very gradual going above 10dB showing 100% approximant behavior. According to the graph, some utterances of /j/ have fricative like property which is context dependent as in some cases /j/ differentiated itself from high vowels /i:, e: and u:/ by taking fricative property.

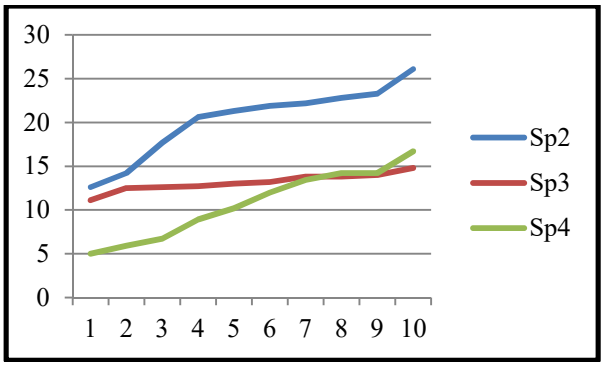


Figure 6 Median graph for /j/ initial position

Figure describes median values for /j/ at medial position. The three graph lines for Sp2, 3 and 4 show behavior of /j/ as approximant at medial position. The graph shows that the values for Sp2 and 4 are significantly high than the other ones. The impact of stress has also seen in Sp4 utterances, as the values of median are very high due to stress. Sp 1 again shows a slight discrepancy than others in the graph as it takes both behaviors with same ratio.

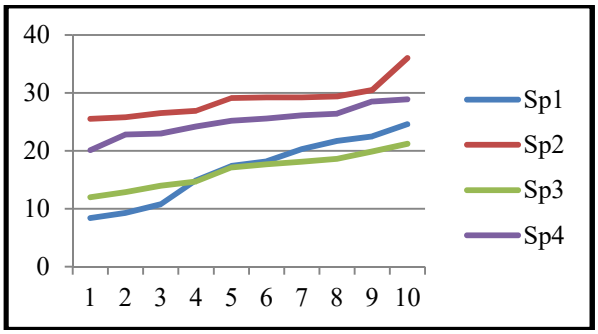


Figure 7 Median graph for /j/ medial position

Figure illustrates the median values of /v/ at initial position. The graph line for Sp3 has stopped at 8th utterance because of the reason that 2 utterances of the Sp3 were having the mix property like /j/ initial. But except those 2, the behavior of /v/ is approximant like in Sp3. /v/ at initial position is both fricative and approximant in other speakers. The fricative behavior of /v/ utterances is also due to the impact of neighboring fricatives on /v/ sound.

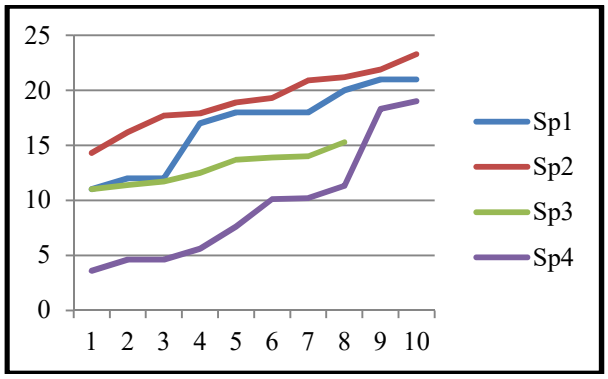


Figure 8 Median graph for /v/ initial position

Figure describes /v/ at medial position. A sharp decrease at utterance 1 below 0 for Sp 2 is due to the devoiced fricative value of /v/ at medial position. Except the 1st utterance of Sp2, other values show approximant behavior of /v/. Similarly, Sp1and 3 has some fricative values but Sp4 shows more gradual behavior above 10dB of /v/ medial as approximant. The median values at each position are also given in Appendix A.

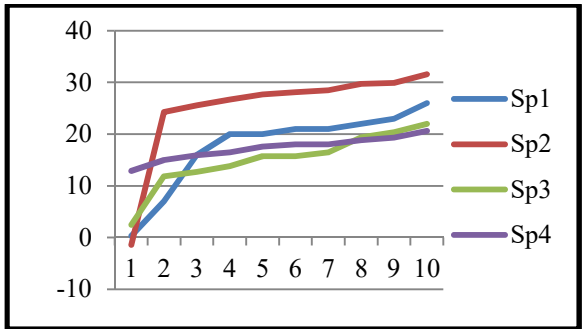


Figure 9 Median graph for /v/ medial position

5. Research Findings

Different behaviors have been observed through the acoustic and HNR analysis of /l, j and v/ in Urdu across speakers. Some features were very speaker specific. The dual property of /j/ at initial level has been observed in only Sp1. It is observed that this property in Sp1 occurs in /j/ initial when it comes with preceded pause specifically when it is stressed. It is also observed when /j/ comes with high vowels like /i:/, /e:/ or /u:/, it takes acoustic property of fricative to differentiate itself from high vowels. Similarly, when /j/ comes with /a:/ sound in unstressed context it becomes approximant. It is also observed that sometimes /j/ take frication because of its neighboring fricative consonant.

It is observed that the variations in the behavior of /l, j and v/ are speaker dependent. It is observed that the speaker 4 is using the fricative quality in case of unstressed context and in stressed context; it is using more approximant like property. The variation of /l/ as fricative has also been observed in speakers. /l/ takes frication when it is followed by a pause, which reduces the periodicity level in /l/. It is also observed that /l/ changes its acoustic property or lose formants when it comes with any fricative sound i.e. /h, x, s/ etc.

The analysis of /v/ reports that /v/ behaves like approximant when it comes at word medial position. However, there are some exceptions to this generalization as in the first utterances of Sp2 and 3; /v/ became voiceless fricative due to neighboring /h/ sound and /r/ respectively indicating neighboring fricatives can influence target sound.

6. Conclusion and future dimensions

/l, l^h, j, v and v^h/ sounds of Urdu have been investigated in this study to find out their acoustic properties as approximants. Two types of experiments have been conducted which report that /l, j and v/ can exist in Urdu both as an approximant and a fricative, although the percentage of approximant behavior is more than the fricative behavior. Analysis also indicates that different variations in these sounds are speaker dependent. Moreover, acoustic analysis indicates that /l/ shows longest duration at final position than others and /v/ shows lowest F2 values than others. Aspirated version of /l/ and /v/ is also studied in this research. Results tells that /l^h/ is now pronounced as /l and h/ and /v^h/ is mostly changed into its un-aspirated version by the speakers. There are other sounds in Urdu inventory like /r/ and /ɽ/ and their aspirated versions which are claimed to have approximant like behavior. In future, acoustic properties of these sounds would be analyzed using scientific methods.

7. References

[1] P. Ladefoged, *A Course in Phonetics*, 4th ed., Bill Hoffman, Ed. Calefornia, Los Angeles, USA: Earl McPeek, 1975.

[2] E. Martinez-Celdran, "Problems in the classification of approximants," *Journal of the International Phonetic*, vol. 34, no. 2, pp. 201-210, Dec 2004.

[3] I. E. Colombo, *On the Phonetic Status of Labial Approximants in Dutch*. University of Amsterdam, 2015.

[4] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, United Kingdom: Cambridge University Press, 1999.

[5] C. Y and E. Wilso, "Acoustic measures for linguistic features distinguishing the semivowels /w j r 1/ in American English," *The Journal of Acoustical Society of America*,vol.92, no.2, pp. 736-757, August 1992.

[6] M. C. Kim and A. J. Lotto, "Acoustic measurements of Korean approximants," *The Korean language in America*, vol.9, pp. 72-77, 2004.

[7] M. Oh, "/W/ Variants in Korean," *Journal of the Korean society of speech sciences*, vol. 2, no. 3, pp. 65-73, 2010.

[8] A. Keerio, L. D. Dhomeja, A. A. Shaikh, and Y. A. Malkani, "Coparative Analysis of Vowels, diphthongs and Glides of Sindhi," *Signal and Image Processing*, vol. 2, no. 4, December 2011.

[9] J. Pierrehumbert and R. Nair, "Implication of Hindi Prosodic Structure," in *Current Trends in Phonology: Models and methods (= Proceedings of the Royaumont meeting 1995),* University of Salford Press, pp. 549-584.

[10] S. Hussain, "Letter-to-sound conversion for Urdu text-to-speech system," in *Association of Computational Linguistics*, Geneva, Switzerland, 2004, pp. 74-79.

[11] *Urdu Lughat: Tarixi Usuul Per*, 1st ed. Karachi, Pakistan: Muheet Urdu Press, 2013, vol. 3.

[12] *Oxford Urdu-English Dictionary*, 1st ed. Karachi, Pakistan: Oxford University Press, 2013.

[13] Z. Bárkányi and Z. Kiss, "Is /v/ different?," in *Proc.Twenty years of theoretical linguistics in Budapest 25.*, Budapest, 2010, pp. 1-5.

[14] S. Hamann and A. Sennema, "Acoustic differences between German and Dutch labiodentals," in *ZAS Papers in Linguistics 42*, Berlin, 2005, pp. 33-41.

Appendix A

Table 1 Median values for /l/ initial position

Sp1	Sp2	Sp3	Sp4
12	17.5	12.19	5.1
16	20.8	12.5	10.1
17	21.6	12.5	10.9
20	23.4	12.9	17.5
20	24.4	13.4	18.5
22	25.3	13.6	18.6
24	25.3	15	19.3
25	25.6	15.6	19.8
25	26	16	19.8
27	26.3	16.5	20

Table 2 Median values for /l/ medial position

Sp1	Sp2	Sp3	Sp4
14	18.7	8.7	13.8
19	19.3	9.1	14.3

20	20.6	11.3	17.9
21	21.8	12.7	19.8
22	23.6	14.5	20.1
23.5	25.6	17.1	20.5
25	26.3	18.4	20.6
26	26.6	19.3	21.5
28	31.8	19.6	22
31	33.3	20.6	23.4

Table 3 Median values for /l/ final position

Sp1	Sp2	Sp3	Sp4
19	19.2	13.3	4.3
21	19.3	15.5	7.6
23	19.5	16	11.8
25	20	16.7	11.9
26	20.4	17.4	15.3
28	21.9	17.7	15.3
28	22.9	18	17.6
28	24.4	18.1	18.2
29	24.9	18.9	18.5
31	25.5	19.1	19.3

Table 4 Median values for /j/ initial position

Sp2	Sp3	Sp4
12.6	11.1	5
14.2	12.5	5.9
17.7	12.6	6.7
20.6	12.7	8.9
21.3	13	10.2
21.9	13.2	12
22.2	13.8	13.4
22.8	13.8	14.2
23.3	14	14.2
26.1	14.8	16.7

Table 5 Median values for /j/ medial position

Sp1	Sp2	Sp3	Sp4
8.4	25.5	12	20.1
9.3	25.8	12.9	22.8
10.8	26.5	14	23
14.9	26.9	14.7	24.2
17.4	29.1	17.1	25.2
18.2	29.2	17.7	25.6
20.3	29.2	18.1	26.1
21.7	29.4	18.6	26.4
22.5	30.5	19.9	28.5
24.6	36	21.2	28.9

Table 6 Median values for /v/ initial position

Sp1	Sp2	Sp3	Sp4
11	14.3	11	3.6
12	16.2	11.4	4.6
12	17.7	11.7	4.6
17	17.9	12.5	5.6
18	18.9	13.7	7.6
18	19.3	13.9	10.1
18	20.9	14	10.2
20	21.2	15.3	11.3
21	21.9	-	18.3
21	23.3	-	19

Table 7 Median values for /v/ initial position

Sp1	Sp2	Sp3	Sp4
0.3	-1.4	2.4	12.9
7	24.3	11.8	15
16	25.6	12.7	15.9
20	26.7	13.8	16.5
20	27.7	15.7	17.6
21	28.1	15.7	18
21	28.5	16.5	18
22	29.7	19.4	18.8
23	29.9	20.4	19.3
26	31.6	22	20.6

Subjective Testing of Urdu Text-to-Speech (TTS) System

Kh.Shahzada Shahid[1], Tania Habib[2], Benazir Mumtaz[2], Farah Adeeba[2], Ehsan ul Haq[3]
Centre for Language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore
[1]{khawaja.shahzada}, [2]{firstname.lastname}, [3]{ehsan.ulhaq}@kics.edu.pk

Abstract

Text-to-speech (TTS) systems for many widely spoken languages have been developed and evolved over the last few decades. Such systems are being used in many different fields. Since these TTS systems have differences in the perceived sound quality, many speech quality test methods have been proposed to compare and evaluate their performance. Test materials for these tests, however, are language specific and hence cannot be used for TTS systems developed for other languages such as Urdu. In this work, we have presented a speech quality test material specially designed for Urdu TTS systems. The proposed test is conducted using the perception of both blind and non-blind native speakers to evaluate naturalness as well as phoneme, word and sentence-level intelligibility of recently developed Urdu TTS system. Furthermore, a qualitative comparison is performed between two most popular methods for building TTS systems.

1. Introduction

Text-to-speech systems (TTS) are commonly used in everyday life, e.g., in navigation devices, public announcement systems [1] and entertainment productions [2] . It also plays a crucial role in the field of telecommunication, industrial and educational applications. TTS systems for foreign languages such as English, German and Japanese, have been developed long ago and are well established today [3]–[5]. However, research on the development of TTS system for the Urdu Language, which is a national language of Pakistan and is spoken by more than 162 million people worldwide [6], is still in its earlier stages [7]. This paper is an attempt to assess the speech quality of recently developed Urdu TTS system [8]. This effort will enhance man to machine interaction possibilities and overcome the literacy barrier for the semi-urban and rural population of Pakistan.

Speech quality is a multi-dimensional term and its evaluation contains several problems [9][10]. Speech quality of a synthesizer is determined by its similarity to the human voice (i.e., *naturalness*), its ability to be easily understood (i.e., *ineligibility*) [11] and its suitability for certain applications [10][12]. Moreover, it is reported that different applications prefer different features' evaluation. For instance, the high speaking rate with speech intelligibility features is usually preferred over naturalness in reading machines for the blind. On the other hand, in multimedia applications or electronic mail readers, prosodic features and naturalness are considered as essential features [13].

Subjective evaluation of speech synthesis is usually done by listening tests according to standards described by ITU-T Rec. P.85 [14]. Several methods have been developed during last decades for assessment of synthetic speech. However, no single evaluation provides a foolproof assessment method that focuses on both naturalness and intelligibility aspects of speech at different levels (phoneme, word, sentence or comprehension) and can provide useful and reliable information about the quality of TTS system. In addition, prior studies indicate that test materials developed for subjective evaluation of TTS need to be language specific [15]. Moreover test material should be large enough to represent a variety of language features (*representativeness*), while at the same time short enough not to distract listeners' attention (*compactness*).

In this study, we have designed both compact and representative subjective testing material for the evaluation of Urdu TTS systems. The proposed tests have been conducted on blind and non-blind Urdu native speakers and results have been reported about speech quality of Urdu TTS system. These results not only evaluate TTS speech quality but also help to figure out areas that need to be considered for further

improvements in TTS. Furthermore, this work also compares the two widely recognized approaches to build speech synthesizers, i.e., unit selection [16] and Hidden Markov Models (HMMs) [17] , with the aim to identify which one is better choice for generating Urdu synthetic speech in terms of both naturalness and intelligibility.

The remainder of this paper is divided into following sections: Section 2 briefly describes the architecture of Urdu TTS system. Section 3 explains the design of subjective quality test and testing materials selected for this purpose. The procedure and comparative results of two voice synthesis approaches are reported in Sections 4 and 5 respectively. Finally, Section 6 concludes the findings of this research.

2. Urdu TTS System Architecture

Urdu TTS system converts Urdu text into synthetic speech waveform as shown in Figure . TTS system generally consists of two main modules, Natural Language Processor (NLP) and Speech Synthesizer. NLP pre-processes the input text including abbreviations, dates, and numbers; and converts into its appropriate phonetic description annotated with prosodic and context dependent information. Speech Synthesizer then generates corresponding speech signal using the description provided by NLP. Overall speech quality of TTS system relies on both of these modules.

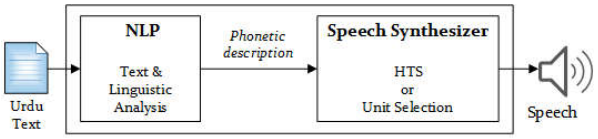


Figure 1 Architecture of TTS system.

Two different types of concatenative synthesis approaches have been used in Urdu TTS system. First, the classical unit selection (US) method that synthesizes speech by concatenating pre-recorded human speech waveforms and hence requires a large amount of speech database [4]. Second is Hidden Markov Model-based synthesis (HTS) that uses statistical models instead of actual speech units [18] and for this reason its footprint is very small (less than 10MB), compared to unit selection approach. More architectural details of Urdu TTS system are available in [18] and [19].

3. Design of Subjective Test

The design of subjective test highly depends on the application domain where TTS system is to be deployed. For example a TTS destined to provide traffic information asks for a more specific type of test materials than TTS to be used as news/screen-reader for the blind, where test materials should cover vocabulary from a wide range of topics (e.g., religion, sports, literature, health etc.) and multiple sentence structures [20]. Urdu TTS system belongs to the second type of category, and hence quality test is designed comprehensively. The test contains a total of 1010 words out of which 496 are unique. These words are taken from news, literature, and daily life conversational vocabulary. Total speaking time of the test is approximately 9 minutes and response time is around 20 minutes.

The theme of this subjective test revolves around four questions: (a) Is Urdu TTS system mature enough to deliver any type of speech content with the acceptable clarity of voice and the underlying message? (b) Is Urdu TTS' voice as pleasant as that of human beings? (c) Is Urdu TTS system suitable for both the blind and non-blind communities? (d) Which one of the two speech synthesis approaches (HTS or US) is a better choice for Urdu TTS based on the criteria set by above questions? To answer these questions, a group of subjective tests is conducted categorized under intelligibility and naturalness tests that are briefly explained below.

3.1. Intelligibility Tests

Intelligibility tests focus on the ability to identify what is spoken regardless of whether it sounds robotic or human-like, noisy or clear. Good quality in intelligibility includes an understanding of spoken utterances with correct perception at each level of speech units from phonemes to sentences [21]. Intelligibility tests designed at segmental, sentence and comprehension levels for Urdu TTS systems are discussed below.

3.1.1. Segmental Test With segmental evaluation methods intelligibility is tested at smallest speech units, like phonemes. In contrast to vowels, consonants are difficult to recognize in synthetic speech, because of sudden spectral transitions and multiple excitation signals [20] and hence test materials usually focus on consonants [13]. Moreover, syllable-initial and syllable-final consonants are perceived differently by listeners [22]. For this reason, it makes sense to break down the segmental-quality evaluation of TTS for

consonants in both initial and final positions within monosyllabic words. For this purpose, a test set is designed containing 64 pairs of confusable rhyme words. Words in a pair differ in their initial or final consonants. The consonants are equally distributed among 4 phonemic distinctive features (8 word-pairs per feature per position). Few examples are shown in **Table** , for complete dataset please refer to Tables A-1 and A-2 in Appendix A.

Table 1 Examples of segmental evaluation test.

Phonemic features	Description	Pairs with different initial consonants	Pairs with different final consonants
Voicing	voiced – unvoiced	پات، /paːʈ/ baːʈ	باپ، /baːp/ baːb/باب
Nasality	nasal - oral	مول، /mol bol/بول	تام، /ʈaːm ʈaːb/تاب
Aspiration	Aspirated – Non-Aspirated	بال، /baːl/ بھال bʰaːl/	باپ، /baːp/ باپھ/baːpʰ
Sibilation	sibilated - unsibilated	چھال، /ʧʰaːl/ کال kaːl/	ساز، /saːz/ ساتھ/saː ʈʰ

These rhyme words are tested through following carrier sentence:

(1) کیا آپ اردو لغت سے لفظ ---- کا مطلب بتا سکتے ہیں؟

kæɑ: ɑːp ʊrɖu lʊɣəʈ se ləfz ----- kɑ: məʈləb bəʈɑ: səkʈe hæ:

What- kæɑ: you- ɑːp Urdu- ʊrɖu dictionary-lʊɣəʈ case marker-se word- ləfz ---- case marker- kɑ: meaning- məʈləb tell- bəʈɑ: can-səkʈe tense aux- hæ:

"Can you inform me the meaning of --- word from the dictionary?"

First, a pair of rhymed words is visually presented to the subject. Then one word of the pair embedded in the carrier sentence is aurally presented and the listener's task is to indicate which of the two words was spoken as part of the sentence. The carrier sentence is sensitive to segmental errors in the word to be tested, as there is a lack of contextual information that can assist listeners to predict the segment not heard. Furthermore, all the cognitive information that is required for this recognition task is provided to the listener before the auditory presentation. Hence, an error in identifying the word can be regarded as a direct measure of TTS systems' inaccuracy.

This diagnostic test can highlight the misidentified phonemes and help to localize the problem points for improvements. The obtained measure of segmental intelligibility is simply the percentage of correctly identified words distributed among 4 phonemic distinctive features.

Table 2 SUS test sets.

Sr. No.	Sentences-Set1	Sentences-Set2
1	میز تیز رفتاری سے بیٹھ گیا۔ meːz ʈeːz rəfʈaːri se bæTʰ gæa:	جہاز کے پتے رونے لگے ʤəhaːz keː pəʈʈeː roːneː ləgeː
2	باغ میں کاغذ بہنے لگا۔ baːɣ mẽː kaːɣəz bæhneː ləgaː	کتاب کی آواز چہکنے لگی kiʈaːb ki aːvaːz ʧəməkneː ləgiː
3	جہاز اپنے پیروں پر لیٹ گیا۔ ʤəhaz əpne pærõː pər leːT gæa:	درخت لو کی طرح اڑنے لگے ɖərəxʈ ləhuː ki ʈarha uɽne ləgeː
4	کتاب کے پتے ٹوٹنے لگے۔ kiʈab ke pəʈʈeː TuːTneː ləgeː	ریت اپنے پیروں پر لیٹ گئی reːʈ apneː pærõː pər leːT gəiː
5	درخت پر سے سرک ٹوٹنے لگی۔ ɖərəxʈ pər se səɾək TuːTneː ləgiː	باغ کا میز پاگل ہو کر بیٹھ گیا baːɣ ka meːz paːgəl ho kər bæTʰ gæɑː
6	ہوا ٹوٹنے کی آواز چہکنے لگی həva: TʊTne ki avaz tʃəməkne ləgi	ہوا بیٹھ کر مہک گئی həvaː bæTʰ kər mʊrʤʰaː gəiː
7	درخت کر سیوں پر ناچنے لگے ɖərəxʈ kʊrsIjõː pər natʃneː ləgeː	کچے انڈے درخت پر ناچنے لگے kətʃtʃeː ənDeː ɖərəxʈ pər naːtʃneː ləgeː
8	ریت کی آواز مہک گئی reːʈ ki aːvaːz mʊrʤʰaː gəiː	خشک جوتیاں میز پر ٹوٹنے لگیں xʊʃk dʒuːʈiːjãː meːz pər TuːTneː ləgĩː
9	خشک جوتیاں لو کی طرح اڑنے لگیں xʊʃk dʒuːʈiːjãː ləhuː ki ʈarhaː uɽne ləgẽː	کچے انڈے سے سرک بہنے لگی kətʃtʃe ənDeː se səɾək bæhneː ləgiː
10	کچے انڈے پاگل ہو کر رونے لگے kətʃtʃeː ənDeː pagəl ho kər roːneː ləgeː	آواز کے بہنے سے میز لیٹ گیا aːvaːz ke bæhneː se meːz leːT gæɑː

3.1.2. Sentence level Test Segmental intelligibility at sentence level is usually evaluated through transcription task of semantically unpredictable sentences (SUS) [23][24]. SUS sentences have grammatically correct syntax, however, they are unpredictable semantically. They have no inherent meaning, therefore minimize the possibility of deriving phonetic information from textual context but the speech signal itself, e.g.,

(2) – میز تیز رفتاری سے بیٹھ گیا

mez tezrɑːftɑːri: se bæt gəɑ:

Table- mez speedily- tezrɑːftɑːri case marker-se sat: bæt tense-gəɑ

"Table sat down speedily"

SUS sentences are constructed using high-frequency words from language specific lexica. Instead

of forced-choice, subjects are asked to transcribe the sentence as they listen. This helps to avoid ceiling effect in listeners' responses. An overall percentage of correct recognition is calculated based on the percentage of correctly transcribed words per sentence. Higher the percentage more intelligible is the synthesized voice.

One inherent problem with sentence level tests is that each sentence can be presented to a subject only once during the test [21]. This fact becomes a major concern when the purpose of the test is to compare two different TTS technologies. In order to avoid learning effect, separate SUS test sets have been designed for both HTS and US voice synthesis and are shown in **Table** . For a fair comparison, the same set of vocabulary is used for both test sets.

Table 3 MOS rating scales [14]

Naturalness (Quality)	**How do you rate the quality of the sound that you just heard**? 1. Bad 2. Poor 3. Fair 4. Good 5. Excellent
Speaking rate	**What was the average speed of delivery?** 1. Much slower 2. Slower 3. Normal 4. Faster 5. Much faster
Pronunciation	**Did you notice any anomalies in pronunciation?** 1. Yes, very annoying 2. Yes, annoying Poor 3. Yes, slightly annoying 4. Yes, but not annoying 5. No.

3.1.2. Comprehension Test Intelligibility test methods discussed so far focus on the accuracy of individual sounds or words, rather than correct reception of the underlying message. For some TTS applications, such as news readers, it is not required to recognize every single phoneme, as long as the meaning of whatever is being spoken is understood [25]. In comprehension tests, synthesized speech sample containing few sentences or paragraph is presented to the subject, followed by a questionnaire about the content of the passage. Hundered percent segmental intelligibility is not needed to answer the questionnaire. Two news paragraphs from BBC Urdu website were selected for testing Urdu TTS. Topic selection was made from the category that is less likely to be familiar to most of the listeners such as latest research reports from health sciences domain.

3.2. Naturalness Test

The goal of an ideal TTS system is to mimic human speech style, so it should also be evaluated against overall speech quality parameters, such as *speaking rate*, *pronunciation,* and *naturalness*, in addition to intelligibility. Naturalness and overall quality of synthetic speech are difficult to quantify as they are abstract subjective attributes and subjects' may have different preferences for these attributes [21]. Mean opinion scoring (MOS), recommended in ITU-T Rec. P.85 [14], is a most widely used method for speech quality evaluations.

Table 4 MOS test set

Sr.	Sentences
1	اس دوران ترکی اور ایران کے مابین مجموعی تجارت کا حجم ۲۱۔۸ بلین ڈالر رہا Is ɖo:ra:n ʈʊrki: ɔr æra:n ke ma:bæn mədʒmu:i: ʈədʒarəʈ ka hʊdʒəm a:Th se Ikki:s bIljən Dɔlər rəha:
2	جہاں خوف کے بعد رش یا د واقع ہو وہ بھی واو معدولہ ہو سکتی ہے dʒəhɑ̃: xʊke: ba:ɖ rəʃ ja:ɖ vaQəja: ho: vo: bhi va:ve dʒ məro:la: ho: səkʈi: hæ.
3	اس کی تاریخ پیدائش ہے ۶/۹/۱۹۸۰ Is ki ʈari:x pe:ɖa:lʃ hæ tʃe: no ʊnni:s sɔ əssi:
4	دونوں کھلاڑیوں نے: ۲،۲ وکٹیں لیں۔ ɖo:nõ: khIlarljõ: ne: do: do: vikTẽ: lĩ:
5	ماہانہ تنخواہ ۲۰ ہزار روپے علاوہ ۱۶٪ کمیشن دی جائے گی فوری رابطہ کریں ۲۴۱۸۲۳۹۴۲۳۰ məha:na: ʈənxa: bi:s həza:r rʊpe: əla:va: ba:ra: fi:səɖ kəmi:ʃən ɖi: dʒa:e: gi: fɔ:ri: ra:bəʈa kərẽ: do: tʃʊ:r e:k a:Th do: tʃʊ:r no tʃʊ:r do: ʈi:n slfər
6	علینا نے کل ۱۹:۰۰ بجے مارکیٹ جانا ہے۔ əli:na: ne kəl ʊnni:s bəje: ma:rkIT dʒa:na: hæ
7	علم صرف میں ف، ع، ل کو حرف کی بجائے کلمہ کہا جاتا ہے۔ Ilme sərf mẽ: fe: æn la:m ko: hərf ki bədʒa:e kəlma: kəha: dʒa:ʈa hæ
8	لاہور میں فروری ۲۰۱۶ء کو خوب طوفان آیا la:hɔr mẽ: fərvəri: tʃo:ɖa sɔ ɖo: ko: xu:b ʈʊfa:n a:ja:
9	آخری وقت اشاعت: اتوار ۹ فروری ۱۷:۳۵ جی ایم ٹی ۲۲:۳۵ پی ایس ٹی ، ۲۰۱۴ a:xəri: vəQʈ əʃa:ʈ lʈva:r no: fərvəri: səʈa:ra: sɔ pæntʲi:s dʒi: æm Ti: ba:i:s sɔ pæntʲi:s pi: æs Ti: do: həza:r tʃɔɖa
10	آجکل لوگ بہت سی لوک داستانوں سے دور ہیں a:dʒ kəl bohəʈ si: lo:k ɖa:sʈa:nõ: se ɖu:r hæ:

3.2.1. MOS Test This method is a grading-based procedure, where subjects are asked to rate given speech samples by asking questions such as "How do you rate the quality of the sound that you just heard?" and responses are collected on a 5-point scale, where high score means better perceived quality. Values from 1 to 5 are presented with descriptions from "bad" to "Excellent", or similar depending on what is asked. Complete range of scales and their descriptions for the subjective attributes are presented in **Table** . The

arithmetic average of scores given by all respondents represents mean opinion score (MOS) and TTS technologies are ranked accordingly. Meaningful sentences covering a wide variety of sentence structures, e.g., sentences with definitions, date, time, contact numbers, and facts & figures are selected (**Table**).

4. Experimental Setup

Total of 23 naïve subjects (3 female, 20 male) aged between 18 and 22 participated in the testing process. Out of 23 subjects 5 were blind males. Blind's subjects' response collected and interpreted separately. All of them were native Urdu speakers. None of them suffered from any hearing problems or dyslexia. All subjects participated as volunteers. Experiments were conducted under control environment where each subject was listening synthesized voices using headphones. Urdu TTS was manually optimized for pronouns and other mispronounced technical terms. The optimization included an adjustment of wrong articulated words and an improvement of pauses between sentences and paragraphs.

4.1. Procedure

The test was composed of four major sections each corresponding to one of the four tests discussed in Sec. 3. In MOS section, each subject's screen displays one sentence at a time synthesized in both HTS and unit selection voices. Voices' identity was kept hidden from the subjects in order to avoid biases. Voices were displayed with names like voice A and voice B. Subjects were asked to listen to a sentence in both voices and rate them according to their naturalness, speaking rate, and pronunciation. Each subject was given a proper explanation of these terms and meaning of rating scale used for voices' quality.

In the comprehension section, one paragraph synthesized in each voice played one by one. After listening paragraphs, respondents were asked to answer three questions taken from the paragraph. Subjects were allowed to listen to the paragraphs again if they need. The third section contains the transcription task for SUS sentences. Fourth section consists of segmental evolution using DRT and MRT test sets, where each respondent has to pick one of the two possible options against the played voice.

5. Results and Discussion

5.1. Intelligibility

5.1.1. Segmental Test This sub-section provides summarized results of segment level tests for both blind and non-blind groups. Table 5 and 6 show that at the segmental level, all features (i.e., voicing, Nasality, Aspiration and Sibilation) are understood better at word-initial place as compared to word-final place for both voices (HTS and US). Moreover, US voice performs better than HTS voice across most of the features except voicing and aspiration. Note: The metric reported in Table 5 and 6 is average percentage of correctly identified words from the pool of word pairs discussed under the heading ***Segmental Test*** in Sec 3.

Table 5 Segmental test results (in percentage) for non-blind group

	Non-Blind			
	HTS		US	
	Word Initial	Word Final	Word Initial	Word Final
Voicing	89.6	65.3	73.5	64.6
Nasality	97.2	95.1	97.9	95.1
Aspiration	95.8	51.4	84.5	62.5
Sibilation	97.9	97.9	100	99.3

Table 6 Segmental test results (in percentage) for blind group

	Blind			
	HTS		US	
	Word Initial	Word Final	Word Initial	Word Final
Voicing	72.5	67.5	75	63.75
Nasality	90	97.5	95	95
Aspiration	77.5	42.5	82.5	52.5
Sibilation	100	85	97.5	95

5.1.2. Sentence level Test Participants were allowed to listen SUS sentences maximum of two times. However, most of them played each sentence for once only. The obtained measure of intelligibility was based on a percentage of correctly recognized words. Results for both voices (HTS and US) are summarized in the graph shown in **Figure** . According to results, intelligibility at word level is better for HTS voice as compared to US and this result is consistent among both subject groups (blind and non-blind).

5.1.3. Comprehension Test Total of three questions was asked per paragraph. Answers to the open-ended questions were scored according to a 3-point scale (0, 0.5, and 1) where 0 points are given to incorrect or unanswered responses; partially correct or too general, yet not wrong answers are given 0.5 points; and only correct and specific answers are marked with 1 point. Results are summarized in the graph shown in **Figure** . Again in this intelligibility test HTS voice's performance is slightly better than US voice.

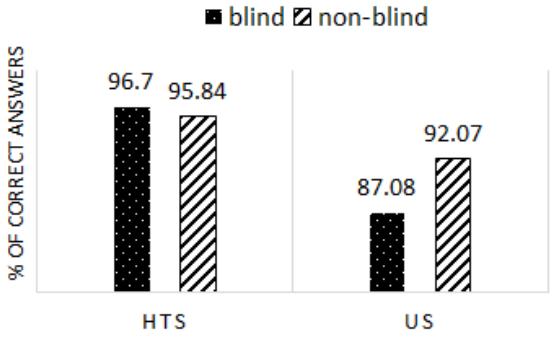


Figure 2 SUS test results.

5.2. Naturalness

For the overall quality rating, subjects were allowed to repeat sentences. Mean rating of both voices w.r.t naturalness, speaking rate and pronunciation are reported in tabular format as shown in **Table** . Entries of this table can be interpreted according to MOS rating scales described in **Table** . According to both (blind and non-blind) groups, US voice is closer to human voice as compared to HTS; US voice speaking rate is almost normal while HTS's is slightly faster than normal; and pronunciation of US is also better than that of HTS voice.

Table 7 MOS test scores.

	Naturalness		Voice Rate		Pronunciation	
	HTS	US	HTS	US	HTS	US
Non-Blind	2.89	3.11	3.28	2.81	2.94	3.32
Blind	2.78	3.22	3.49	3.08	2.94	3.54

6. Conclusion

From the results it is clear that both synthesized voices (HTS and US) are reasonably intelligible for humans and most of the respondents easily understood the synthesized sentences. Moreover, this work also pinpoints the shortcomings of Urdu TTS, e.g., from **Table** and 6 we can see that these voices are weak in modeling aspiration feature as compared to nasality feature. Improvements of these aspects will be done in future work. When it comes to overall intelligibility, i.e., how accurately message is understood, HTS synthesis approach performs better than US, the reason is when pre-recorded speech units are concatenated in US approach they get affected by sudden changes in pitch values that create distractions for listeners.

From the naturalness point of view, however, US is preferable among both types of subjects (blind and non-blind). The reason is that in US approach speech waveform is synthesized by concatenating actual human voice units while in the case of HTS it is generated through statistically trained models. Currently the speech corpus used for training is annotated at phoneme, word, syllable, stress and break index levels only and the prosodic information, which is essential for naturalness effect in synthetic speech, still has not been incorporated. In future, the prosodic structure of Urdu language for various types of sentences and role of grammatical and prosodic information in the high-quality speech synthesis should also be investigated.

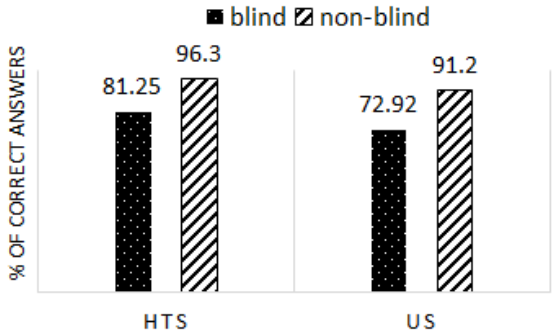


Figure 3 Comprehension test results.

7. Acknowledgment

This work has been conducted as part of the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers, supported through a research grant from ICT RnD Fund, Pakistan.

8. References

[1] S. Arndt, J.-N. Antons, R. Gupta, R. Schleicher, S. Moller, T. H. Falk, and others, "Subjective quality ratings and physiological correlates of synthesized speech," in *Quality of*

Multimedia Experience (QoMEX), 2013 Fifth International Workshop on, 2013, pp. 152–157.

[2] T. Dutoit, *An introduction to text-to-speech synthesis*, vol. 3. Springer Science & Business Media, 1997.

[3] M. W. Macon, A. Kain, A. Cronk, H. Meyer, K. Mueller, B. Saeuberlich, and A. W. Black, "Rapid prototyping of a german tts system," 1998.

[4] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, vol. 1, pp. 373–376.

[5] A. N. S and S. T, "Article: Text to Speech Synthesis of Hindi Language using Polysyllable Units," *IJCA Proc. Natl. Conf. Power Syst. Ind. Autom.*, vol. NCPSIA 201, no. 3, pp. 15–20, Dec. 2015.

[6] G. F. S. Lewis M. Paul and C. D. F. (eds.), Eds., *Ethnologue: Languages of the World*, 19th ed. Dallas, Texas: SIL International, 2016.

[7] S. Hussain, "Phonological Processing for Urdu Text to Speech System," in *Contemporary Issues in Nepalese Linguistics (eds. Yadava, Bhattarai, Lohani, Prasain and Parajuli)*, 2005, vol. ISBN 99946.

[8] "Online Urdu TTS." 2016.

[9] U. Jekosch, "Speech quality assessment and evaluation," in *Third European Conference on Speech Communication and Technology*, 1993.

[10] A. Mariniak, "A global framework for the assessment of synthetic speech without subjects," in *Third European Conference on Speech Communication and Technology*, 1993.

[11] S. Suryawanshi, R. Itkarkar, and D. Mane, "High quality text to speech synthesizer using phonetic integration," *Int. J. Adv. Res. Electron. Commun. Eng.*, vol. 3, no. 2, pp. 77–82, 2014.

[12] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[13] S. Lemmetty, "Review of Speech Synthesis Technology." 1999.

[14] I. Rec, "P. 85. A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices," *Int. Telecommun. Union, Geneva*, 1994.

[15] I. McLoughlin, "Subjective intelligibility testing of Chinese speech," *Audio, Speech, Lang. Process. IEEE Trans.*, vol. 16, no. 1, pp. 23–33, 2008.

[16] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Commun.*, vol. 49, no. 4, pp. 317–330, 2007.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0.," in *SSW*, 2007, pp. 294–299.

[18] Nawaz Omer; Habib Tania, "Hidden Markov Model (HMM) based Speech Synthesis for Urdu Language," in *the Proceedings of Conference on Language and Technology 2014 (CLT14), Karachi, Pakistan*, 2014.

[19] F. Adeeba, S. Hussain, T. Habib, Ehsan-Ul-Haq, and K. S. Shahid, "Comparison of Urdu Text to Speech Synthesis using Unit Selection and HMM based Techniques," in *the Proceedings of Oriental COCOSDA Conference 2016, Bali, Indonesia (accepted)*.

[20] V. J. van Heuven, R. van Bezooijen, and others, "Quality evaluation of synthesized speech," *Speech coding Synth.*, vol. 21, pp. 707–738, 1995.

[21] T. Ojala, "Auditory quality evaluation of present Finnish text-to-speech systems," Ph.D. thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, 2006.

[22] M. A. Redford and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in CVC syllables," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1555–1565, 1999.

[23] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener," *Speech Commun.*, vol. 16, no. 3, pp. 225–244, 1995.

[24] L. C. W. Pols and others, "Multi-lingual synthesis evaluation methods," 1992.

[25] Y.-Y. Chang, "Evaluation of TTS systems in intelligibility and comprehension tasks," in *Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing*, 2011, pp. 64–78.

Appendix A

Table A - 1 Phonetic characteristics at word initial and final

Phonemic features	Description	consonant pairs to be tested	Pairs with different initial consonants	Pairs with different final consonants
Voicing	voiced - unvoiced	/p/-/b/	پات/pɑːʈ/ ، بات baːʈ/	باپ/baːp/ ، باب/baːb/
			پول/poːl/ ،بول/boːl/	آپ/ɑːp/ ،آب/ɑːb/
		/ʈ/-/ɖ/	تار/ʈaːr/ - دار/ɖaːr/	رت/roʈ/ ، رد/roɖ/
			تاب /ʈaːb/ ،داب/ ɖaːb	داد/ɖaːɖ/ ،دات/ɖaːʈ/
		/t/-/d/	ٹاٹ/taːt/ ، ڈاٹ daːt/	چاٹ/ʧɑːt/ ، چاڈ/ʧɑːd/
			ٹال/taːl/ - ڈال/daːl/	روٹ/rot/ ،روڈ/rod/
		/ʧ/-/ʤ/	چار/ʧaːr/ ، جار/ʤaːr/	سیج/seʤ/ ، سیچ/seʧ/
			چاٹ/ʧaːt/ ، جاٹ ʤaːt/	بیج/beʤ/ ،بیچ/beʧ/
Nasality	nasal - oral	/m/-/b/	مل/mɪl/ ،بل/bɪl/	تام/ʈaːm/ ،تاب/ʈaːb/
			مول/mol/ ،بول/bol/	آم/ɑːm/ - آب/ɑːb/
		/m/-/p/	مان/maːn/ ،پان paːn/	کم/kɪm/ ،کپ/kɪp/
			پول/pol/ ،مول/mol/	آم/ɑːm/ - آپ/ɑːp/
		/n/-/ʈ/	تان/ʈaːn/ ،نان/naːn/	رات/raːʈ/ ،ران/raːn/
			تال/ʈal/ ،نال/nal/	پات/paʈ/ ،پان/pan/
		/n/-/ɖ/	دام/ɖaːm/ ،نام/naːm/	باد/baːɖ/ ،بان/baːn/
			دال/ɖaːl/ ،نال/naːl/	بد/bɪɖ/ ،بن/bɪn/

Table A - 2 Phonetic characteristics at word initial and final

Phonemic features	Description	consonant pairs to be tested	Pairs with different initial consonants	Pairs with different final consonants
Aspiration	Aspirated - Non-Aspirated	/p/-/ pʰ/	پٹ/pəʈ/ ، پھٹ/pʰəʈ/	باپ/baːp/ ، باپھ/baːpʰ/
			پل/pəl/، پھل/ pʰəl/	-
		/b/-/ bʰ/	بال/baːl/ ، بھال bʰaːl/	لوب/lob/ ،لوبھ/lo bʰ/
			بات/baːt/ ، بھات bʰ	گاب/gaːb/ ،گابھ/gaːbʰ/
		/ ʧ/-/ ʧʰ /	چاپ/ʧaːp/ ، چھاپ ʧʰaːp/	مچ/moʧ/ ، مچھ/moʧʰ/
			چیڑ/ ʧiɽ / ، چھیڑ/ ʧʰiɽ	پاچ/paːʧ/ ،پاچھ/paːʧʰ/
		/ ɖ/-/ ɖʰ /	دانی/ɖaːniː/ ،دھانی ɖʰaːniː/	سگند/sɪƊɖ/ ، سگندھ/sɪƊɖʰ/
			دات/ɖaːɖ/ ، دھات ɖʰaːɖ/	گد/gɪɖ/،گدھ/gɪɖʰ/
Sibilation	sibilated - unsibilated	/ ʧʰ/-/K/	چال/ʧaːl/ ،کال kaːl/	بچ/bɪʧ/ ،بچھ/bɪʧʰ/
			چان/ʧaːn/ ، کان kaːn/	باک/baːk/ ،باچھ/baːʧʰ/
		/Z/ - / ʐʰ /	زال/zaːl/، تھال/ʈʰ aːl/	ساز/saːz/ ، ساتھ/saːʈʰ/
			زم/zəm/، تھم/ʈʰəm/	رز/rəz/،رتھ/rəʈʰ/
		/S/-/ ʂ/	ساس/saːs/،طاس/saːʈ/	سات/saːs/، ساتھ/saːʈʰ/
			سان/saːn/،تان/ʈaːn/	راس/raːs/،رات/raːʈ/
		/ʃ/-/k/	شام/kaːm/ ، ʃaːm/کام	نوش/noʃ/،نوک/nok/
			شان/kaːn/،کان/ʃaːn	شاک/ʃaːk/ ، شاش/ʃaːʃ/

Development of Sindhi Lexical Functional Grammar

Mutee U Rahman and Hameedullah Kazi
Department of Computer Science, Isra University, Hyderabad, Sindh 71000, Pakistan
muteeurahman@gmail.com , hkazi@isra.edu.pk

Abstract

This paper presents an on-going work on Sindhi morphology and grammar development. An LFG (Lexical Functional Grammar) model for Sindhi is developed where morphological constructions are modeled in Xerox Lexicon Compiler (LEXC), and syntactic constructions are modeled in LFG by using Xerox Linguistic Environment (XLE). Development of various grammatical constructions of Sindhi is discussed. Morphological constructions considered for development include: Nouns, Pronouns, Adjectives, Verbs, Adverbs, Postpositions and pronominal suffixation of verbs. Syntactic constructions include noun phrase, verbal complex, verbal subcategorization, adjuncts, coordination, and subordination. While developing morphology and syntax of Sindhi, Tense, Aspect, Mood and Agreement are also considered wherever applicable.

1. Introduction

The paper presents Sindhi computational grammar development project in which finite state morphology and LFG (Lexical Functional Grammar) frameworks are used to implement Sindhi morphology and syntax respectively. Finite state morphology is implemented in XFST (Xerox Finite State Technology) tools [1] and Syntax is implemented in XLE (Xerox Linguistic Environment [2]. Sindhi is a resource poor language in Computational Linguistics and Natural Language Processing domains. Neither Sindhi Morphology nor the Syntax is studied by researchers with computational linguistics perspective. Sindhi has rich inflectional and derivational morphology. Nouns adjectives and pronouns have number gender and case inflections [3]. Verb morphology includes number, gender, tense, aspect and mood inflections. Sindhi syntax features include free constituent ordering, agreement, complex noun phrase constructions, coordination, subordination, syntactic case formations, and pro-drop. LFG rules defined for Sindhi syntax quite reasonably handle these syntactic constructions. Following sections give an overview of finite state morphology and lexical functional grammar frameworks.

1.1. Finite State Morphology

Finite state transducers (FSTs) play an important role in language processing applications [4] and computational studies of morphologically complex languages. Efficient morphological parsers can be implemented by combining these FSTs and computational lexicon (repository of words). FSTs convert/translate lexical level constructs to surface level words by applying morphotactics (morpheme ordering rules. Their reversible nature makes reverse conversion/translation possible. This two level (lexical and surface) morphology plays important role in implementation of morphological analyzers for natural languages [17]. Figure-1 shows the process of two level (lexicon and surface) morphology modeling using FSTs. A sample orthography FST rule is given in Table-1. This rule is used by FSTs to convert intermediate level word into surface word. Finite state morphological models based on these FSTs are well known models and successfully been used for morphological modeling of many languages. They handle concatenative and non-concatenative morphology very well [5].

1.2. Lexical Functional Grammar

Lexical Functional Grammar (LFG) is a natural language syntax representation formalism based on generative grammars [6] [18]. LFG defines the structure of language and relationship among different aspects of linguistic structure. Various relations are de-fined at lexicon level as LFG has a rich lexical structure. LFG represents linguistic structure at different levels which include lexicon, constituency structure (c-structure) and functional structure (f-structure) levels.

Table 1. A sample orthograpghy rule

Singular	Intermediate	Plural	Rule
Mango	Mangos	Mangoes	ε→e /^____s#

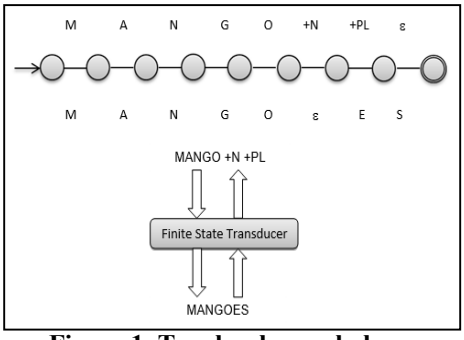


Figure 1. Two level morphology

The lexicon contains list of words or parts of words (smaller meaning bearing units) along-with information about these words including their distribution or syntax and morphology. Thus a lexical entry in LFG may include part of speech, number, gender, case, and argument structure in case of verbs and some postpositions and adjectives. Syntactic structure information in LFG is represented at two different levels. C-structure representation handles word or phrase grouping and their precedence in a phrase structure tree along-with some grouping and order constraints. F-structure represents more abstract relations between different functional constructs like subject, object, secondary object, oblique, complement, and open complement.

Subsequent sections discuss related work, implementation of finite state morphology and lexical functional grammar for Sindhi with nominal and verbal elements, pronominal suffixes followed by conclusion and future work.

2. Related Work

Apart from (Rahman and Bhatti) [7] one cannot find any work in finite state morphology and LFG frameworks for Sindhi morphology and grammar developments. In this work Sindhi noun morphology is discussed and few basic FSTs are presented. However, Sindhi syntax representation efforts in Context Free Grammars and Linear Specification Language can be found in (Rahman and Shah) [8] and (Rahman, Shah and Memon) [9]. First study tries to represent selected Sindhi sentence structures in CFG rules which have over generation problems. Second study tries to cope with over generation by using LSL (Linear Specification Language) but again lacks the agreement problem solution and feature representations. The only comparatively comprehensive research study available but not yet published is "Implementing GF Resource Grammar for Sindhi" [10]. The study tries to investigate the Sindhi morphology and syntax from computational perspective in grammatical framework [11]. However, the study does not cover the most of the parts of Sindhi morphology and syntax. Neither the morphological analyzer nor the syntax analyzer is proposed or designed; only the resource grammar library is made available as a shared resource.

Among south Asian languages Urdu is extensively studied with LFG perspective [12]. Urdu became part of parallel grammar project (Butt Helge and King) [13] and was analyzed with large scale grammar development perspective. It was found that basic analysis decisions made for European languages are applicable to typologically different language Urdu. In Pargram project parallel computational grammar of different languages is being developed within LFG framework. Various research articles discussing different syntactic issues in Urdu LFG including complex predicates, clitics, argument structure, argument srambling in noun phrases and verb phrases can be found on official website of Urdu Pargram [14]. Jafar Rizvi in his PhD thesis [15] also presented Urdu syntax analysis in LFG.

3. Implementation of Finite State Morphology and Lexical Functional Grammar for Sindhi

Overall implementation model is shown in Figure-2. Survey of Sindhi language and linguistics provides foundations of the work. Based on these foundations Sindhi grammar is analyzed and studied with LFG perspective and Sindhi morphological constructions are studied with finite state morphology perspective.

Later Xerox Finite State Tools Lexicon Compiler and Xerox Linguistic Environment are used to develop Sindhi morphology and Syntax respectively. Different components are interfaced with each other in XLE to parse and analyze Sindhi sentences. Apart from finite state morphology full form lexicon for postpositions is also developed in LFG. As a result, parse tree and functional structure analysis are generated.

Following sections discuss morphology and syntax implementation details.

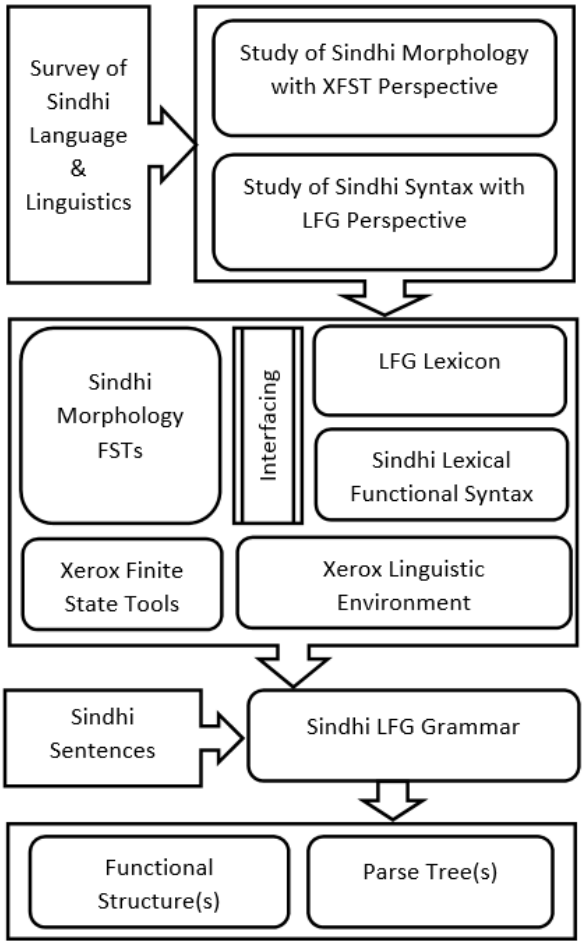


Figure 2. Sindhi grammar development model

3.1. Implementing Morphology

Inflectional morphology of various word classes of Sindhi is implemented by incorporating the inflection rules in finite state models using Xerox LEXC (Lexicon Compiler) [16]. Different morphological paradigms of nouns, pronouns, adjectives, adverbs and verbs are represented in finite state transducers in LEXC syntax. These scripts are compiled to generate finite state machines which represent Sindhi lexicon.

Sindhi nouns are inflected by number gender and case. Different paradigms are modeled in FST rules and resulting transducers act as function machines in which either upper side represents the input and lower side represents the output or vice versa. The reversible property of these FSTs makes them very useful. When lower side becomes input these FSTs function as morphological analyzers and when upper side is input these will function as surface form generators.

While defining finite state lexicon for Sindhi nouns in LEXC, a Root lexicon named Nouns is defined which is further extended to various sub-lexicons. These sub-lexicons actually model various inflectional paradigms of nouns. LEXC script fragment defining Sindhi noun morphology is shown in figure 3.

```
!SINDHI NOUN MORPHOLOGY
Multichar_Symbols
+Noun +Adjective +Adverb +Verb
+Common +Proper +Abstract !Noun Types
+Animate +Inanimate !Noun Concept
+Accusative +Dative +Ergative
+Genitive +Instrumental + Locative
+Nominative +Oblique +Vocative !Noun
Cases
+Count +Mass +Gerund +Measure +City
+Country +FirstName +LastName
+FullName +Name
+Fem +Masc !Gender
+Sg +Pl !Number
+1st +2nd +3rd !Person

LEXICON Root
Nouns;

LEXICON Nouns
       !Boy (Animate Common Noun)
CHOkirO+Noun+Common+Count+Animate:CHO
kir N_Cat1;
       ...
LEXICON N_Cat1
       +Sg+Masc+Nominative:O      #;
       +Sg+Masc+Oblique:E         #;
       +Sg+Masc+Vocative:A        #;
       +Sg+Fem+Nominative:Ia      #;
       +Sg+Fem+Vocative:I         #;
       +Sg+Fem+Oblique:I          #;
       +Pl+Masc+Nominative:A      #;
       +Pl+Masc+Oblique:ani       #;
       +Pl+Masc+Vocative:aO       #;
       +Pl+Fem+Nominative:yUN     #;
       +Pl+Fem+Oblique:yani       #;
       +Pl+Fem+Vocative:yUN       #;
```

Figure 3. LEXC fragment of Sindhi noun morphology

It can be seen that script starts with multi character symbol declarations which are used to define morphological tags. Stem forms of noun are placed in root lexicon (Nouns in this case) followed by sequence of tags representing different features of noun. Stem along-with these features will produce intermediate word form shown after colon (:) following the tag

sequence. This intermediate form is further inflected based on various feature sequences defined in sub-lexicon N_Cat1. For example, consider the stem form and tag sequence given below:

```
CHOkir+Noun+Common+Count+Animate
```

This will produce intermediate animate common count noun form “CHOkir”, this transducer is followed by another transducer in series (via N_Cat1 sub-lexicon link) which takes further input tags as shown below:

```
+Sg+Masc+Nominative
```

This tag sequence produces the singular masculine nominative morpheme “O”. The overall concatenated tag sequence preceded by stem (upper side) and concatenated output (lower side) are given below:

```
Upper:        CHOkir+Noun+Common+Count+
              Animate+Sg+Masc+Nominative
Intermediate: CHOkir          O
Lower:        CHOkirO
```

While going from upper to lower side surface form “CHOkirO” of stem “CHOkir” with features specified in tag sequence is generated; going from lower to upper will give following morphological analysis of noun “CHOkirO”.

```
CHOkir {"+Noun" "+Common" "+Count" "+Animate" "+Sg" "+Masc" "+Nominative"}
```

Above morphological analysis says that “CHOkirO” is a morphological form of stem “CHOkir” which is a common animate count noun in singular masculine form with nominative case. In the same way oblique morphological form (used as base for various syntactic cases of nouns) “CHOkirE” is generated by producing and concatenating the oblique morpheme “E” by input tag sequence given below and output sequence “CHOkir” and “E”.

```
CHOkir+Noun+Common+Count+Animate+Sg+Masc+Oblique
```

Total twelve (12) different inflections of stem “CHOkir” are taken care of. A total of 21 different common noun categories are identified according to their inflectional properties. For every category a different sub-lexicon is defined. Usually proper nouns are not inflected therefore their entries only contain the feature tags. For example, the proper noun Pakistan has following entry in the lexicon.

```
Pakistan+Noun+Proper+Inanimate+Country+Masc+Sg+3rd:pAkistAn#;
```

It says that Pakistan is an inanimate masculine singular proper noun which is a country and has surface form “pAkistAn”. Most of the proper nouns have this type of entry. However, in Sindhi there are exceptional cases of proper noun inflections. For example, a person name “dOdO” can have number, and case inflections “dOdA” (plural or singular vocative) and “dOdE” (oblique form). A sub-lexicon is defined to handle these inflections.

Verb in Sindhi is a morphologically complex word class. Verbs are marked by number, gender, case, tense, aspect and mood. Various categories of auxiliary verbs are also inflected by number, gender, and case; auxiliaries may also be used as tense and aspect markers with inflections. Copula verbs also undergo morphological changes. Due to many different categories of verbs reasonably good number of tags is used while implementing verb morphology. Verb lexicon covers auxiliary verb, copula verb and main verb morphology. Morphological analyses show that a verb in Sindhi can have up to 75 different morphological forms. Implementation strategy of verb morphology is identical to noun morphology discussed above. Pronoun, Adjective, and adverb morphology is also modeled on same lines.

3.2. Implementing Syntax

Different syntactic constructions of Sindhi are implemented in XLE by defining Sindhi LFG rules. Morphology defined in LEXC scripts is compiled to finite state transducers (discussed in previous section) and integrated to LFG grammar via morphology syntax interface in XLE environment.

3.2.1. Nominal Elements. Nominal elements include nouns, pronouns, adjectives, adverbs and phrases constituted by these elements. Different NP constructions implemented include: pronoun-noun, adjective-noun, and pronoun-adjective-noun combinations. These noun phrase combinations are further complicated by coordination, postpositional phrases and relative clauses. F-structure analysis of a noun phrase with demonstrative pronoun-noun combination is shown in Figure 4. Demonstrative pronoun “ihO” (this) is treated as a determiner in noun SPEC (specification). Different cases of nominal elements including nominative, accusative, dative, ablative, locative, instrumental, participant, genitive/possessive, agentive and vocative are taken care of. Different complications of syntactic case

marking are handled by defining a special case phrase KP [14] which represents case marked noun phrase constructions.

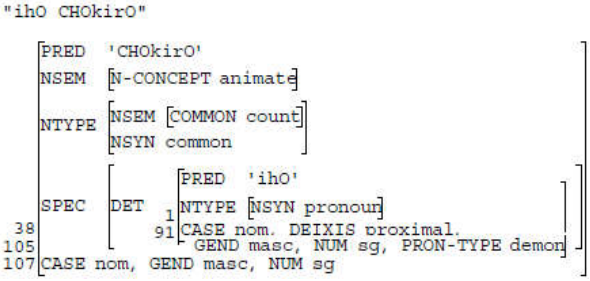


Figure 4. Noun phrase with demonstrative pronoun.

Figure 5 shows an example of a case phrase with dative and accusative case marking. F-structure chart shows two possibilities of case "dat" and "acc"; the proper noun "Ali" therefore can either be in dative or accusative case as "khE" is case marker for both these cases. However, ambiguity of case of "Ali" will be resolved when other syntactic elements in the sentence are present and depending on whether "Ali" is direct object or indirect object or sometimes a dative subject.

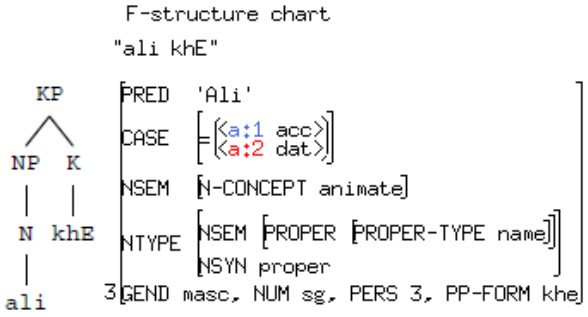


Figure 5. Dative and accusative case marking of noun.

For genitive case separate phrase KPPoss (possessive case phrase) is defined which reflects special agreement features required for agreement by different constituents of a sentence. LFG definition of KPPoss in XLE format is given below:

```
KPPoss -->    NP: {(! N-FORM)=c obl |
                   (! NTYPE NSYN)= proper}
                   ^=!;
              KPoss: ^=!.
```

LFG lexicon entry of KPoss (possessive case marker) "jO" showing extra attributes is given below.

```
jO        KPoss * (^ PP-FORM)=of
                  (^ K-NUM)=sg
                  (^ K-GEND)=masc
                  (^ K-FORM)=nom
                  (^ CASE)=gen.
```

7. It may be noted that extra attributes K-NUM, K-GEND, and K-FORM (K represents case) are introduced here to reflect the possessive case marker attributes to be agreed with possessed noun attributes. An example of KPPoss with genitive noun marking is shown in figure 6.

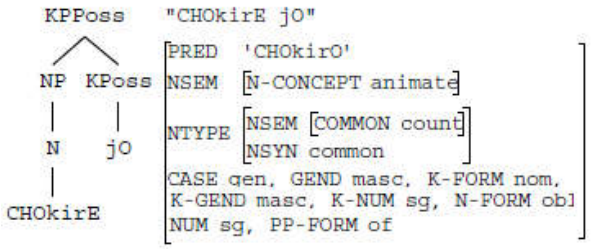


Figure 6. Genitive case marking with possessive case phrase.

3.2.2. Verbal Elements. Verbal elements include verbs which subcategorize (require arguments) for different grammatical functions. These grammatical functions include subject (SUBJ), object (OBJ), secondary object (OBJ2), oblique (OBL), PREDLINK, complement (COMP) and open complement XCOMP. Noun phrases (including all nominal elements) either define these functions or play essential role in their definition within a sentence. Sentence constituents therefore include verbs, their arguments and adjunct (ADJUNCT) elements which do not subcategorize for verbs. Different Verb categories include predicative verbs (main verbs and copula verbs), modal verbs and auxiliary verbs. In Sindhi main, auxiliary and modal verbs are combined to make verbal complex. Auxiliaries are also used to mark tense, aspect and mood. LFG implementation of verbal syntax includes verbal subcategorization for different grammatical functions listed above, verbal complex, and tense-aspect-mood analysis. Tense coverage include aorist formations, present, past and future tenses. Aspectual formations including perfective, imperfective-habitual and imperfective-continuous are analyzed by implemented LFG rules. Verb mood is also analyzed, coverage of different mood constructions includes: subjunctive, presumptive, imperative, declarative or indicative, permissive, prohibitive, capacitive, suggestive, and compulsive moods. A short version of sentence definition in LFG format is given below:

```
S--> NP:(^   SUBJ)=!   (!   GEND)=(^
GEND);)
       (KP: (^ OBJ2)=! (! CASE)=c dat)
```

```
(KP: (^ OBL)=! {(! CASE)=c inst
| (! CASE)=c agent})
(KP: (^ OBJ)=! {(! CASE)=c acc
| (! CASE)=c nom})
VC: (! NUM)=(^NUM) (! GEND)=(^
GEND) ^=!.
```

Above rules define sentence S as a sequence of noun phrase (NP) which is a subject, followed by optional case phrases (KPs) which include indirect object (OBJ2), oblique (OBL) and direct object (OBJ) followed by verb complex which may include combinations of different verb types. Above given rule defines the general structure of Sindhi sentence. Different constraints like (! GEND) = (^ GEND) and (! CASE=c dat) are placed to ensure gender case and number agreement. Consider following present tense sentence where subject and object are in nominative case.

Ali	KHatu	likhE	thO
Ali.Nom.M	Nom.M.Sg	Write.Aorist	Aux.Pres
Ali	Letter	Write	Be

Ali writes a letter.

Parse tree and functional structure analysis of above sentence are shown in figure 7 and figure 8 respectively. Subject and object case is identified morphologically, tense form in combination with present tense auxiliary "thO" identifies the tense, and aspect is also identified by morphological form of main verb "likhE" which is neither progressive nor perfective. Aspect is undefined therefore.

Consider following sentence:

Ali	CHOkirE-khE	KHatu
Ali.Nom.M	boy.Obl.Sg.M-Dat	letter.Nom.M.Sg

likhE	payO
write.Aorist.Sg	Aux.Cont

Ali is writing a letter to the boy.

It can be seen in above sentence that there are three verbal arguments, a subject "Ali", an indirect object "CHOkirO" in oblique form with dative case marker, and a direct object "KHatu" with nominative case. Verb aspect is continuous / progressive identified by "payO" auxiliary used as a continuity marker. Main verb "likhE" is in aorist form. Figure 9 and figure 10 show parse tree and f-structure analysis of above sentence respectively. Indirect object is subcategorized as OBJ2 with dative case. Auxiliaries "thO" and "payO" also define the indicative mood.

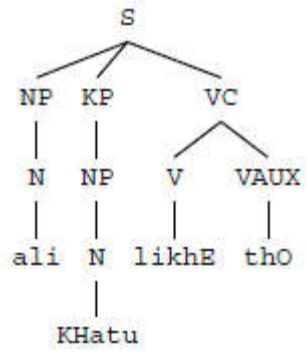


Figure 7. Sample present tense sentence with unspecified aspect.

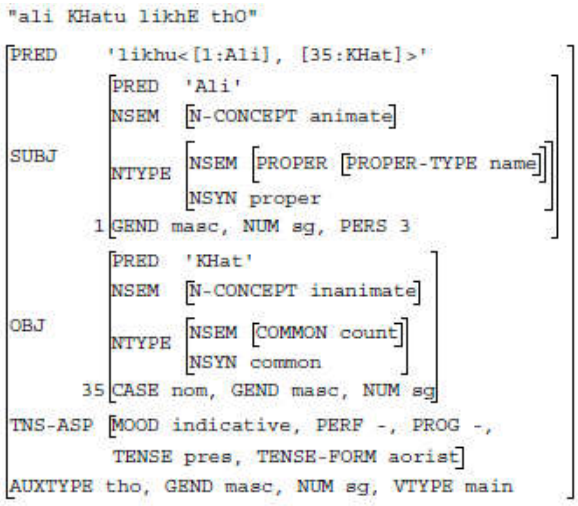


Figure 8. LFG analysis of sample sentence with SUBJ, OBJ subcategorization in present tense with undefined aspect.

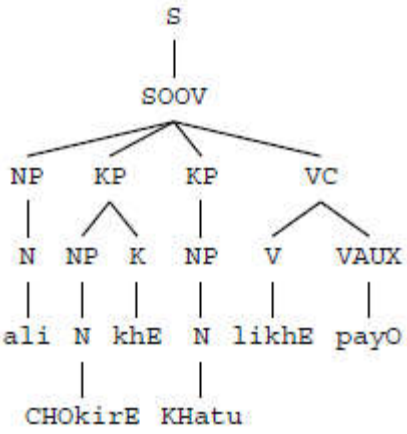


Figure 9. Sample sentence with imperfective continuous aspect.

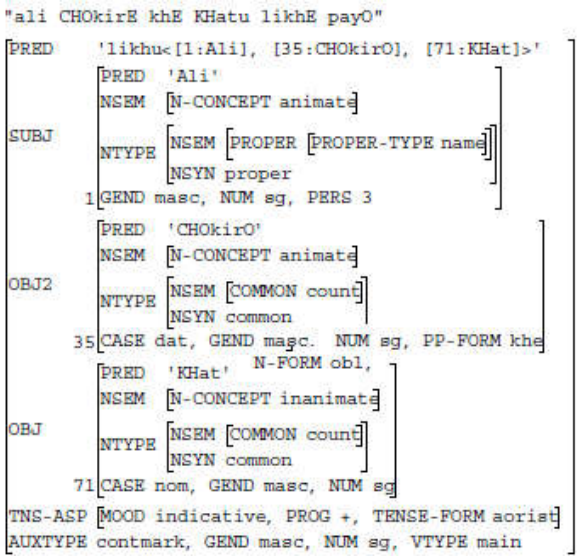


Figure 10: LFG analysis of sample sentence with SUBJ, OBJ, and OBJ2 subcategorization in aorist tense form with imperfective continuous aspect.

4. Pronominal Suffixes

Sindhi pronominal suffixes may appear with nouns, verbs, postpositions, and adverbs of place. Pronominal suffixes are treated as special lexical entries in lexicon. For example, consider transitive verb "likhu" (write); when appears with 1st person pronominal suffix "-iyami" becomes "likh-iyami" (I wrote). Morphological analysis of "likhiyami" is given below:

```
{likhiyami "+Token" | likhu "+Verb" "+Psx" "+SSg" "+S1P" "+SMF" "+SObl" "+Sg" "+PastPart"}
```

Above morphological analysis says that "likhiyami" is a morphological form of root "likhu". +Psx attribute says that this is a pronominal suffixed form. The tag pattern "+Sxxx" represent different attributes of subject reflected by pronominal suffix. +PastPart tag says that verb form is a past participle. F-structure analysis of "likhiyami" is shown in figure 11. It can be seen that different attributes of verb "likhu" in f-structure are extracted from morphological tags given in above morphological analysis. Pronominal suffixation may cause a complete sentence replaced with single word form with all its verbal and nominal elements. Syntax analysis therefore needs to extract / reconstruct this information from morphology. In this case the sentence "mUN likhiyO" (I wrote) is replaced by "likhiyami". This reconstruction can be seen in verbal subcategorization of "likhu" where the SUBJ argument contains the value 'pro' which represents a pronoun with gender, noun form, number and person attributes (feminine, oblique, singular, 1st person). Oblique singular 1st person pronoun in Sindhi is "mUN" which can either be feminine or masculine. As the verb is in past participle form therefore its aspect is perfective.

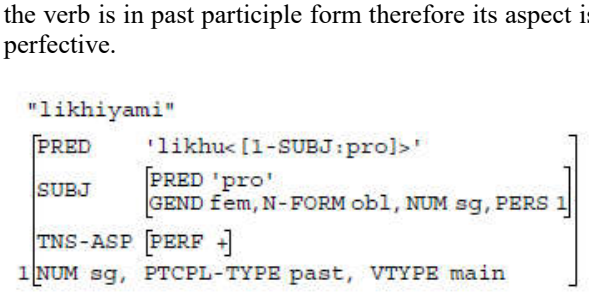


Figure 11: F-structure of pronominal suffixed verb "likhiyami".

5. Coverage

Morphological coverage includes: finite state models of nouns, pronouns, adjectives, adverbs and verbs. Full form LFG lexicon of postpositions, conjunctions and few adverbs. Case, mood, tense and aspect morphology of nominal and verbal elements is also implemented. Table 2 shows some figures about morphology coverage. Interestingly adjectives have more average inflections per stem as compared to nouns. This is due to degree change inflections of native Sindhi adjectives where inflections are doubled as compared to nouns. For example, a masculine noun with 'O' ending can have up to 12 inflections and an adjective with 'O' ending will also have almost 12 inflections. However, with internal morphological change when degree changing morphology is applied number of inflections becomes double. For example, adjective 'naNdHO' (small) becomes 'naNdHaRO' all inflections of naNdHO will also be applied to naNdHaRO as well and this will double the number of inflectional forms. Pronoun inflections per stem is also 3.58 due to number gender and case inflections (mostly in wh-pronons).

Syntax coverage include noun phrase constructions with all nominal elements, verbal subcategorization with SUB, OBJ, OBL, OBJ2, COM, XCOMP, ADJUNCT, XADJUNCT, and PREDLINK, coordination, subordination, mood, case, aspect, tense, and agreement.

5. Conclusion and Future Work

Development in current state covers the morphological and syntactic constructions discussed in above sections. Basic morphology and syntax

constructs in Sindhi are identified and modeled. Morphological analysis shows interesting results like adjectives have more average inflections than nouns, and pronouns have 3.58 average inflections per word. Also verb can have up to 75 different morphological forms. Though the basic constructs of Sindhi morphology and Syntax are implemented yet many complexities are subject to further research and development including: pronominal suffixation with nominal elements, pronominal suffixation with postpositions, NP coordination model, verbal complex constructions which form complex predicates, and pro-drop. Also the morphological lexicon size and coverage requires more enhancements. Developed model will be tested against synthesized test-suit covering all morphology and syntax patterns implemented and real time corpus test suit being developed.

Table 2: Morphology coverage

Word Class	Stems	Morphological Forms / Inflections	Average Inflections / Stem
Verbs	100	5013	50.13
Nouns	323	1729	5.35
Pronouns	79	283	3.58
Adjectives	71	394	5.55
Adverbs	38	38	1.00
Total	**611**	**7457**	**12.20**

6. References

[1] K. R. Beesley, and K. Lauri. "Finite-state morphology: Xerox tools and techniques." CSLI, Stanford (2003).

[2] C. Dick, M. Dalrymple, R. Kaplan, T. H. King, John Maxwell, and Paula Newman. "XLE documentation." Palo Alto Research Center (2008).

[3] M. U. Rahman. "Sindhi Morphology and Noun Inflections", in proc. *Conference on Language & Technology (CLT-09)*, pp. 74-81. 2009.

[4] R. Emmanuel, and Y. Schabes. Finite-state language processing. MIT press, 1997.

[5] K. R. Beesly. "Arabic Morphology Using Only FinateState Operations", in proc. *Workshop on Computational Approaches to Semetic languages*, Montreal, Quebec, pp. 50-57. (1998).

[6] M. J. Steedman. A Generative Grammar for Jazz Chord Sequences. Music Perception 2 (1): 52–77. JSTOR 40285282. (1989).

[7] M. U. Rahman., and M. I. Bhatti. "Finite State Morphology and Sindhi Noun Inflections.", in proc. *Pacific Asia Conference on Language, Information and Computation (PACLIC 24)*. Sendai, Japan. pp. 669 – 676 (2010).

[8] M. U. Rahman, A. Shah "Grammar Checking Model for Local Languages.", in proc. *SCONEST (Student Conference on Engineering Sciences and Technology)* Karachi. (2003).

[9] M U. Rahman, A. Shah, R.A. Memon. Partial Word Order Syntax of Urdu/Sindhi and Linear Specification Language. *Journal of Independent Studies and Research (JISR) Volume 5*, Number2, July 2007. pp. 13 – 18.

[10] J. D. Oad. *Implementing GF Resource Grammar for Sindhi*. Unpublished Master's Thesis. Department of Applied Information Technology Chalmers University of Technology Gothenburg, Sweden. (2012).

[11] A. Ranta. "Grammatical Framework: A Type-Theoretical Grammar Formalism", *Journal of Functional Programming* 14 (2): 145–189. (2004).

[12] M. Butt, *The Structure of Complex Predicates in Urdu,* CSLI Publications, Stanford. (1995).

[13] M. Butt, D. Helge, T. H. King, H. Masuichi, and C. Rohrer. "The parallel grammar project." in proc. "*Workshop on Grammar engineering and evaluation*" Volume 15, pp. 1-7. Association for Computational Linguistics, 2002.

[14] T. Bögel,, M. Butt, A. Hautli, and S. Sulger. "Urdu and the modular architecture of ParGram", in proc. *Conference on Language and Technology*, vol. 70. Lahore. 2009.

[15] S. M. J. Rizvi. *Development of Alorithms and Computational Grammar for Urdu*, PhD thesis. ept. of Computer and Information Science PIAS. Islamabad. 2007

[16] L. Karttunen, *Finite-State Lexicon Compiler*. Technical Report, ISTL-NLTT2993-04-02, Xerox Palo Alto Research Center, Palo Alto, California (1993)

[17] L. Karttunen, and K. R. Beesley. "Twenty-five years of finite-state morphology." *Inquiries Into Words*, a Festschrift for Kimmo Koskenniemi on his 60th Birthday (2005): 71-83.

[18] M. Dalrymple. *Lexical-Functional Grammar*. John Wiley & Sons, Ltd, 2001.

A Comprehensive Image Dataset of Urdu Nastalique Document Images

*Qurat-ul-Ain Akram Aneeta Niazi Farah Adeeba Saba Urooj

Sarmad Hussain Sana Shams

Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
**ainie.akram@kics.edu.pk*

Abstract

Reporting the standard image dataset along with ground truth information has become important in pattern recognition and Optical Character Recognition (OCR) research. Nastalique writing style is mostly used to write Urdu books, magazines and newspapers. In this paper, a large image dataset of Urdu document images written using Nastalique writing style has been reported. This data has been collected to cover the range of font sizes from 14 to 40. The ground truth typed corpus has been developed along image corpus. A total of 2,912 document images are scanned from 413 books, among them 593, 595, 150, 149, 151 ,461, 202, 186, 226 and 199 images are scanned for 14, 16, 18, 20, 22, 24, 28, 32, 36 and 40 font sizes respectively.

Keywords— *Urdu Image Dataset, Urdu Document Images, OCR, Noori Nastalique, Ligature, Main body, Diacritic*

Introduction

The research on the development of Optical Character Recognition (OCR) has long history. OCRs for printed and handwritten text of different languages including English, Russian, Chinese, Devanagari, Urdu and Arabic have been developed [1-4]. The accuracy of the OCRs depends on thorough analysis of the variations of the document images and effectiveness of the developed techniques on large dataset. In addition, a large amount of standard data is also required covering all real life varieties of document images, to evaluate and compare different techniques. Recently, using standard datasets of different languages, different competitions have been organized to compare and evaluate different techniques of OCR including document analysis and classification and recognition [5,6]. These annual competitions play an important role for the development and maturation of the algorithms which result in overall performance improvement of OCR systems.

Based on the above discussion, standard corpus is required for pattern recognition and OCR research. The accessibility of the benchmark corpus not only facilitates the researchers to do research but also provides platform to evaluate different techniques.

English is written in Latin script. The development of OCR for printed English text is quite easy as compared to the cursive scripts. It has 52 letters in characters set, each character has single shape. Normally projection profile methods are used to segment English document image into lines, words and characters. The recognition of handwritten English text is a challenging task. Martin and Bunke [7] report the English handwritten dataset of Lancaster-Oslo/Bergen (LOB) text corpus. The text lines extracted from different domains are printed in a proper format on a form. The form is designed properly so that domain of the printed text, text number, printed text, handwritten text and name of the writer can be extracted easily. The writers are asked to write the printed text in the specified area of the form. Filled forms are scanned at 300 DPI at gray scale resolution of 8-bit using HP-Scanjet 6100. The scanned images are saved in tiff format with LZW compression. A total of 556 forms are filled by 250 writers. After pre-processing, total of 4,881 handwritten line images are extracted. These lines have 43,751 words instances and 6,625 words vocabulary. Each line has 8 to 9 words on average. A separate ASCII file is maintained containing the information of each printed and handwritten text line.

Marti and Bunke [8] report English corpus twice as large as corpus [7]. The reported handwritten data set is extracted from 1,066 filled forms, written by approximately 400 different writers. The dataset is comprised of 92,85 handwritten lines and 82,227 word instances covering 10,841 vocabulary words.

This dataset has on average 8.59 lines per form. The average number of words per text line is 8.98.

Arabic language belongs to cursive script. In cursive languages, one or more characters form a ligature. The main stroke of the ligature is called RASM (or main body) and secondary stroke(s) is called IJAM(s) or diacritic(s) [9]. Normally, Naskh writing style is used to write Arabic text. In Naskh, the characters in ligature are written along the baseline. Each character has at most four shapes based on the position in a ligature, such as initial, medial, final and isolated shapes, shown in Figure1.

Due to the cursive nature of the Arabic text, the development of the OCR for Arabic is a challenging task. The main reason of limited research on Arabic OCR is the unavailability of dataset along with character level ground truth information for Arabic language.

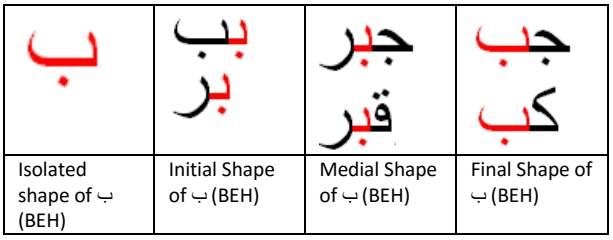


Figure 1. Consistent isolated, initial, medial and final shapes of ب (BEH) in Naskh

Margner and Pechwitz [10] report a process of genrating the synthetic Arabic dataset. The ARABTex is used to generate the Arabic documents from ASCII text. Ground Truth (GT) information consists of character code, font style, size and positional information of character. They also develop statistical HMM based OCR using 946 Tunisian town names dataset, and report reasonable accuracy.

Al-Ma'adeed et al. [11] develop Arabic handwritten database (AHDB). They first design the form to automatically extract information of respective handwritten text. Twenty nine high frequent words and sixty seven words used to write a cheque are printed on the form. In addition, the writers have to write three sentences representing numbers and quantities of cheque. Total of 104 forms written by 104 writers are scanned at 600 DPI using Hewlett Packard 6350 scanner. Mozaffari et al. [12] present handwritten dataset of 52,380 isolated characters and 17,740 numbers which are extracted from filled exam forms of schools. The filled forms are scanned at 300 DPI in gray scale format. Each character image is stored as a 77x95 BMP image. The presented IFHCDB database consists of 52,380 isolated characters and 17,740 numbers, and is divided into training (70%) and testing (30%) data. Kharma et al. [13] develop Arabic handwritten database which has Arabic words, numbers, signatures and complete sentences. Five hundred students are selected to develop this dataset. Each student is instructed to copy a predefined numbers, digits and sentences. The filled forms are scanned in both gray scale, and Black and White (BW) formats using HP scanner. The handwritten digits, words, sentences and signatures are cropped from original grayscale images and saved in BMP file formats. Digits, words, sentences and signatures are also extracted from BW images. This database includes 37,000 Arabic words, 10,000 digits, 2,500 signatures and 500 Arabic sentences. Pechwitz et al. [14] report another publically available IFN/ENIT-database of Arabic. This handwritten dataset contains 946 Tunisian town/villages names along with postcode. A total of 411 writers filled 2,265 forms. The filled forms are scanned at 300 DPI with BW format. During scanning, the page numbers and other information is maintained manually. This dataset contains 26,459 Tunisian town/village names and 212,211 handwritten characters. The GT information including postcode, global word, character shape sequence, baseline (y1,y2), baseline quality, number of words, number of PAW (Part of Arabic Word), number of characters and writing quality is stored against each town/village name image.

Urdu is cursive language which belongs to Arabic script. Normally, Nastalique writing style is used to publish Urdu content such as books, magazines and newspapers in Pakistan. Nastalique is written diagonally and has complex rules to place marks and diacritics. Nastalique has context sensitive character shaping [15]. Based on the shapes similarity, the Urdu character RASMs are divided into 21 classes. Unlike Naskh, Nastalique has character as well as ligature overlapping (Figure 2). The shape of a character depends on the context in which it appears, illustrated in Figure . The detailed analysis of Nastalique is discussed in [15, 16, 17].

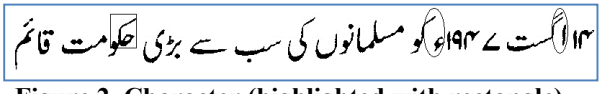
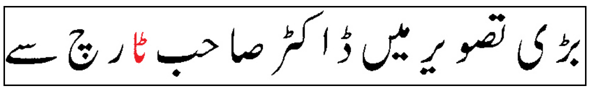


Figure 2. Character (highlighted with rectangle) and Ligature (highlighted with circles) Overlapping

Figure 3. Contextual Character Shaping of Character ب (BEH) at initial position in Nastalique

Nastalique writing style is a compact writing style and due to paper, ink and printing qualities, sometimes the diacritics and RASM are attached as shown in Figure 4. Nastalique has thick-thin-thick transitions of stroke while writing the character/ligature. Due to printing qualities, sometimes the Urdu ligatures are broken at the thin stroke (Figure 5).

(a) Example of RASMs attachment



(b) Example of RASM and diacritic attachment

Figure 4. Diacritics and RASM attachment, the attached connected components are highlighted with red color

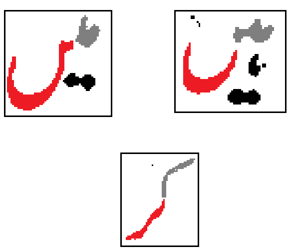


Figure 5. Broken connected components of ligatures (broken connected components with different colors)

The development of OCR for the Nastalique is a challenging task due to the aforementioned characteristics of Nastalique and behavior of paper and printing qualities. Limited research exists in the literature for recognition of Urdu document images written using Nastalique writing style. Real image corpus of published Urdu Nastalique documents is not available for the development of Urdu document processing algorithms and also for the evaluation of different research approaches. Usually researchers use their own corpus of Nastalique writing style. Most of them have been developed manually for large font sizes. [4, 18-25].

In this paper, a comprehensive dataset of Urdu Nastalique document images scanned from Urdu books is presented. This dataset is publically available for the researchers to do research in Urdu document analysis, pattern recognition and OCR.

Methodology

A survey has been conducted to analyze font style and font sizes of Urdu books and magazines for the development of Urdu image corpus. According to the survey, most of the Urdu books and magazines are written using Nastalique writing style having 14 to 40 font sizes. The normal text of the Urdu books is written using 14 and 16 font sizes. In children books, the normal text is written in larger font sizes range from 18-22 font sizes. The remaining font sizes are normally used to write headings. Therefore, based on this analysis, three categories of the document images are defined (1) normal text, (2) normal text for children and poetry books and (3) headings text. The complete process for books collection, corpus acquisition, corpus labeling, and ground truth data generation has been designed for the development of this image corpus. Each of the sub-processes is detailed in subsequent sections.

1.1 Corpus Collection

Corpus collection process is divided into two main phases i.e. corpus design and corpus development. Corpus design deals with selection of books from which the selected document pages will be scanned. Books for each font size category are selected on the basis of defined criterion to ensure variety of domains, paper quality, print quality, paper transparency, and publisher and publication date. Corpus development involves scanning, organization and GT generation of the scanned images. Details are presented in subsequent sections.

Image corpus for 14 to 40 font sizes has been developed for the coverage of different available font sizes. To generate presented image corpus, first step is the selection and purchasing of Urdu books. The selection criterion is detailed below.

1. **Character Set and Symbols**: The image corpus should cover the following:
    a. Urdu alphabet given in Fig.6 below
    b. Latin digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
    c. English characters (A-Z, a-z)
    d. Urdu digits (۰, ۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹)
    e. Urdu aerab (ً ُ ٗ ٍ ٰ ِ ٖ َ)
    f. Other symbols of Urdu, as follows: (ؐ ؑ ؒ ؓ ؔ ٓ ٔ ٕ ء ٘ ؟ () " ' - ؛ : ،)

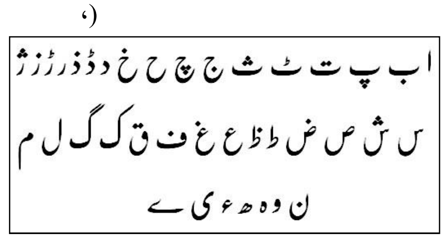


Figure 6. Urdu character set

2. **Font Size and Style:** Based on the survey findings books written using Noori Nastalique should picked up. The selected font sizes are 14-40.
3. **Multiple Domains:** Books are selected from multiple domains for each font size category to address the coverage of a balanced corpus.
4. **Publishers and Publication Date Variety:** Books published from multiple publishers of different cities are selected. In addition, variety of publishers within a city is also considered. Other than publisher, publication date also affects printing as well as paper quality of books. Therefore, while selecting the books, this parameter is also considered and books having variety of publication dates are selected.
5. **Page/Printing Quality:** Paper and printing qualities also affect image quality. All these varieties are included in the Urdu text image corpus to have the standard dataset for Urdu images.

1.2 Corpus Development from Books

Based on the availability of books according to the above mentioned criteria, the number of books and pages are selected according to font size category. To estimate font size of the printed text, Urdu character set and two characters high frequent Urdu ligatures are typed at multiple font sizes range from 14 to 40. These are then printed on transparencies which are placed on the printed text of books to find font size. For normal font size i.e. 14 and 16 font sizes, at least 100 books are selected and five pages from each book are scanned to generate image. In addition, table of content (TOC) page and page with or image/figure are also scanned for the researchers who want to do research on document layout analysis. For the second category of font size i.e. children books which are available less in frequency as compared to the first category, at least 30 books for each of the selected font size i.e. 18, 20 and 22 font sizes, are selected and from each book at least 5 pages are selected to scan. To generate image corpus of third category i.e. heading text, at least 20 books for each of 24, 28, 32, 36 and 40 font sizes are selected and at least 10 headings from each book are marked to scan. The number of books, number of scanned images, and domains coverage for each font sizes are provided in Section 3.

The selected pages of each book are scanned at 300 DPI using HP Scanjet G3110 scanner. During scanning, both BW and gray scale versions are generated for each font size except for 16. For each version, two types of images are scanned; (1) image without cropping the region of interest and (2) image with cropping the region of interest, both samples of gray scale and BW are shown in Figure 7. The images without cropping the region of interest is developed for the researchers who want to do research on page frame detection of Urdu document images. To generate image corpus for headings, the heading textual area is extracted and saved during scanning. All the images are saved in JPG or BMP file format.

1.3 Corpus Organization

The intelligent labeling of the image corpus is essential for the research and development. This is normally done manually and is time consuming task to ensure error free data labeling. The data labeling helps to extract the desired data automatically. Image corpus for each font size is maintained separately to maintain Urdu image corpus in an orderly manner. Moreover, gray scale and BW images are placed separately. Each version of cropped (edited) and un-cropped (unedited) of BW and gray scale are also maintained.

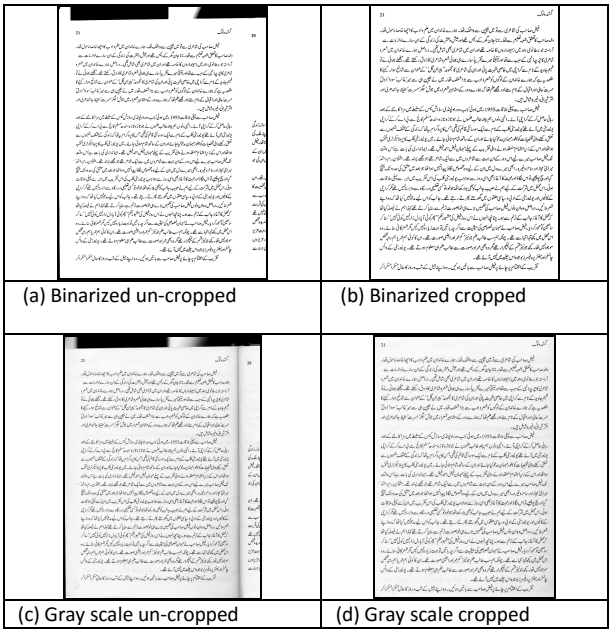


Figure 7. Sample of binarized and gray scale cropped and un-cropped region of interest

The naming convention of images is defined to automatically extract information related to the book, font size, editing and color versioning etc. Each image name for normal text (for 14 to 22 font sized text images) has following tags.

A_B_*C_D_E_F_G.jpg

e.g. BW_UE_B13_R_P26_F14.jpg where

1. ***A*** represents the image format information i.e. gray scale represented by **G** or Black and White represented by **BW**.
2. ***B*** tag represents whether the scanned image is cropped (edited) represented by **E**, or un-cropped (unedited) represented by **UE**.
3. ****C*** When ***B*** tag is E then **C** tag is used to indicate editing type which is cropped for this corpus. Image name does not have C tag when ***B*** tag is UE.
4. ***D*** tag defines the book number (assigned manually) of the image from which it is scanned. The book number has **B** as prefix letter indicating book. This book number can be used to get further information about book including book name, author, publisher, publications date and domain which is maintained in separate file.
5. ***E*** indicates content type of the scanned image. The image can have normal (or regular) text represented as **R**, figure represented as **I**, table of contents represented as **T.**
6. ***F*** is correspond to the page number of book which is scanned to generate the image. The page number is defined with letter **P** as prefix.
7. ***G*** is last tag used for the font size of image. Depending on the font sizes appeared in the text of the image, there can be multiple entries of font size, each is defined with prefix F.

The image corpus for each font size of headings is also maintained separately. As heading images are actually cropped from the document image during scanning. Therefore, cropped and un-cropped versions are not maintained explicitly. The naming convention for the heading image is defined as follows:

A_D_H_F_H#_G.jpg

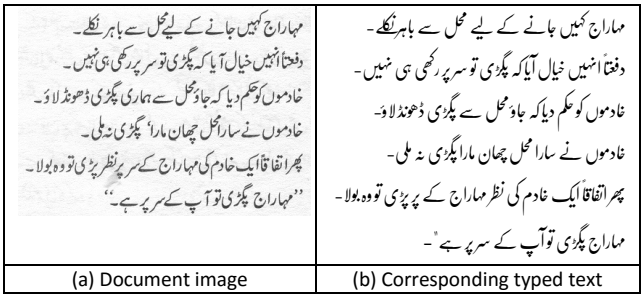
e.g. G_B149_H_P34_H1_F32.jpg where

A, ***D***, ***F*** and ***G*** tags are same as mentioned above. The ***H*** is used to define the type of the image i.e. H indicating the image is of heading. There can be more than one heading on same page. The **H#** is used to define the sequence number of heading in the document image from which the heading image is extracted. The heading number is defined with prefix letter **H** as can be seen in above example.

The complete information of each font size corpus is also maintained manually in separate file during scanning of images. This file provides information related to the book ID, book name, author name, publisher, year of publication, city, total number of pages, domain, image name, available font sizes in image, and columns (either 1or 2) of each scanned image. This information is cross verified to generate the error free details of an image.

1.4 Text Corpus of Images

Parallel typed version of each image is also generated as ground truth data, to process and recognize document images of the reported image corpus. This GT data will assist the researchers to extract training and testing data for classification and recognition by developing segmentation of lines and ligatures systems. Furthermore, this parallel text corpus of the reported image is also helpful for the researchers to develop language models using contextual information for post-processing of OCR system to improve the accuracy. Each scanned document image is typed by two typists. They are given instruction to type text as is and enter carriage return where required to have exact mirror of the image. This means number of lines in text files must be same as number of lines in document image (Figure 8). A total of 2,843 images are typed. Both versions of typed data are manually verified and mistakes are removed. During verification of the text pages, it has been observed that in some pages, typist typed correctly but in document image there were typo mistakes. Therefore, for the training and recognition of Urdu OCR it has been ensured that text corpus should be the mirror of image and those typo errors are remained in the text version. The detailed statistics of text corpus are given in Section 3.



(a) Document image | (b) Corresponding typed text

Figure 9. Sample of image and corresponding typed text

Corpus Statistics

The image corpus has been developed covering variety of domains for each font size. During development, complete information about the page is maintained. The summarized information of number of books, domains and authors is given in Table 1.

Table 1: Statistics of Urdu image corpus

Font size	Book/Magazine count	Number of document images	Domains	Authors
14	101	593	18	76
16	116	595	19	100
18	30	150	10	23
20	45	149	2	24
22	56	151	2	21
24	21	461	18	24
28	21	202	6	21
32	23	186	9	21
36	31	226	7	22
40	26	199	7	22

The scanned document images contain normal text, TOC and figures for the Urdu documents layout analysis research and development. The presented corpora contain total of 29 document images which have figures and 84 document images of TOC. The document images having normal text also have variation of paper, printing, headings, headers and footers etc. The sample layouts of figures, normal text and TOC are shown in Figure 9 and Figure 10.

Table 2. Additional characters' Unicode of Urdu

Identical Form	Decomposed form
ئ (U+0626)	◌ٔ+ی (U+06CC + U+0654)
ؤ (U+0624)	◌ٔ+و (U+0648+ U+0654)
أ (U+0623)	◌ٔ+ا (U+ 0627+ U+0654)
ۂ (U+06C2)	◌ٔ+ہ (U+06C1+ U+0654)
ۃ (U+06C3)	

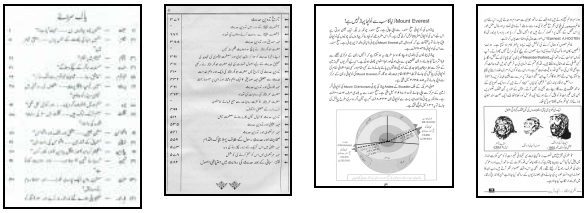


Figure 10. Samples of document images having figures and TOC

Figure 11.Sample layouts of normal text having variation of paper and printing qualities

Table 3. Font wise Urdu Letters, Urdu digits, English letters and digits, Urdu Aerab and symbols statistics

Font Size	Total Characters	Unique Urdu Characters	Unique Urdu Digits	Unique English Characters	Unique Latin Digits	Unique Urdu Aerab	Unique Symbols
14	726,385	45	10	52	10	12	32
16	579,730	45	10	51	10	13	31
18	111,178	44	10	34	10	10	22
20	101,559	44	10	20	10	9	15
22	81,718	43	8	3	10	10	18
24	7,807	43	3	10	6	7	18
28	2,730	42	3	16	4	6	12
32	5,519	40	0	4	10	9	11
36	3,462	42	4	11	3	7	13
40	2,949	42	0	0	0	9	12

Table 4. Font wise lines and ligature statistics of corpus

Font Size	Total document images	Lines	Total Ligatures	Unique Ligature	Average Lines per image	Average Ligatures per Line
14	591	13,712	386,648	6,452	23	28
16	528	11,080	306,080	5,938	20	27
18	150	2,622	60,056	2,872	18	23
20	149	2,318	54,657	2,204	16	24
22	151	1,857	43,121	1,865	12	23
24	461	463	3,961	883	1	9
28	202	203	1,424	502	1	7

32	186	274	2,874	616	2	11
36	226	260	1,776	537	1	7
40	199	222	1,510	498	1	7

Conclusion

In this paper, a comprehensive image corpus of Nastalique writing style is presented. The complete process to select books according to the define criteria, scan and organize the images in orderly manner is defined. In addition, ground truth typed data is also developed. A total of 2, 912 images are selected from 413 books. Among these, 593, 595, 150, 149 and 151 images are scanned for 14, 16, 18, 20 and 22 font sizes. The image corpus for headings contains 461, 202, 186, 226 and 199 heading images for 24, 28, 32, 36 and 40 font sizes respectively. The subset of the reported document image corpus for 14, 16, 18, 20, 22, 24, 28, 32, 36 and 40 are publically available for researchers at [26-35]. Moreover, the typed corpus of each font size is prepared as ground truth information which is also publically available at [36-45].

Acknowledgements

This work has been supported by Urdu Nastalique OCR research project grant by ICTR&D Fund, Ministry of IT, Govt. of Pakistan. See www.UrduOCR.net for details.

References

[1]. R. Smith, "An Overview of the Tesseract OCR Engine," in Ninth Int. Conference on Document Analysis and Recognition (ICDAR), 2007.

[2]. R. Smith, D. Antonova and D.-S. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in International Workshop on Multilingual OCR, Barcelona, Spain, 2009.

[3]. N. Sankaran and C. V. Jawahar, "Recognition of printed Devanagari text using BLSTM Neural Network," in 21st International Conference on Pattern Recognition (ICPR), 2012.

[4]. N. Sabbour and F. Shafait, "A Segmentation Free Approach to Arabic and Urdu OCR," in SPIE, Volume 8658, 2013.

[5]. Antonacopoulos, S. Pletschacher, C. Clausner and C. Papadopoulos, "Competition on Historical Newspaper Layout Analysis (HNLA2013)," in 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.

[6]. D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez-i-Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan and L. P. Heras, "ICDAR 2013 Robust Reading Competition," in 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.

[7]. U. -V. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition," in Fifth International Conference on Document Analysis and Recognition, 1999.

[8]. U. -V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, pp. 39-46, 2002.

[9]. M. Davis, "Unicode Text Segmentation," Addison-Wesley Professional, 2013.

[10]. V. Margner and M. Pechwitz, "Synthetic data for Arabic OCR system development," in Sixth International Conference on Document Analysis and Recognition, 2001.

[11]. S. Al-Ma'adeed, D. Elliman and C. A. Higgins, "A data base for Arabic handwritten text recognition research," in Eighth International Workshop on Frontiers in Handwriting Recognition, 2002.

[12]. S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban and S. M. Golzan, "A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research," in Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule (France), 2006.

[13]. N. Kharma, M. Ahmed and R. Ward, "A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing," in 1999 IEEE Canadian Conference on Electrical and Computer Engineering, 1999.

[14]. M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze and H. Amiri, "IFN/ENIT - database of handwritten Arabic words," in CIFED 2002, 2002.

[15]. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," in International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE), 2006.

[16]. S. Hussain, "www.LICT4D.asia/Fonts/Nafees_Nastalique," in 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore, 2003.

[17]. S. Hussain, S. Rahman, A. Wali, A. Gulzar and S. J. Rahman, "Grammatical Analysis of Nastalique Writing Style of Urdu," Center for Research in Urdu Language Processing, FAST-NU, Lahore, 2002.

[18]. Q. Akram, S. Hussain, A. Niazi, U. Anjum and F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique," in 11th IAPR Workshop on Document Analysis Systems, Tours, France, 2014.

[19]. Hasan, S. B. Ahmed, S. F. Rashid, F. Shafait and T. M. Breuel, "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," in International Conference on Document Analysis and Recognition, 2013.

[20]. Muaz, "Urdu Optical Character Recognition System," Unpublished, MS Thesis Report, National University of Computer and Emerging Sciences , Lahore, 2010.

[21]. S. A. Sattar, "A Technique For The Design And Implementation Of An OCR For Printed Nastalique Text," Unpublished, Degree of Doctor of Philosophy

Thesis Report, N.E.D University of Engineering and Technology, Karachi, Pakistan, 2009.
[22].D. A. Satti, "Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach".
[23].F. Shafait, D. Keysers and T. M. Breuel, "Layout Analysis of Urdu Document Images," in INMIC'06. IEEE, 2006.
[24].S. Tariq and S. Hussain, "Segmentation Based Urdu Nastalique OCR," in 18th Iberoamerican Congress on Pattern Recognition (CIARP 2013), Havana, Cuba, 2013.
[25].M. Naz, Q. Akram and S. Hussain, "Binarization and its Evaluation for Urdu Nastalique Document Images," in INMIC, Lahore, 2013.
[26]."CLE Urdu Image Corpus 14 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurduimage corpus14pt.htm. [Accessed 30 9 2016].
[27]."CLE Urdu Image Corpus 16 Point Size," Center for Language Engineering (CLE), 30 06 2012. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus16pt.htm. [Accessed 30 09 2016].
[28]."CLE Urdu Image Corpus 18 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus18pt.htm. [Accessed 30 09 2016].
[29]."CLE Urdu Image Corpus 20 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus20pt.htm. [Accessed 30 09 2016].
[30]."CLE Urdu Image Corpus 22 Point Size," Center for Language Engineering (CLE), 31 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus22pt.htm. [Accessed 30 09 2016].
[31]."CLE Urdu Image Corpus 24 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus24pt.htm. [Accessed 30 09 2016].
[32]."CLE Urdu Image Corpus 28 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus28pt.htm. [Accessed 30 09 2016].
[33]."CLE Urdu Image Corpus 32 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus32pt.htm. [Accessed 30 09 2016].
[34]."CLE Urdu Image Corpus 36 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus36pt.htm. [Accessed 30 09 2016].
[35]."CLE Urdu Image Corpus 40 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus40pt.htm. [Accessed 30 09 2016].
[36]."CLE Urdu Text Corpus 14 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus14pt.htm. [Accessed 30 09 2016].
[37]."CLE Urdu Text Corpus 16 Point Size," Center for Language Engineering (CLE), 30 06 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus16pt.htm. [Accessed 30 09 2016].
[38]."CLE Urdu Text Corpus 18 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus18pt.htm. [Accessed 30 09 2016].
[39]."CLE Urdu Text Corpus 20 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus20pt.htm. [Accessed 30 09 2016].
[40]."CLE Urdu Text Corpus 22 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus22pt.htm. [Accessed 30 09 2016].
[41]."CLE Urdu Text Corpus 24 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus24pt.htm. [Accessed 30 09 2016].
[42]."CLE Urdu Text Corpus 28 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus28pt.htm. [Accessed 30 09 2016].
[43]."CLE Urdu Text Corpus 32 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus32pt.htm. [Accessed 30 09 2016].
[44]."CLE Urdu Text Corpus 36 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus36pt.htm. [Accessed 30 09 2016].
[45]."CLE Urdu Text Corpus 40 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus40pt.htm. [Accessed 30 09 2016].

Clustering Urdu News Using Headlines

Samia Khaliq, Waheed Iqbal, Faisal Bukhari and Kamran Malik
Punjab University College of Information Technology, University of the Punjab, Lahore, Pakistan.
E-mail: {mscsf13m014,waheed.iqbal,faisal.bukhari, kamran.malik} @pucit.edu.pk

Abstract

In this paper, we proposed and evaluated a new algorithm to automatically cluster Urdu news from different news agencies. This task is challenging as we do not have language processing libraries for Urdu language. Our experimental dataset consists of news from famous Pakistani media houses including Jang, BBC Urdu, Express, UrduPoint, and Voice of America Urdu (VOA). The proposed algorithm only uses headlines to cluster the news. News headline provide a concise summary of the news which motivates us to use it instead of using the entire news story. Our experimental evaluation shows micro and macro averages for precision 0.45 and 0.48 respectively for identifying similar news using headlines.

Keywords—*Urdu News, Clustering, Similar News, News Aggregation*

1. Introduction

Nowadays, most of the media houses publish news online to their websites and social networking sites to rapidly attract readers. Majority of the news readers are interested to read specific type of news according to their own interests. For example, politicians and businessmen are interested to remain updated with latest news especially related to politics as an abrupt change in a political scenario, not only influences the reputation of country but may also have a drastic impact on the economy of the country. Moreover, some of the media houses provide different perspective, biased or unbiased, based on a same news. Therefore, it may be helpful to read same news from different broadcasting agencies. A news aggregation tool is required to automatically aggregate news from various news sources and cluster them for the readers to read same news from different sources.

Urdu language is spoken by more than 100 million people all over the world. However, a limited research is conducted to develop Natural Language Processing (NLP) tools and APIs to process Urdu text easily. In this paper, we propose and evaluate a new algorithm to process Urdu news and cluster similar news. News articles and clustering techniques are widely used over Internet for various language, however, there is no automated service that aggregates and clusters the Urdu news from various Urdu news agencies. Finding similar or related news is beneficial for reader. As these news come from different news sources and reader can easily browse through the same news coming from different sources. Moreover many news agencies provide different perspective based on the same news. So it is important for readers to read the same news from different news agencies.

We have developed a crawler that scraps Urdu news from various Urdu news broadcasting agencies. We have an online portal (htt://www.newslink.pk) that aggregates and show the updated news from various news sources. Now, in this paper we have presented and evaluated our algorithm to automatically cluster Urdu news using only headlines. News headline provide a concise summary of the news which motivates us to use it instead of using the entire news story.

The rest of this paper is organized as follows. Section 2 provides related work of performing clustering using news. Section 3 describes our technique for clustering of Urdu news using headlines. Section 4 provides experimental evaluation and results. Section 5 concludes the paper and explains our future work.

2. Related Work

- Many clustering techniques are used by various researchers that use different criteria for clustering of News articles or reports. In [1] focuses on event centric clustering of news. They get news reports from different news sources and

cluster them according to events. Their system work in online incremental environment. They have used RSS feed as a source of news report and their system clusters news reports from separate RSS feed. Their results show that their system gives much better results while using fine grained clustering technique rather coarse grained technique, which gives poor performance. They have used modified K. Means clustering technique for incremental approach of news clustering.

- Another criterion for clustering news is based on Topic. Shah and Elbahesh [2] discussed their approach for clustering web based news articles into topic wise categories. They have pointed out that search engines that search for news articles have some drawback as these search engines only search on basis of keywords and ignores the whole content of a news article. So they focused on applying text mining techniques and proposed a system that will do clustering of news articles on basis of their topic. For this purpose they have used three clustering algorithms on their dataset. K means, single link clustering algorithm and Hybrid clustering algorithm. Their results show that Hybrid algorithm outperforms other two algorithms and gives much better results than other algorithms.

Some other approaches use WordNet [3] for clustering of news articles. Similarly event centric clustering is focused on events mentioned in articles [4]. Word N-grams is also used for enhancing document clustering by the same researchers who have used WordNet [5]. Cluster centric approach is also followed to extract events from online news where already created clusters are used for extraction purpose [6].

Various news analytical tools use different techniques to classify news articles. Classification techniques are used by many researchers. A news headline has positive and negative effect on its viewers. It also contains emotions. Text of news headlines makes it positive or negative news. Santos, Ramos and Marques work on Portuguese News Headlines. They have classified news headlines as positive, negative and neutral. For this purpose they have used supervised approach and trained a classifier. This classifier classifies news headlines in above three categories. They have used two classification algorithms, Sequential Minimal Optimization (SMO). It is SVM (Support Vector Machine Method). Second classifier they used was Random Forest. These algorithms were used to recognize the features. The experiments were done on syntactic features which they explained as argument1 verb argument2 relation. Results of their experiments show that using these relations as features improves sentiment classification of news headlines [7].

Bergen and Gilpin also worked on classification of news article headlines in positive, negative and neutral news. They tried to uplift positive news headlines so it can positively affect the readers. Their goal was to develop an algorithm that distinguishes between positive and negative stories for this purpose they classify news articles headlines as positive or negative.

In [8] they have collected data for news articles from RSS feeds and sources used were Google News, CNN, Fox News, The New York Times. Bergen and Gilpin used feature extraction for both positive and negative news and two different classification algorithms were used: Naïve Bayes and Support Vector Machine also dataset was also divided in two forms, first in which only headline data was included while other include headline with text. On both datasets both classification algorithms were applied. Their results represent that Naïve Bayes classifier is better than Support Vector Machine. Naïve Bayes have accuracy around 70% while Support Vector Machine accuracy was around 68% [8].

In [9] classification of Thai news was done through structural features. In this paper news web documents from two different sources was collected. The main purpose of this paper was to formulate a simple method that extracts news articles from web collections. They have explored machine learning methods which helped in distinguishing article pages from non-article pages of web collection. After separating article pages they have compared these articles in fine grained manner so that they can identify informative structures present in the articles. In both phases of article extraction from web documents and extraction of informative structures, they have used structural features. For classification they have used three classification algorithms i.e. SVM, Naïve Bayes and C4.5. Their experimental results show that SVM works better than other two algorithms but the difference was not extraordinary.

Ramdass and Seshasai [10] also worked on classification of news articles. They have used dataset of news articles from MIT newspaper The Tech. the Tech archive requires classification of news articles into sections like News, Sports and Opinions. So the main objective of their project was to investigate and implement techniques that classify those articles into their relevant sections. They have used already classified documents so they make use of supervised classification techniques. They have split those documents for training and testing purpose. For experiments they have used different natural language feature sets and also some statistical techniques that used these feature sets. Naïve Bayes classification and Maximum Entropy Classification techniques were used for the experiments.

Their results show that news articles have two different directions one is news content and other is opinion content. The first half of their study which focused at Naïve Bayes and Maximum Entropy classifiers used the content, while the second half looked at grammatical structure. Results of both halves proved that Naïve Bayes and Maximum Entropy classifiers outperformed the results of second half.

Apart from classification, text similarity is used in document clustering. Also techniques are used for clustering of news articles. Anna Huang in [11] discussed and analyzed clustering of documents. The author has done comparison of different measures used to determine similarity of text between different documents. They have used K means algorithm for clustering and for results are compiled on seven different text datasets. Five similarity measures were evaluated by the author. Similarity measures that the author discussed are: Euclidean Distance, Cosine Similarity, Metric, Pearson Correlation Coefficient, Jacard Coefficient. Results of their experiment represent that among similarity measures Euclidean distance performs worst while Jacard and Pearson Coefficient perform better than other similarity measures and produce much better

clusters.

- Thilagavathi, Anitha, and Nethra also worked on clustering of documents that is based on sentence similarity. For clustering of documents they have used fuzzy algorithm instead of hard clustering. They have mentioned that fuzzy clustering algorithm is more flexible and it allows a pattern to belong to the entire produced cluster but the degree of their membership will be different for every cluster. Fuzzy clustering algorithm works on *Expectation-Maximization Framework*. This framework helps in determining the probability of membership of a sentence in a cluster. The authors concluded that FRECCA that is a fuzzy clustering algorithm can be used for any relational problem of clustering. Their results show that fuzzy algorithm helps in avoiding the overlap and gives much better performance [12].

3. Methodology

To achieve our goal of formulating an algorithm for similar news identification following steps are used:

A. Data Gathering:

We have developed a crawler web crawler which gathers headlines from different media houses websites. We have targeted renowned news channels of Pakistan. The crawler is capable to scrap news headline, news image, news date and time, and news story. Currently our crawler is scraping news from the following media houses websites:

- BBC Urdu
- ARY News
- Nawa eWaqt
- Express News
- Jang
- GEO News
- UrduPoint

B. Pre-Processing:

In pre-processing phase we cleaned the news headlines by removing stop words and identify tokens for each news headline. Table I shows the pre-processing of few news headlines. The Table provides news, stop words, and tokens identified for some sample headlines.

Table I: Pre-processing news headlines to identify tokens by removing stop words.

News	Stop Words	Tokens
حکومت، جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	پر، کے، ہے	حکومت، جوڈیشل، کمیشن، قیام، رضامند، عارف، علوی
وزیر اعلیٰ پنجاب سے وزیر داخلہ چوہدری نثار علی خان کی ملاقات	سے، کی	وزیر، اعلیٰ، پنجاب، وزیر، داخلہ، چوہدری، نثار، علی، خان، ملاقات
سعید اجمل کا ورلڈ کپ میں حصہ نہ لینے کا فیصلہ	کا، میں، نہ، لینے	سعید، اجمل، ورلڈکپ، حصہ، فیصلہ
مہند اکمسی، وینڈلیہ ہموشیہ کی ساتوں بری حقیت واحترام سے منائی گئی	کی، و، سے، گئی	مہند اکمسی، وینڈلیہ، ہموشیہ، ساتوں، بری، حقیت، احترام، منائی
ایم کیو ایم نے تنظیم کاونڈر ٹیریم پالیسی پر سفارشات کو حتمی شکل دیدی	نے، پر، کو، دیدی	ایم، کیو، ایم، نے تنظیم، کاونڈر ٹیریم، پالیسی، سفارشات، حتمی، شکل
ملیشیا کی نجی ہوائی کمپنی کا طیارہ لاپتہ ہوگیا	کی، کا، ہوگیا	ملیشیا، نجی، ہوائی، کمپنی، طیارہ، لاپتہ
میلبورن ٹیسٹ: کوہلی اور رہانے ڈٹ گئے	اور، گئے	میلبورن، ٹیسٹ، کوہلی، رہانے، ڈٹ

C. News Clustering Algorithm:

We have design an algorithm to identify similar news. Following are the notations and formulas uses to design the algorithm:

- n_i – news i
- n_j – news j
- $S_{i,j}$ – similarity score between news i and news j
- tl_i - token list of news i
- tl_j – token list of news j
- st_i – size of token list
- st_{avg} – average size of token lists of news i and j
- $m_{i,j}$ – count of similar tokens of news i and news j
- t – a constant threshold value

Where;

$$st_{avg} = \frac{st_i + st_j}{2} \quad (1)$$

$$S_{i;j} = \frac{m_{i,j}}{st_{avg}} \quad (2)$$

The threshold variable is used by the algorithm to identify similar news based on matching number of tokens between given two news headlines. For example, if the calculated similarity score $S_{i;j}$ is greater than or equal to threshold value then both news headlines will be considered as similar. The main part of the algorithm is to identify the list of similar news on a given news headlines. We model this algorithm as a function named getRelatedNewsList() and explained in Algorithm 1.

4. Experimental Evaluation

To evaluate results of clustered news we used a dataset consisting on 500 news headlines. These news headlines are distributed on the following five categories equally:

- International
- National
- Health
- Business
- Entertainment

First we run our algorithm on these categories which give clusters of related news headlines. Now these clusters are compared with the ground truth.

A. Formatting Ground Truth

To compare results of system generated clusters we have defined ground truth for each category. Ground truth is devised manually. We have given each category news headlines to three persons and ask them to mark related news. Then we have compared related news marked by all participants and select those clusters of related news that are common or marked by at least two participants. In this way we get ground truth of clusters for each category.

Algorithm 1: getRelatedNewsList

Input: n_i,news headline which is a source headline and we need to identify list of similar news to this news from our dataset

Result: Algorithm return a list of news similar to given news n_i

```
begin
    relatedNewsList ← null
    t ← 0.5
    datasetList ← getDatasetList()
    for j = 0 to datasetList.size do
        n_j ← datasetList[j]
        S_{i,j} ← getSimilarityScore(n_i, n_j)
        if S_{i,j} >= t then
            relatedNewsList.add(n_j)
        end
    end
    return relatedNewsList
end
getSimilarityScore (n_i, n_j)
begin
    tl_i = _getTokensList(n_i)
    tl_j = _getTokensList(n_j)
    st_i = tl_i.size
    st_j = tl_j.size
    S_{i,j} ← 0
    m_{i,j} ← 0
    for x = 0 to st_i do
        for y = 0 to st_j do
            if tl_i[x] matches tl_j[y] then
                m_{i,j} ← m_{i,j} + 1
            end
        end
    end
    if m_{i,j} > 0 then
        avg = (st_i + st_j)/2
        S_{i,j} = m_{i,j}/avg
    end
    return S_{i,j}
end
```

B. Evaluation Criteria:

After getting ground truth for each category we have evaluated system generated clusters and computed different evaluation measures like precision, recall, F measures. A confusion matrix for category "National" is shown in Table II. Similarly we computed confusion matrix for each category before computing the evaluation metrics.

Table II: Confusion Matrix for Category National.

		Predicted	
		UNC	CN
Actual	UNC	TN =91	FP =1
	CN	FN = 3	TP = 5

Following are the symbols used in the equations to calculate different evaluation metrics:

- TN: True negative are the number of news that are not included in any cluster by ground truth and by our system
- FN: False Negative number of news that our system did not find any cluster but in ground truth these news are included in clusters.
- FP: False Positive number of news that our system includes in cluster but in ground truth these news are not clustered
- TP: True Positive number of news clustered by both ground truth and by system.
- NP: News Population is total number of news headlines in a category.
- TFN: Total False Negative number of news that our system did not find any cluster but in ground truth these news are included in clusters.
- TFP: Total False Positive number of news that our system includes in cluster but in ground truth these news are not clustered
- TTP: Total True Positive number of news clustered by both ground truth and by system.
- UNC: Number of news that are not in any cluster
- CN: Number of news that are part of specific cluster.

For each category we computed Precision (P), Recall(R) and F1 Measure (F1) from this confusion matrix using the following formulas:

$$P = \frac{\text{TP}}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2\text{PR}}{P + R} \tag{5}$$

Overall Macro average Precision (M_{acp}), Macro average Recall (M_{acR}), Macro average F-Measure (M_{acF1}), Micro average Precision (M_{icp}), Micro average Recall (M_{icR}) and Micro average F1 Measure (M_{icF1}) of all categories are calculated using the following formulas:

$$M_{acp} = \frac{1}{n}\sum_{i=1}^{n} P(C_i) \tag{6}$$

Where $P(C_i)$ represents the Precision for specific i^{th} category.

$$M_{acR} = \frac{1}{n}\sum_{i=1}^{n} R(C_i) \tag{7}$$

Where $R(C_i)$ represents the Recall for specific i^{th} category.

$$M_{acF1} = \frac{1}{n}\sum_{i=1}^{n} F1(C_i) \tag{8}$$

Where $F\,1(C_i)$ represents the F-Measure for i^{th} category.

$$M_{icp} = \frac{TTP}{TTP+TFP} \quad (9)$$

$$M_{icR} = \frac{TTP}{TTP+TFN} \quad (10)$$

$$M_{icF1} = \frac{2*M_{icp}*M_{icR}}{M_{icp}+M_{icR}} \quad (11)$$

C. Results:

Table III shows category wise Precision, Recall, and F-Measure. We also show micro and macro averages for each of these measures. The highest Precision (0.84) is achieved in National category, however, the lowest Precision (0.26) is found in Business category. The micro average for Precision is obtained is 0.45, however, the macro average for Precision is obtained 0.48. The result looks far better than a random probability of (0.2) for Precision. Therefore, the proposed algorithm is effective to cluster similar news. We show some sample clusters identified by our proposed algorithm in Table IV.

Table III: Category wise Precision, Recall F-Measure

Category	Precision	Recall	F-Measure
International	0.62	0.5	0.55
National	0.83	0.63	0.71
Health	0.43	0.5	0.46
Business	0.26	0.35	0.3
Entertainment	0.29	0.33	0.31
Micro Average	0.45	0.46	0.46
Macro Average	0.48	0.46	0.47

Table IV: Clustering Results

Sr#	Related News	Source
Cluster1	بنگلادیش: جماعت اسلامی کے ایک اور رہنما کے لیے سزائے موت	VOA
	بنگلہ دیش: جماعتِ اسلامی کے رہنما اظہر الاسلام کو سزائے موت	BBC
	بنگلہ دیش میں ٹریبونل نے جماعت اسلامی کے ایک اور رہنما کو سزائے موت سنادی	Express
Cluster2	ذکی الرحمن لکھوی کی نظربندی کا حکم معطل	BBC
	ذکی الرحمن لکھوی کو ایک مقدمے میں گرفتار کرلیا گیا	BBC
	ذکی الرحمن کی نظربندی کا حکم معطل	VOA
	ذکی الرحمن لکھوی ایک اور مقدمے میں گرفتار	VOA
Cluster3	موٹروے پولیس کی تنخواہوں میں 20 فیصد اضافے کا اعلان	Jang
	وزیراعظم کا موٹروے پولیس کی تنخواہوں میں 20 فیصد اضافے کا اعلان	Express
	وزیراعظم کا موٹروے پولیس کی تنخواہوں میں 20 فی صد اضافے کا اعلان	UrduPoint
Cluster4	4 ارب سال قبل مریخ زمین جیسا تھا	BBC
	4ارب سال قبل سیارہ مریخ کا ماحول زمین جیسا تھا، ناسا کا دعویٰ	NawaeWaqt
	4 ارب سال قبل مریخ کا ماحول زمین جیسا تھا	BBC
Cluster5	سینٹ لوشیا: پاکستان کا ہدف 243 رنز	BBC
	سینٹ لوشیا: پاکستان کی پہلے دھنگ	BBC
	سینٹ لوشیا: ویسٹ انڈیز کا ہدف 230 رنز	BBC
	سینٹ لوشیا: پاکستان کا ہدف 262 رنز	BBC
	سینٹ ولسنٹ: ویسٹ انڈیز کا پاکستان کو جیت کیلئے 153 رنز کا ہدف	Jang
Cluster6	انسداد دہشت گردی کے قومی ادارے "نیکٹا" کو بحال کرنے کا فیصلہ	UrduPoint
	انسداد دہشت گردی کے قومی ادارے "نیکٹا" کو بحال کرنے کا فیصلہ	Jang
	اانسداد دہشت گردی کے قومی ادارے کی فوری بحالی کی ہدایت	VOA
	انسداد دہشت گردی کے قومی ادارے نیکٹا کو بحال کرنے کا فیصلہ	GEO
Cluster7	حکومت انتخابی دھاندلی کی تحقیقات کیلئے جوڈیشل کمیشن کے قیام پر رضامند ہوگئی، ڈاکٹر عارف علوی	Express
	حکومت جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	GEO
	حکومت جوڈیشل کمیشن کے قیام پر رضامند ہے، عارف علوی	UrduPoint

5. Conclusion and Future Work

Urdu language readers are presented all over the world. One of the popular things for these readers is to read Urdu news from the Internet. In this paper, we presented and evaluated an algorithm to automatically cluster similar news from various Urdu news media houses. Our experimental evaluation shows an average micro average for Precision measure is 0.45 and the macro average for Precision is 0.48. We believe that the work will help us to provide a tool for Urdu news readers to read same news from different broadcasting services to understand different perspective on the same news.

Currently, we are integrating this work with our on line portal (newslink.pk). We are also looking to improve the processing time of the algorithm by using Apache Hadoop.

6. References

[1] J.Azzopardi and C.Staff, "Incremental clustering of news reports," *Algorithms*, vol. 5, no. 3, pp. 364–378, 2012.

[2] N. A. Shah and E. M. ElBahesh, "Topic based clustering of news articles," in *Proceedings of the 42Nd Annual Southeast Regional Conference*, ser. ACM-SE 42. New York, NY, USA

[3] C. Bouras and V. Tsogkas, "W-kmeans: Clustering news articles using WordNet," in *14th International Conference on Knowledge Based and Intelligent Information andEngineering Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 379–388.

[4] J. Borglund, "Event centric clustering of news articles," Uppsala University, Department of Information Technology, Tech. Rep. 13 072, 2013.

[5] C. Bouras and V. Tsogkas, "Enhancing news articles clustering using word n-grams," in *2nd International Conferenceon DataManagement Technologies and Applications*, 2013.

[6] J. Piskorski, H. Tanev, M Atkinson and E. Van Der Goot, "Cluster centric approach to news event extraction," *in Proceedings of 2008 Conference on New Trends in Multimedia and Network Information Systems*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2008, pp.276–290. http://dl.acm.org/citation.cfm?id=1565754.1565782

[7] C. Ramos and N. C. Marques, "Sentiment classification of Portuguese news headlines" in International Journal of Software Engineering and its Applications, vol. 9, Portugal,2015, pp. 9-18. http://dx.doi.org/10.14257/ijseia.2015.9.9.02

[8] K. Bergen and G. Leilani, "Negative news no more: Classifying news article headlines," Stanford University, USA, Tech. Rep., 2012.

[9] S. Tongchim, S. Virach, and H. Isahara, "Classification of news web documents based on structural features," in *Advances in Natural Language Processing*, vol. 4139, 2006, pp. 153–160

[10] D. Ramdass and S. Shreyes, "Document classification for newspaper articles," 2009.

[11] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, 2008, pp. 49–56.

[12] K. G.Thilagavathi, J.Anitha, "Sentence similarity based document clustering using fuzzy algorithm," *International Journal of Advance Foundation and Research in Computer*, vol. 1, 2014.

Database Schema Independent Architecture for NL to SQL Query Conversion

Saima Noreen Khosa
NCBA&E
saimakhosa@yahoo.com

Muhammad Rizwan
Khwaja Fareed University of Engineering and IT
rizwan2phd@gmail.com

Abstract

NL (Natural Language) to SQL (Structured Query Language) query conversion overcomes communication gap between databases and non-technical user. It is very easy way to write question in NL and system automatically convert this question into a SQL query and show result to the user. Here we propose the database schema independent architecture for NL to SQL query conversion. User is free to give input in its own way in English. System is not restricted user to give input in a specific pattern. Existing proposed solutions are usually schema naming structure dependant and does not apply other than their specific schemas, but we propose a generic architecture which is independent of any particular schema and rather work on all schemas uniformly within the defined scope of the proposed architecture.

1. Introduction

For many long time computer programmers try to overcome the communication gap between causal user and computer. In most of places computer are used for sorting data and retrieval (according to the need and requirement of the user). During last decades there is lots of work done in NL (Natural Language) to SQL (Structured Query Language) conversion. NL is a natural language which is used in a daily life for the communication between human. And SQL is a query Language which is used to retrieve data from relational databases. The process of conversion of NL to SQL is divided into step by step levels. For example morphology deals with the meaning of smallest part of word. Lexical Level deals with sentence structure and tokenization. Here we also deal with the possible meaning and extract the one which is suitable for this use programmatic knowledge. Semantic and syntactic deals with one the limit of sentence but in some scenario the exact meaning comes beyond the limit of one sentence.

2. Related work

In this section we review the existing work regarding NL to SQL conversion.

2.1. Midway Query Generation

An already proposed system changes English statement enter by novice user in all midway query. So users can choice a one of that intermitted query [1] to find which one is more close to her requirement. After selection of final requirement, system fairs a SQL query. If any error occurs in system then user often frizzed. To overcome this problem writer gives a recommendation framework in future. Understanding of any language by machine is a difficult task. For this purpose use parsing rules which convert any natural language statement into computer understandable statement.

Generally NLP has following steps to process Natural language: first one is morphological analysis in which every single word is analyzed and split punctuation also. Then Syntactic analysis was checks the rules of language. If sentence are syntactically incorrect it will rejected. After syntax semantic are checked. Sometimes impact of previous sentence come on the upcoming sentence, it called Discourse integration and in last Pragmatic analysis phase comes. System facilitate user insert and delete value.[1]

2.2. Opportunity of modifies the system

Some systems give opportunity to user that they extend or modify system in any other language as English, German, and Greek. Under discussed system allow user to enter input near the SQL format. System gets three inputs. Firstly get SQL schema. Writer gives a specific pattern for make a file of SQL schema for the system. Second is SQL keys file. This file contains the relation between word phrases with SQL keys words. There is also a specific pattern to write this key file. Third input to the system is user query. User query enters also in a reserved given pattern. This is near to SQL query format, but not exact. [2]

After entering first two inputs in the system, Lexical Analyzer analyzes them. After that Lexical Analyzer give its output to the Syntax Analyzer. User query is also an input of Syntax Analyzer. Syntax Analyzer checks both inputs. After checking final process was held and SQL query is generated. Writer also makes it feasible about the extension process of the system. In which, system is extended easily in other languages also. For this proved a Key words file to the system. This file maps word phase of that language with SQL key words. And the remaining process are same as discussed earlier.[2]

2.3. Conversion of Urdu question into SQL

NLIDBs are one of the mechanisms which accomplish this task. User give input to NLIDB in its daily routine language and the answer is also given in same language. Authors discuss NLIDB for Urdu language. Algorithm discuss in this paper is efficiently maps the given Urdu question into SQL queries. Discuss algorithm implemented in c#.NET and test on student Information System and Employee Information System [3].

In this paper writer discuss some requirement of natural language interface to database. The query placed by user is either a request or question. Presented system evaluation the query must match one given category. It is also important to identify the rule of parameter in query. Parameters are table attribute values. Division of sentence is important for the understanding of computer. The parts of sentence are called tokens. To divide the sentence into tokens in called token formulation. After token formulation the process come is syntactic markers to make query semantically correct it is important to define and remove syntactic markers. It is important to extract the necessary parameters, for the successful translation of query. These parameters are table name, attributes and value.

Parser identifies the parameters and constructs. Construction of dictionary is important which keep the synonyms of columns and tables names. Inclusion of synonyms makes it possible for user to write a sentence in different natural way. The main propose of the constructed system is to track the correctness of query semantically. For this propose constructed a semantic dictionary. [3]

2.4. Natural Language Interface for DB

Generally computers are used to storing data and retrieve data according to the need requirement of the user. For the retrieval of data from the data base user employed SQL Language. If anyone works in relational databases then he/she must know SQL query structure how to write a query in SQL. Causal users don't know how to write query in database. Here writers proposed a Web Database System which creates an ease for novice user to access data from the database. Without having knowledge about SQL or internal structure of database, user just write query in natural Language and the system convert it into SQL query. Retrieve data and show result in natural language. This system gives a suitable answer for single database. This system restricted user to give input in a specific pattern. For example users enter a question system check that words are comprised in data dictionary of system if yes then further proceed and if not then show a massage to the user enter a right question. This is relevant to the database or data dictionary. After this, entering a correct form of input, system creates tokens of the input and removes extra words. There are some rules already defined. After removing extra words system mapping with that rules when rules are mapped a SQL query is created and run in the system and retrieve a required data from the database. And show arranged result to the user. [4]

3. Scope of Proposed System

In this paper we propose a generic system for NL to SQL conversion. Firstly we present the scope of the system for better understanding of proposed system architecture which is presented later on. As we know that scope of the system is very important to accurately understand the working of the system. Our system would work best by fulfilling the following constraint / scope w.r.t database and its schema:

1. System accepts only English language query.
2. Table names and column names should be complete English words within database schema. It should not be abbreviation or short word as it decreases the efficiency and accuracy of proposed system.
3. If any schema entity name is multiword, it should be connected with underscore.
4. Multiword name should have maximum two words.
5. Columns name should be unique within the schema.
6. System can accept only single line input query which can be extendable in future work.

4. Proposed System architecture

First we present architectural diagram of the proposed system as under:

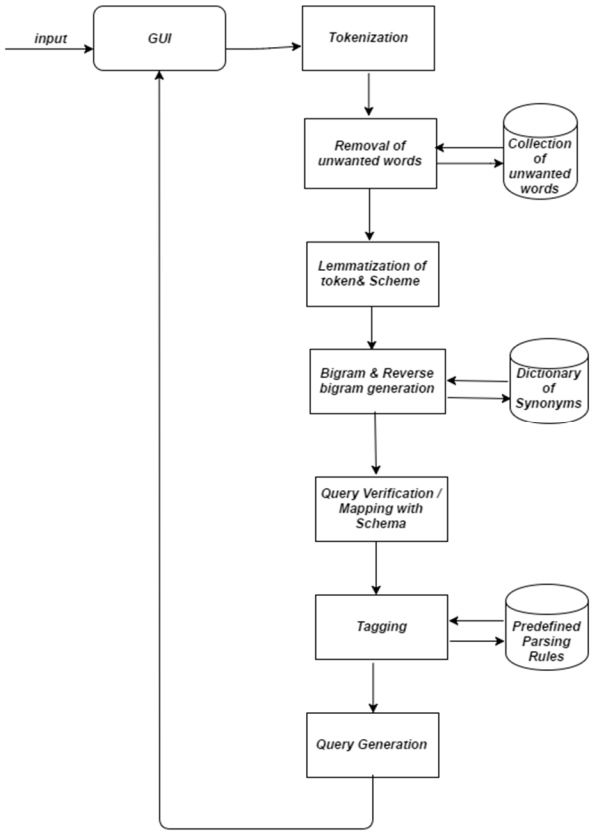


Figure 1: Schema Independent NL to SQL System Architecture

As we can see the Figure 1, proposed system architecture is divided in this sequence of steps mention as follows:

- Tokenization
- Removal of unwanted words
- Lemmatization of query tokens as well as schema entities
- Generation of word bi-gram entities and their reverse combination
- Query verification / Mapping with schema
- Tagging
- Query generation

5.1. Tokenization

In tokenization phase we break text stream into single word tokens. Upcoming phases require morphology processing so we need to break up user input query into lexemes or single word tokens so that we could deal with each token separately. At this stage we not only tokenize the user input but also the schema entities name. Tokenization is very essential part of NLP. It helps to understand semantics of smallest parts of a sentence. Entered query spilt into single word tokens, for further processing. Consider the following example user input:

"Find out employees name whose salary is equal to 5000"

After tokenization, we will have following tokens with their corresponding indexes.

Table : 5.1

Token	Find	Out	Employees	Name	Whose	Salary	is	equal	To	5000
Index	1	2	3	4	5	6	7	8	9	10

The reason for generating index is that later on we can refer these tokens w.r.t to their index and this approach would also help to implement this system as well.

5.2. Removal of unwanted words

User enters query in natural language (we consider English language; other languages are not in the scope of this paper). After tokenization, the proposed system removes the unwanted words in the question from the user query. Only those words would remove which have no semantic importance regarding NLP to SQL query conversion. After filtering unwanted words e.g. and, or, is etc which should be predefined in "Escape Word Dictionary" this phase gives output as under for the same query cited in previous phase.

Table. 5.2

Token	Find	Out	Employees	Name	Whose	Salary	equal	5000
Index	1	2	3	4	5	6	8	10

Here token number 7 and 9 has been removed as they are escaped words.

5.3. Lemmatization of tokens and schema

As it is very difficult to exactly match the words with different forms so we need to perform lemmatization before any verification with schema. Lemmatization means to cut off the last part of the word which occur in different shapes & spellings, so that all form of the word within particular text would be same.

e.g.

am, are, is → *be*

Car, Cars, Car's → *Car*

The target schema keywords also need to convert in lemmatized form so that the verification process in the upcoming verification phase would give us the maximum accuracy. Now by applying the lemmatization on Table 5.2, the output table would be as under:

Table :5.3

Token	Find	Out	Employee	Name	Whose	Salary	equal	5000
Index	1	2	3	4	5	6	8	10

Form above table we can see that Employees converted to Employee after lemmatization, and rest of the words already in lemmatized form, so there is no change in these words.

5.4. Generation of Bi-Gram and Reverse Bi-Gram

In our scope we already mentioned that schema entities having two words attributes or table name are allowed. So in the upcoming verification phase we also need to match bi-gram query tokens with bi-gram database schema entities tokens. We will generate dictionary of Bi-gram and their reverse combination. The bi-gram tokens of synonyms would also be generated. The wordnet can be very handy for this task[8]. e.g.

Table 5.4 .Bi-Gram from Query Tokens

find_out	out_employee	employee_name	name_whose	whose_salary
1_2	2_3	3_4	4_5	5_6

salary_equal	equal_5000
6_8	8_10

Table 5.5. Reverse Bi-Gram of Query Tokens

out_find	employee_out	name_ employee	whose_ name	salary_ whose
2_1	3_2	4_3	5_4	6_5

equal_ salary	5000_ equal
8_6	10_8

5.5. Query verification / mapping with Schema

In this phase we have to verify schema entities tokens (table names and column name) as per the output of previous phase with the query tokens. We also extract synonyms of each lemma and make lemma dictionary of use input. We then generate bi-gram and their reverse combination lemma and add in the dictionary too. After that we verify each token Bi-grams and their combination with the lemmatized schema. So that we can identify the columns and tables name in the user input. We tag each column and table in the input which is exactly matched with the lemmatized schema.

5.6. Tagging

After verification we tag lemmatized tokens for the use of next phase and to identify context of that particular token with schema. Tagged can be done with respect to three ways:

i. Attribute or Column tag
ii. Table tag
iii. Index Number

Attribute tag shows that word is an attribute or column in that schema, table tag shows that word exist in the schema as table. Index number shows the indexing number in tokenized array. e.g

Table 5.6

Token	*Tags*
Employee_Name	Attribute
	Employee (Table Name)
	3_4 (index number)

Table 5.7

Token	*Tags*
Salary	Attribute
	Employee (Table Name)
	6 (index number)

From table 5.6 & 5.7, we can see that we assign each uni-gram and bi-gram token to their corresponding tags after verification. These tagged_tokens would be the output of this phase and would help to make actual SQL query fragments against the user input in upcoming phase. We also tag numbers as "cardinal number".

There are many scenarios in natural text which can be tagged / identified by using state-of-the-art NLP algorithms and tools such as Apache Open NLP [5], Stanford NLP [6], LingePipe [7] which provides implementation of state-of-the-art algorithms of tokenization, POS tagging, sentence splitting, named entity extraction and different NLP task. Form future

point of view, we can also tag named entities (date time and place), conditional words and booster words as the field of NLP has significant success and accuracy in these areas.

5.7 Query Generation

5.7.1. Fragment Creation

In this phase different fragments of SQL query are generated against the natural language query.

SQL query "where" part is extracted and matched from each fragments and their corresponding rules. In this phase, we create fragments which consist of tagged words and suffix and prefix of tagged word. We do not include suffix and prefix if it is another tag word. Tag word in the query which refers to the column or table of target schema.

5.7.2. Query Parts Creation

In this phase we try to create different query parts from fragments. FROM parts is identify based on table tagged tokens. If multiple tables exist then we try to join them based on some common attribute between two successive tables. If no table exists then we try to match attribute in all table of schemas.

"Select" part of SQL query is extracted based on all attribute tagged tokens. If there is no tagged attribute then * is considered as select part. The "select" and "from" clause can be extracted by the help of tagged attributes and tagged tables.

5.7.3. Parsing Rules Mapping

We create some morphological rules. We map these rules with tagged input. We extract different fragment from the input which match according the rules. There are different types of fragments like WHERE clause, SELECT clause, FROM clause. The following table shows the rules for "where" clause of SQL query.

Table 5.8

	Scenario	SQL condition	*<A> stands for attribute <CN> stands for cardinal number <BW> stands for booster word*
1	<A><CN>	<A> = <CN>	
2	<A>greater than<CN>	<A> > <CN>	
3	<A>less than<CN>	<A> < <CN>	
4	<BW><A>	<BW><A>	
5	<A>equal to<CN>	<A> = <CN>	

From table 5.8, we can see that what possible parsing rules can be extracted from user natural text input.

According to each fragment and its type we generate SQL fragments. By combing all SQL fragments we can finally generate complete SQL query.

6. Conclusion

This paper shows that the concept of NL to SQL query conversion in broader scope w.r.t multiple database schemas support because other existing system depend on specific single schema which narrow their scope but our proposed system works schema independently. System gets input in plain English language. User is not restricted to follow any pattern. Advancement in this field can give benefit to large number of novice user or businessman who wants to directly explore their organization databases without technical knowledge of SQL.

In future work we try to modify the architecture so that it might have multilingual support and more sophisticated morphological rules with latest NLP advancement which can map more SQL fragment like "group by" & "order by" etc as well.

7. References

[1] Bhadgale, Anil M., et al. "Natural language to SQL conversion system." IJCSEITR 3.2 (2013): 161-6.

[2] Papadakis, Nikos, Pavlos Kefalas, and Manolis Stilianakakis. "A tool for access to relational databases in natural language." Expert Systems with Applications 38.6 (2011): 7894-7900.

[3] Ahmad, Rashid, Mohammad Abid Khan, and Rahman Ali. "Efficient Transformation of a Natural Language Query to SQL for Urdu." Proceedings of the Conference on Language & Technology. 2009.

[4] Alexander, Rukshan, Prashanthi Rukshan, and Sinnathamby Mahesan. "Natural Language Web Interface for Database (NLWIDB)." arXiv preprint arXiv:1308.3830 (2013).

[5] OpenNLP, Apache. "a Machine Learning Based Toolkit for the Processing of Natural Language Text." URL http://opennlp. apache. org (Last accessed: 2016-09-18).

[6] Manning, Christopher D., et al. "The Stanford CoreNLP Natural Language Processing Toolkit." ACL (System Demonstrations). 2014.

[7] Carpenter, Bob, and Breck Baldwin. "Text analysis with LingPipe 4." LingPipe Inc (2011).

[8] Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. "WordNet:: Similarity: measuring the relatedness of concepts." Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004.

Sentence Level Sentiment Analysis Using Urdu Nouns

Faiza Hashim
Department of Computer Science,
University of Peshawar, Pakistan
f1.shah@yahoo.com

Mohammad Abid Khan
Department of Computer Science
University of Peshawar, Pakistan
mabid6@gmail.com

Abstract

Some pioneer effort has been made for Urdu sentiment analysis and a lot more is needed to explore this field. The goal of this research study is to contribute in the field of sentiment analysis for Urdu language by extending the scope of sentiment expressions from adjectives to nouns in the domain of news. In this study, a sentiment analyzer for Urdu opinionated data is developed which works at sentence level to analyze public opinions given on news headlines. A lexicon based approach is used by the developed sentiment analyzer which exploits Urdu lexicon of adjective and noun class expressions. Urdu opinionated corpus construction and Urdu sentiment lexicon construction are part of this study. The main novelty of this study is the use of nouns as sentiment carriers along with adjectives. Experimental results are used to evaluate and show better performance of the system compared to previous approaches used for Urdu sentiment analysis.

1. Introduction

Sentiment analysis (SA) is the fastest growing field of Natural Language Processing and Text Mining under the umbrella of Artificial Intelligence (AI). The field of sentiment analysis emerged from human behavior of decision making by consultation and asking friends or family about their opinions in daily life. Examples are buying a product or a service, observing public response about a specific law or policy approved by government; increasing demands of customers and employees performance by their companies and political reviews during elections.

The increasing availability and use of online resources for opinion sharing on social media, such as news, blogs, review sites, has greatly facilitated the decision making process of several parties such as customers, governments and companies. A huge volume of opinionated data is entered daily by exploiting the World Wide Web over the internet in different local and regional languages along with English. As compared to English language, these local and regional languages are observed as resource poor languages as data and tools are scarcely available. Developing automated sentiment analyzers for such resource poor languages is a challenging task these days.

Sentiment analysis can be performed on three different levels, depending on the nature of the data and choice of the users. These three levels of analyses are:

- Document level sentiment analysis
- Sentence level sentiment analysis
- Aspect level sentiment analysis

The document level sentiment analysis classifies the overall sentiments of a document as being positive or negative. A single result, positive or negative, holds for the whole document which fails to detect sentiments about individual aspects of the topic [1].

The sentence level sentiment analysis classifies each sentence as positive, negative or neutral. At this level, sentence must be a separate unit, expressing a single opinion.

In aspect level, the data for analysis comprises those entities which have one or more attribute(s)/feature(s). The opinions on these features can be expressed by different opinion holders on a single feature. The aspect level sentiment analysis produces a feature-based opinion summary of multiple reviews, classifying sentiments of each attribute/feature as positive or negative.

The two main approaches used for the problem of sentiment analysis are machine learning approach and lexicon-based approach. In machine learning approach, the text classification is performed by training the classifier on labeled data. Any text classification algorithm can be used for this approach, such as Naïve

Bayes, SVM (Support Vector Machine), Maximum Entropy, Multilayer Perception, and Clustering.

The lexicon-based approach works on sentiment words dictionary, i.e. the lexicon. The lexicon consists of predefined list of sentiment words associated with their polarity and intensities. The lexicon varies according to the context and it is difficult to maintain a unique lexicon which can work in different contexts. As compared to machine learning approach, lexicon-based approach is considered as a simple approach as it does not require labeled data [2].

Corpus-based and Dictionary-based approaches are also used by some researchers. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases. Dictionary-based approaches use synonyms and antonyms in WordNet to determine word sentiments based on a set of seed opinion words [3].

The role of adjectives as sentiment carrier is central and cannot be denied. However, in some specific situations e.g. news domain, subjective nouns can also work as sentiment carriers and can increase the performance of the system as nouns appear frequently in news data. Example (1), (2), (3), (4) and (5), taken from [4], show the importance of nouns as sentiment expressions. In these examples, the underlined terms are nouns, which are used to classify the opinions as positive or negative.

(1) بلاول کی نقل و حرکت کو محدود کرنے کی سازش ہے۔
Bilawāl ki naql o harkath ko mehdud karne ki sāzish hai.
"This is a conspiracy to restrict Bilawal's mass compaign".

(2) کچھ نہیں سب ڈرامے بازی ہے۔
Kuch nahi sab dram e bāzi hai
"It is nothing but a stage act".

(3) اللہ حامد میر کو جلد صحتیاب کرے۔
Allah Hāmid mir ko jald sehathyāb kary
"May Allah make Hamid Mir well soon".

(4) میں جماعت اسلامی کے سیاسی کردار سے مکمل اتفاق کرتا ہوں
Mein Jamāt-i Islami ke siyāsi kirdār se mukam'al ithifāq kartha hun
"I fully endorse Jamat e Islami's political action".

(5) ایسے لوگ صدیوں میں پیدا ہوتے ہیں۔صائمہ شہادت کے مرتبےپر فائز ہوئی ہیں۔صائمہ کو سلام۔
Aisy log sadiyo mein paida hothy hain. Sāimā shahādad ke marthaby par fāiz hui hain. Sāima ko salām
"Such people are born in centuries. Saima embraced martyrdom. We pay our tribute to her".

2. Motivation

While dealing with sentiment analysis for English language and other international languages, this field appears as an already greatly explored one. The reason is that a lot of tools and resources are available and much research is done and is going on. However, moving to regional languages, sentiment analysis appears as a newly emerging field.

Urdu is the national language of Pakistan and is spoken by 60.5 million speakers in the Indian subcontinent [5]. Urdu is written with Arabic script from right to left and the recommended writing style is Nastalique [6]. The research progress in such resource poor languages is a challenging task. The main motivation for the present research work is to facilitate sentiment analysis for Urdu language and develop automated sentiment analyzer to mine Urdu opinionated data using nouns as sentiment expressions along with adjectives.

3. Related Work

Sentiment analysis is a text classification technique which is used to classify opinions as positive, negative or neutral. Usually, this classification starts at word or phrase level [7, 8, and 9] and moves to sentence level [10. 11, 12, 13, 16 and 18] and to document level [14, 15].

Some effort was also done to draw a comparison for better sentiment analysis between different levels of analyses [16, 17].

In some cases the output of one level is used as input to other levels, as reported in [17, 18, 19].

The lexicon based approach works on sentiment words dictionary, i.e. the lexicon [14, 35].

In sentiment lexicons are available which include WordNet [18, 20] and SentiWordNet [21, 22, and 23].
In the field of sentiment analysis, adjectives are traditionally considered the center of attention as adjectives are sentiment carriers which are used to determine the polarity of given text [9, 15, 26, 29, 35, 36 and 37]. Some effort has been made to consider other parts of speech along with adjectives as sentiment expressions. The work includes the use of adjectives and verb class information [18, 38], adjectives and adverbs [15, 39], "adjective verb adverb" framework [33] and non-effective adjectives and adverbs [34].

While dealing with Urdu sentiment analysis, only adjectives are considered as the sentiment expressions in opinionated text [9, 25, 26, 28 and 29].

The tools and automated systems developed so far for English language cannot be used exactly for Urdu data due to the vast orthographic, morphological and grammatical differences between both the languages [9].

Mukund and Srihari made a pioneer effort in Urdu sentiment analysis and developed a classifier to distinguish subjective sentences from objective sentences of Urdu language [24].

Urdu sentiment lexicon was developed by [9] for the first time to performed lexicon based sentiment analysis to mine Urdu opinionated data using shallow parsing.

The work reported in [25] directed the core issue in analyzing sentiment i.e. negation handling and handled negation at phrase level.

Sayed et al. highlighted the importance of adjectival phrases in sentiment analysis and used the term "SentiUnits" for expressions containing sentiment information [26].

Mukund and Srihari used the method of structural correspondence learning (SCL) to transfer sentiment analysis learning from Urdu newswire data to Urdu blog data exploiting the code switching and code mixing techniques [27].

The work reported in [28] performed lexicon based sentiment analysis at aspect level and associated targets with sentiment expressions and the result was a better performance compared to their earlier work [9].

A bilingual lexicon was developed by [29] using bilingual datasets (English and Roman Urdu) and performed lexicon based approach for bilingual sentiment analysis of tweets.

4. Methodology

The current study uses a lexicon based approach for Urdu sentiment analysis which works at sentence level to categorize public opinions posted on news headlines [4]. Although machine learning approaches give promising results but due to the unavailability of Urdu opinionated labeled data, the current research adopted a lexicon based approach. Sentence level sentiment analysis gives better performance than document level sentiment analysis because at sentence level sentiment analysis each opinion is analyzed and considered for being positive or negative, however in document level sentiment analysis a single result holds for whole document. Considering the nature of data, aspect level sentiment analysis cannot be used as in news data no direct mapping of entities with their attributes exists. Therefore sentence level sentiment analysis is more useful for this study. Urdu lexicon and Urdu corpus construction are additional tasks of this study. Figure 1 shows a complete step by step workflow for Urdu sentiment analyzer.

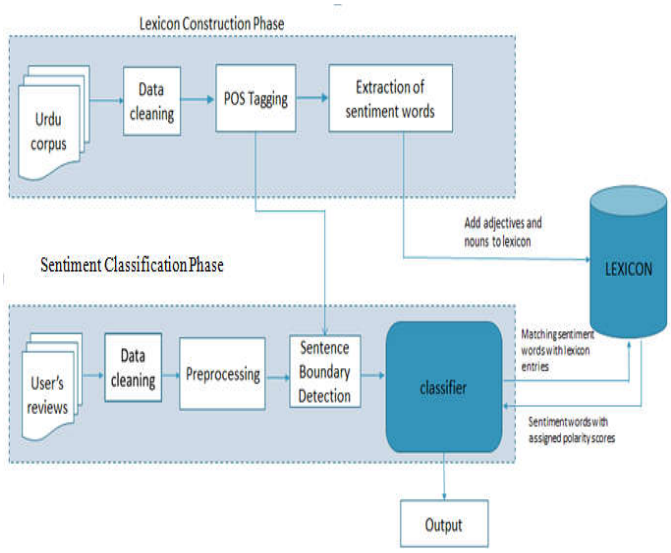


Figure 1: Urdu Sentiment Analyzer Flow Chart

The proposed Urdu Sentiment Analyzer works in two phases.

- Lexicon Construction Phase
- Sentiment Classification Phase

The lexicon construction phase involves preprocessing the corpus data and extraction of sentiment words from corpus and construction of lexicon using these words.

The sentiment classification phase performs the actual function of sentiment analysis i.e. mining the opinions as positive, negative or neutral. The input to this phase is taken from previous phase and consults the lexicon for making decision.

4.1. Sentence Boundary Identification

The proposed method is performed at sentence level. At this level, sentence must be a separate unit expressing a single opinion given by a single opinion holder. Therefore, a "#" sign is manually inserted at the end of each opinion in the collected corpus, to separate it from other opinions and make it a complete single unit.

4.2. Urdu Corpus Creation and Data Cleaning

A corpus of public opinions on Urdu news headlines has been collected to be used for sentiment analysis [4]. About 1000 opinions (positive, negative and neutral) have been collected from different domains including current affairs, politics, sports, health and entertainment from daily jang blog [4].
The corpus collected so far is in raw form; therefore, the corpus is preprocessed by removing unwanted information (opinion holder's names, addresses, date of posting and dashed lines), soft and hard spaces and opinions/words given in foreign language.

4.3. Parts of Speech Tagging

Part of Speech (POS) tagging is the process of assigning a tag showing the part of speech of a given word e.g. the symbol NN for a noun, ADJ for an adjective and PN for proper nouns. Part of speech tagging plays a vital role in enabling the algorithm to extract and match information using POS tags associated with the words.

After preprocessing the corpus, the data is tagged using tagger for Urdu data developed by [30]. Until this implementation work, the above mentioned tagger [30] was not officially released and was not publically available for personal use. Therefore, the corpus tagging has been performed by these authors of [30] themselves as a response to a request. This tagger shows the performance up to 88.74% accuracy level. However, proper noun (PN) tags are corrected manually to achieve better accuracy, which included the pronoun.

4.4. Lexicon Construction

Lexicon construction is an important phase of the proposed algorithm. The lexicon is constructed from the collected opinionated Urdu corpus. The lexicon consists of two tables i.e. one for nouns and second for adjectives. Words having NN (noun) or ADJ (adjective) POS tags are extracted from corpus and are transferred to their appropriate tables in the lexicon.

Frequencies are assigned to words in the lexicon. Words of lexicon are arranged according to the descending order of their frequencies. The importance of this arrangement of frequencies is to speed up the execution of algorithm as higher frequency words will appear in the start of the table and hence will be matched first. Polarities are also assigned to lexicon entries. A two scale polarity is used i.e. +1 for positive entry and -1 for negative entry. However the absence of adjective and nouns in a sentence gives a neutral opinion. Due to scarcely availability of Urdu opinionated data, the lexicon constructed in this work contains a total number of 316 distinct nouns and 133 distinct adjectives.

After the completion of the lexicon construction phase, the algorithm starts with the sentiment classification phase which performs the actual functionality of the sentiment analyzer. The detail of each step of this phase is given in the following sections.

4.5. The Classifier

A classifier is a programming module of sentiment analyzer which classifies the given opinion as being positive, negative or neutral. At this stage, the classifier takes an opinion as input, which is preprocessed with sentence boundaries identified; and examine this opinion for adjectives and nouns. If any text is identified with an adjective (ADJ) or noun (NN) tag, the classifier takes this piece of information and consults the appropriate table in the lexicon. If the text is tagged NN (or ADJ) then the item is searched in the nouns (or adjectives) table to find its matching entry in the lexicon. If match is found, its polarity is extracted.

Two counters are used for storing polarity value of words i.e. P variable for positive polarity and N variable for negative one. The polarity counter is incremented when its respective polarity is extracted. However, the polarities assigned in the lexicon can be altered by the presence of negation words, which need to be carefully handled. After solving the negation (Section 4.6.1), the appropriate counters are incremented. The entries are matched with lexicon and polarities are extracted until the "#" symbol is encountered, indicating the end of opinion. Finally, both the counters are compared and result is given depending on the value of the counters, i.e. if P counter value is greater than N counter value then the algorithm gives a "positive opinion" result otherwise "negative opinion". The classifier gives a "neutral opinion" for opinions which lack any adjective/noun or both.

However, the situation where P and N values are the same (i.e., P == N), then the algorithm considers the N variable and give a negative result. The reason behind this decision is the strong influence of negative opinion words over the positive opinions by examining the opinions in the collected corpus. Examples (6) and (7) justify this action [4].

(6) یہاں اچھےکام پر سزا ملتی ہے۔

Yaha achy kām par sazā milthi hai

"Good work is given a negative here".

(7) قادری صاحب ناکامی مبارک، عوام کو بے وقوف بنانے کا شکریہ۔
Qādri sahib nākāmi mubārak, awām ko bevaqvvf banāny ka shukriyā
"Congratulations on failing Qadri Sahib. Thank you for fooling the public".

In above mentioned examples the underlined terms are sentiment words and the presence of equal number of positive and negative sentiment words (P ==N) give negative results.

4.5.1. Handling Negation. Negation words alter the polarity of associated sentiment words. In Urdu language, 'نہیں' and 'نہ' act as negation words. The position of the negation words is very important. The current classifier is considering the negation words at one word distance i.e. before or after the target item. If negation word is present in any of these locations then the polarity value is altered otherwise it is taken as original one.

The negation at one word difference is chosen for correct sentiment classification. Exceeding the one word distance in most opinions gave wrong results. Consider the following examples [4] for this justification.

(8) نہیں نواز حکومت کی یہ پالیسی اچھی ہے
nahi nawāz hakumath ki yeh pālisi achi hai
"No, this policy of Nawaz government is fine".

(9)۔کچھ نہیں سب ڈرامے بازی ہے
kuch nahi sab dram-e bāzi hai
"It is nothing but a stage act".

(10)۔ایسی بدعنوان حکومت نہیں ہونی چاہیے
'aisi bad'unwān hakumath nahi honi chāhiye
"A corrupt government has no right to rule".

In examples (8), (9), and (10) negation words are present at two or more words distance from the underlined sentiment word. Therefore, the presence of these negation words does not have any effect on the polarity of these sentiment words. For example, if negation at two words difference is considered then in example (9) and (10), the underlined sentiment words polarity will be altered (from negative to positive) and hence will give wrong sentiment result.

However, examples (11), (12) and (13), taken from [4], show the presence of negation word at one word distance (before or after) from sentiment word. This occurrence of negation word affects the polarity of preceding or following sentiment words. Thus the fetched polarity of sentiment word is altered by the classifier.

(11) آپریشن مسلے کا حل نہیں، اس سے بہت سے لوگ متاثر ہونگے۔
Āpreshan masle ka haln nahi, is se bohuth se log muthāsir hongy
"Military is not the solution as it adversely affects people".

(12) سیاست دان قابل اعتماد نہیں
Siyāsathdān qābil –yi 'aithmād nahi
"Politicians cannot be trusted".

(13) ۔ بھارت کبھی پاکستان کا دوست نہیں ہو سکتا
Bahārat kabi Pākistān ka dost nahi ho sakthā
"India can never be a friend to Pakistan".

5. Experimental Results

This section explains the experiment performed on corpus of Urdu opinionated data taken from the domain of news [4].

In the field of sentiment analysis, the performance of a system varies from one domain to another and similar is the case with the level of analysis selected. The performance of an algorithm that is good in one domain may not be the same when switched to another domain. [9] Performed sentiment analysis on two different corpora (product reviews corpus and movie reviews corpus). The same analyzer achieved 72% accuracy in the movie review domain while it achieved 78% accuracy in the product review domain. Therefore, there is no such agreed upon performance level in this field.

5.1. Experiment 1.

This experiment is performed to observe the opinions which have been classified as positive or negative on the basis of noun (NN) only, adjectives (ADJ) only or both nouns and adjectives together. Figure 2 shows a comparison chart of sentiment words (NN, ADJ, NN+ADJ), highlighting the importance of Nouns in the field of sentiment analysis. The corpus has been divided into sections each having 100 opinions. Each section is analyzed and figures are collected on the basis of following:

1. Opinions which contain only nouns
2. Opinions which contain only adjectives
3. Opinions which contain both nouns and adjectives

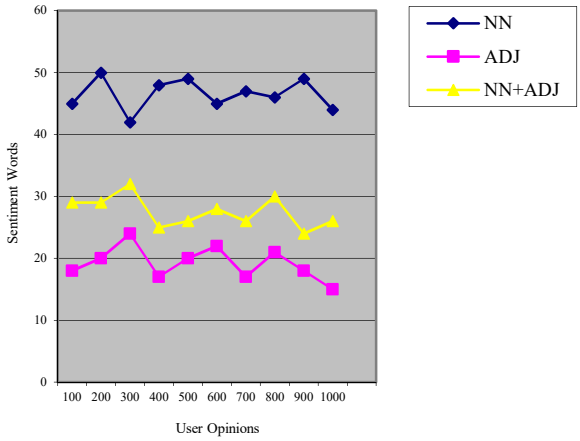


Figure 2: Sentiment Words Comparison

The experimental results show the usefulness of subjective nouns, where most of the decision is made by exploiting the nouns when there is lack of adjectives in data. However, nouns + adjectives also give better performance than only adjectives.

5.2. Experiment 2.

This experiment is performed to check the overall performance of the developed Urdu sentiment analyzer. A total of 1000 opinions have been analyzed consisting of positive, negative and neutral opinions from news domain. *Accuracy* is used as the system performance metric. It is the measure of how much close is the document classification suggested by our system to the actual sentiments present in the review. It is the percentage of correctly classified objects by the system, calculated by the following formula [32].

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Where

- A is the accuracy of sentiment classifier.
- TP (True Positive) is the number of positive sentences that are correctly classified as positive.
- TN (True Negative) is the number of negative sentences that are correctly classified as negative.
- FP (False Positive) is the number of positive sentences that are incorrectly classified as positive.
- FN (False Negative) is the number of negative sentences that are incorrectly classified as negative.

The figures mentioned in Table 1 have been taken from the corpus exploiting the current Urdu sentiment analyzer:

Table 1: System Performance

TP	TN	FP	FN	Accuracy
448	420	62	70	86.8%

5.3. Analysis.

Considerable amount of work is done in the field of Urdu sentiment analysis. In [9] the analysis is performed on a corpus of movie and product reviews and achieved 74-76% of accuracy. [28] Extended their previous work mentioned in [9] by associating the targets with SentiUnits and improved the performance to 85%. The targets are the attribute features for which an opinion is made. The work reported in [29] adopted a lexicon-based approach for bilingual sentiment analysis of tweets using bilingual dataset (English and Roman Urdu) and achieved 76% accuracy. The current research, that uses the lexicon-based approach for Urdu opinionated data, has achieved a better state-of-the-art performance with 86.8% accuracy (Table 1). Due to the unavailability of prior Urdu opinionated data, the current classifier was tested on the corpus constructed in this research work.

6. Conclusion

Discussion in the previous sections clearly reflects that sentiment analysis is a growing field of Natural Language Processing which aims at developing automated tools for categorizing opinions as positive, negative or neutral that greatly help timely and effective decision making. There are also indications that sentiment analysis for Urdu language is in its infancy and needs to be explored. Though these are big challenges for young researchers but there are great opportunities too.

This study explains the process of developing Urdu sentiment analyzer by using nouns, in the domain of news, as sentiment carriers in addition to the traditional use of adjectives which are considered as universal sentiment carriers. It is proved that the importance of nouns as sentiment expressions cannot be denied especially in news domain and thus nouns help improve the performance of sentiment analyzer in the absence of adjectives.

•

7. Future Work

In this study, the scope of sentiment expressions has been extended to subjective nouns in the domain of news. However, in future other domains can be considered for the presence of nouns as sentiment expressions e.g. product and movie reviews. Similarly, experiments on different domain-POS combinations can be carried out.

The lexicon entries need to be expanded to cover maximum sentiment words of Urdu language in order to improve the performance of the system in future. Other parts of speech can also be examined and can be included in the lexicon. A synonym column can be employed which contains synonyms of sentiment words, so that if exact matching is not available then the decision can be made on the basis of appropriate matching with the help of synonym entry in the lexicon.

8. References

[1] T. Nasukawa and J. Yi. "Sentiment analysis: Capturing favorability using natural language processing." In pro. *2nd international conference on Knowledge capture*, ACM, 2003.

[2] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. "Comparing and combining sentiment analysis methods," in proc. *of the first ACM conf. on Online social networks*, ACM , 2013.

[3] X. Ding, B. Liu, and PS. Yu. "A holistic lexicon-based approach to opinion mining. "In proc. *of the International Conference on Web Search and Data Mining*, ACM, 2008.

[4] www.dailyjang.com, [access date 25 July 2014].

[5] N. Durrani, and S. Hussain. "Urdu word segmentation." In proc. *of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[6] S. Hussain. "www. LICT4D. Asia/Fonts/Nafees_Nastalique." In proc. *12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore*. 2003.

[7] Y. Jo, and A. H. Oh. "Aspect and sentiment unification model for online review analysis."In proc. *fourth ACM international conference on Web search and data mining*, ACM, 2011.

[8] T. T. Thet, J. C. Na, and C. S. G. Khoo. "Aspect-based sentiment analysis of movie reviews on discussion boards." *Journal of Information Science.* 2010.

[9] A. Z. Syed, M. Aslam, A. M. Martinez-Enriquez, "Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits," In proc. *of the 9th Mexicon Int. Conf. of Artificial intelligence,* Mexico, 2010.

[10] A. Meena and T. V. Prabhakar. "Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis". Springer Berlin Heidelberg, 2007.

[11] J. Zhao, K. Liu, and G. Wang. "Adding redundant features for CRFs-based sentence sentiment classification." In proc. *of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2008.

[12] G. Gezici, B. Yanikoglu, D. Tapucu, and Y. Saygın. "New features for sentiment analysis: Do sentences matter." In *SDAD 2012 The 1st international workshop on sentiment discovery from affective data*, p. 5. 2012.

[13] K. Eguchi and V. Lavrenko."Sentiment retrieval using generative models." In proc. *of the 2006 conference on empirical methods in natural language processing*, Association for Computational Linguistics, 2006.

[14] B**.** Pang, L. Lee and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In proc. *of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, 2002.

[15] P. D. Turney. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In proc. *of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002.

[16] A. Khan. "Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and Blogs."*International Journal of Computer Science & Emerging Technologies* 2, no. 4, 2011.

[17] N. Farra, E. Challita, R. A. Assi, and H. Hajj. "Sentence-level and document-level sentiment mining for Arabic texts." In *IEEE International Conference on Data Mining Workshops (ICDMW), 2010*, IEEE, 2010.

[18] S. Kim and E. Hovy. "Determining the sentiment of opinions." In proc. *of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, 2004.

[19] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. "Structured models for fine-to-coarse sentiment analysis." In *Annual Meeting-Association For Computational Linguistics*, vol. 45, no. 1, p. 432. 2007.

[20] M. Hu, and B. Liu."Mining and summarizing customer reviews."In proc. *of the tenth ACM SIGKDD international*

conference on Knowledge discovery and data mining, pp. 168-177.ACM, 2004.

[21] K. Denecke. "Are SentiWordNet scores suited for multi-domain sentiment classification?." In *Fourth International Conference on Digital Information Management, 2009. ICDIM*, pp. 1-6. IEEE, 2009.

[22] B. Ohana and B. Tierney." Sentiment classification of reviews using SentiWordNet."In *9th. IT & T Conference*, p. 13. 2009.

[23] S. Mohammad, Cody Dunne, and Bonnie Dorr. "Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus." In proc. *of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 599-608. Association for Computational Linguistics, 2009.

[24] S. Mukund, and R K. Srihari. "A vector space model for subjectivity classification in Urdu aided by co-training." In proc. *of the 23rd International Conference on Computational Linguistics: Posters*, pp. 860-868. Association for Computational Linguistics, 2010.

[25] A. Z. Syed, M. Aslam, A. M. Martinez-Enriquez, "Sentiment analysis of Urdu language: Handling phrase level negation," In proc*. of the 10th int. conf. of Artificial intelligence,* Mexico, 2011.

[26] A. Z. Syed, M. Aslam, A. M. Martinez-Enriquez, "Adjectival phrases as the sentiment carriers in Urdu text,"*Journal of American Science,* vol. 7, no. 3, pp. 644-652, 2011.

[27] S. Mukund and RK. Srihari. "Analyzing urdu social media for sentiments using transfer learning with controlled translations." In proc. *of the Second Workshop on Language in Social Media*, pp. 1-8.Association for Computational Linguistics, 2012.

[28] A. Z. Syed, M. Aslam, A. M. Martinez-Enriquez, "Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text," *Artificial intelligence review*, pp. 1-27, March 2012.

[29] I. Javed, H. afzal, A. Majeed and B. khan, "Towards Creation of Linguistic Resources for Bilingual Sentiment Analysis of Twitter Data", 2014.

[30] B. Jawaid, A. Kamran, O. Bojar, "A tagged corpus and a tagger for Urdu," In proc*. of the 9th Int. Conf. on language processing and evaluation,* Reykjavík, LREC 2014, Iceland, 2014.

[31] B. Jawaid and O. Bojar,."Tagger Voting for Urdu." In *24th International Conference on Computational Linguistics*, p. 135. 2012.

[32] R. Prabowo and M. Thelwall. "Sentiment analysis: A combined approach." *Journal of Informetrics* 3, no. 2, p. 143-157, 2009.

[33] Subrahmanian, V. S.; Reforgiato, D. "AVA: Adjective-verb-adverb combinations for sentiment analysis." *Intelligent Systems, IEEE* 23, no. 4, pp. 43-50, 2008.

[34] Sokolova, M.; Lapalme, G. "A systematic analysis of performance measures for classification tasks." *Information Processing & Management* 45, no. 4, pp. 427-437, 2009

[35] Hatzivassiloglou, V.; McKeown, K. R. "Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning." In *Proc. of the 31st annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1993, pp. 172-182

[36] Riloff, E.; Wiebe, J. "Learning extraction patterns for subjective expressions." In *Proc. of the 2003 Conf. on Empirical methods in natural language processing*, Association for Computational Linguistics, 2003, pp. 105-112.

[37] Bloom, K.; Argamon, S. "Unsupervised extraction of appraisal expressions."In *Advances in Artificial Intelligence*, Springer Berlin Heidelberg, pp. 290-294, 2010.

[38] Chesley, P.; Vincent, B.; Li Xu; Srihari, R. K. "Using verbs and adjectives to automatically classify blog sentiment", *in AAAI symposium on computational approaches to analysing weblogs,* AAAI-CAAW, pp. 27-29, 2006.

[39] Benamara, F.; Cesarano, C. A.; Picariello; Recupero, D. R.;Subrahmanian V. S. "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," In *Proc. of the int. Conf. on weblogs and social media,* ICWSM, 2007.

A Management and Evaluation Framework for English to Urdu Translation

Aneeta Niazi, Saba Urooj
Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
firstname.lastname@kics.edu.pk

Abstract:

The success of a translation project is largely dependent on the overall project management and quality assurance of the translated text. In this paper, we present a framework for carrying out English to Urdu translation projects systematically and efficiently. Different processes involved during the life cycle of a translation project have been explained in detail. To make the translation process significantly efficient, we have developed automatic systems for managing and evaluating translations. We have implemented the framework on 2164 sentences. A detailed analysis and classification of the glossary usage errors have been carried out on 540 sentences. We have obtained a very timely and high quality translation output at the end of the project, indicating the robustness of the presented framework.

1. Introduction:

Translation is defined as the communication of the meaning of a source-language text by means of an equivalent target-language text [1]. The translation process is generally assumed to be a very simple activity, in which translators carry out the task with the help of references, such as dictionaries, thesaurus, online resources etc. [2]. However, the translation activity has undergone significant changes in the last few decades, which has made a translation project more than just a translation task [3]. Over the years, translation has become an economic activity, which is concerned with the output of a specified product, based on existing corpora and information technology [4]. As a result of this evolution, translators are now expected to deliver larger translation volumes within shorter deadlines [2].

Urdu is the national language of Pakistan, as well as one of the twenty three official languages of India [5]. There are around 11 million native speakers. Urdu, and around 105 million speakers who speak it as a second language [6]. The official language of Pakistan is English, since the country was under British rule as a part of British India [7]. But recently, the supreme court of Pakistan has ordered the government to adopt Urdu as an official language [8]. This shift of language requires the translation of all official documentation from English to Urdu. Translation is also a basic requirement for making English content available for the Urdu speaking population, who do not understand English language. A considerable amount of translated data is also required for the development of English to Urdu parallel corpora, which is a prerequisite for many NLP applications, such as machine translation, multilingual information retrieval [9], automatic construction of lexicons and comparative analyses of the structure of languages [10]. Such translation projects may require very large volumes of English text to be translated into Urdu, which are difficult to manage by translators. The quality assurance of the translated text also becomes a very challenging task, when large amount of data is required to be translated.

In this paper, we have discussed different tasks that are involved in the life cycle of a translation project in detail. We present a semi automated management and evaluation framework for the efficient and accurate translation of English text into Urdu.

2. Literature Review:

In the recent years, the translation activity has become very popular for cross cultural understanding and communication, and has emerged as a very significant field of study. Dunne [11] has pointed out

the following eight characteristics for the overall success of a translation project as a learning activity:

i. Providing clear instructions.
ii. Managing people (by creating a schedule and monitoring).
iii. Quality assessment to clarify expectations.
iv. Reliable documentation.
v. Risk management by reducing uncertainty.
vi. Efficiency.
vii. Energy and focus.

In order to complete a translation project successfully, a set of guidelines should be provided to the translators for clarifying client's requirements and the purpose of translation. Such guidelines should help the translators in decision making, and give them an insight about the user demands. The performance of the individual team members should be evaluated for managing the translation deliverables according to committed deadlines. Regular review of the translated text should be carried out, and feedback should be given to the translators to maintain the quality of translation according to client's expectations. The client should be informed about the progress of the project by documenting interim reports. The chances of uncertainty and surprises should be reduced through regular quality assessment and better communication with the client and the team. The project should be made efficient by extracting glossaries for specific terminologies and developing a detailed plan at the beginning of the project. The team will be able to work with great energy and focus when the issues regarding purpose, terminology, communication, deadlines and deliverables will be resolved with the development of guidelines, glossaries and a realistic project plan.

Lauffer [12] has carried out a detailed analysis of the translation process. A translation process can be broken down into three general stages:

i. Understanding and reasoning
ii. Searching
iii. Revising

During the "Understanding and reasoning" stage, the translators read and familiarize themselves with the source text to produce translated text. The translation is usually done at word or sentence level. The translators have to undergo a lot of decision making about the structure of the translated text, to ensure consistency. During the "Searching" stage, the translators look for words, terms and expressions using resources such as dictionaries and glossaries. The "Revising" stage consists of re-reading the translated text for accuracy, grammatical structure, word order, idiomatic language, lexical choices, syntax and the flow of the text.

Sprung [13] has emphasized on the need of organizing translation projects for time and budget management. Other than the linguistic processes involved in a translation project, processes such as preparing products for translation, extracting sentences to be translated, scheduling, testing and evaluation of the translated text are also very critical for the successful completion of a translation assignment. Software engineering and service management can play a very important role for delivering a high quality of translated text and making the translation process efficient.

Pérez [4] has outlined the framework for translation project management. A translation project's life cycle consists of different phases, including commissioning, planning, ground work, translation and wind up. The ground work consists of terminology extraction and research, segmentation of source text into sentences, and the preparation of work packages. The translation phase is usually carried out with the help of Computer Aided Tools (CAT) and translation memories. During the wind up phase, the consistency and completeness of the translated text are evaluated.

Makoushina [14] has conducted a survey for comparing the approaches and functions of quality assurance CAT tools, available for conducting translation projects. The tools included in the survey are Deja Vu, SDLX QA Check, Star Transit, SDL Trados QA Checker, Wordfast, Error Spy, QA Distiller, and XBench. The survey results have shown that over 81% of the respondents use QA Automation tools for translation. The survey demonstrates an urgent need of providing support for more languages, encodings and file formats by CAT tools. These tools lack the optimal support that is required for right to left languages, such as Urdu. The users generally lack skills required to perform additional customization for achieving higher level of error detection. The error level of the tools is very high, and users have to spend a lot of time for deciding whether a reported error needs correction or not.

Moses is a machine translation tool, used for automatic translation that requires a large parallel corpus for training. The existing English to Urdu parallel corpora are not large enough to train a system for automatic machine translation.

Sere [2] has described translation project as a succession of fast-paced activities, which requires issues to be resolved as rapidly as possible in order to meet the deadlines. A translation project is considered a failure if it is not delivered to the client on time or the quality of the translation does not meet the client's requirements.

There are also some challenges faced by the translators during the process of translation. These challenges include language specific challenges and

are cited by [15] during the process of mapping Urdu words from Urdu WordNet to Princeton WordNet. These challenges include morphological challenges e.g. Causitivization in which causative infixes such as لا/lɑ: and وا/ vɑ: change the verb into its causative form such as سونا (so: nɑ:/sleep) is changed into سلانا (sʊlɑ:nɑ:/ to make someone sleep). These morphological causative are not found in English hence making it difficult for translators to find a single translation of such verbs from source to target language. The translation of such verbs can result into multiple translations by the translators and can affect the quality of the output by causing inconsistency. There are other similar syntactic and semantic challenges cited in the work [15]. These issues can be solved by identifying them at the stage of guidelines development and setting a formula for the translations of such words.

3. Methodology:

For the translation of English source text into Urdu, we have developed a framework for the overall translation project management, as well as the quality evaluation of the translated text. The presented framework comprises of a list of sequential tasks, in order to carry out the translation project in an efficient and orderly fashion.

The presented translation framework consists of the following tasks:

i. Acquisition of the source data to be translated.
ii. Glossary extraction.
iii. Development of translation guidelines.
iv. Word counts computation.
v. Division of source sentences into work packages.
vi. Translation.
vii. Automatic evaluation of translated work packages.
viii. Automatic evaluation of glossary usage in translated text.
ix. Manual review of translated text.
x. Automatic reformatting of the translated text according to source text.

The detail of these tasks is given as follows:

i. Acquisition of the Source Data to be Translated:

As a first step, the English text to be translated is acquired. The source data may consist of a single or multiple folders, containing a single or multiple files. The text inside the source files is usually in the form of paragraphs containing multiple sentences. Sometimes, there are empty lines present in between these paragraphs. For the presented framework, the files containing source text should be in .txt format.

ii. Glossary Extraction:

A detailed analysis of the source data is carried out by translators for the manual extraction of glossary items. The specific terms in the data are included in the glossary, along with their standard Urdu translations. This is a very critical step, as it involves a lot of decision making for selecting the most appropriate translations of glossary items.

The glossary items can be divided into the following categories:

a) Frequently occurring terms (having greater than or equal to 10 instances). For example, the word "title" occurs 225 times in the source text. It is included in the glossary with a standard translation, "عنوان"(*unwaan*). The glossary usage will ensure that all the 255 instances get consistently translated as "عنوان"(*unwaan*), instead of any other alternative translation.

For example, consider the translation of the following two sentences:

English:

"The guests arrived at the closing ceremony."

"They were closing the windows."

Urdu:

"مہمان اختتامی تقریب پر پہنچ گئے۔"(*mehmaan ikhtataami taqreeb per puhanch gae.*)

b) Terms having multiple senses. For example, the word "approve" occurs 8 times in the source text. It has two senses i.e. "Officially agree to or accept as satisfactory" and "archaic Prove; show" [16]. In the glossary, it is included with the standard translation, "توثیق"(*tauseeq*), meaning "Officially agree to or accept as satisfactory", based on the context of the source text.

There can be terms occurring multiple times, and having multiple senses in the source text. "وہ کھڑکیاں بند کر رہے تھے۔"(*woh khirrkian bund ker rahe thay.*)

In the above example, the word "closing" has been correctly translated as "اختتامی"(*ikhtataami*) and "بند"(*bund*) in two different sentences. For the word "closing", the glossary contains both "اختتامی"(*ikhtataami*) and "بند"(*bund*) as standard translations, separated by a comma.

c) Terms without standard Urdu translations. For example, the term, "under the patronage" occurs only 2 times in the source text. Since, there is no standard Urdu translation for this term, the translators themselves have coined its standard translation, "زیر سرپرستی"(*zair-e-sarparasti*), and added it in the glossary.

The extracted glossary is evaluated by language experts for finalization. The format of glossary is given as follows:

English	Urdu
Title	عنوان
Core	مرکزی، بنیادی
Responsible	ذمہ دار

iii. Development of Translation Guidelines:

A set of guidelines is developed for the translators by language experts in order to ensure consistency in translated text. These guidelines are developed in accordance with the requirements of a particular project, and can vary for different projects. The guidelines provide clarification related to the following issues:

a) The usage of diacritics. For example, "مکمل"(*mukammal*)" should be used in the translated text instead of "مکمّل"(*mukammal*, containing a diacritic "ّ").
b) Transliteration of English words (such as proper nouns etc.). For example, "United Group", being a proper noun, should be transliterated as "یونائٹڈ گروپ"(*united group*), instead of being translated as "متحد گروہ"(*muttahid groah*).
c) Spelling consistency. For example, "کے لیے" (*ke liay*, with a space between "*ke*" and "*liay*") should be used in the translated text instead of "کیلیے" (*keliay*, with all joined characters).
d) Date number format. For example, "11/03/2005" should be translated as "۲۰۰۵/۰۳/۱۱", instead of "۱۱ مارچ،۲۰۰۵ء".
e) Numerical representations. For example, the alphanumeric characters "a, b, c" should be transliterated as "اے، بی، سی".

An identifier is assigned to each sentence e.g. S1, S2, S3 etc., and a sentence log file is automatically generated. Each row of the sentence log file contains a sentence identifier, along with its source sentence and its corresponding source file and folder information. The format of the sentence log file is given as follows:

f) Text that will not be translated, such as abbreviations, web addresses, email addresses etc. For example, "UET", "www.abc.com" and "xyz@yahoo.com" should not be translated.
g) Treatment of special symbols. For example, "&" should be translated as "اور"(*aur*), "?" should be translated as "؟" and "*" should not be translated.
h) Consistency of the translated sentence structure with the source sentence structure i.e. if a source sentence is in passive voice, its translation should also be in passive voice. For example, the sentence, "A road was constructed." should be translated as "ایک سڑک تعمیر کی گئی۔"(*aik sarrak taameer ki gae*) instead of "انہوں نے ایک سڑک تعمیر کی۔" (*unhoon ne aik sarrak taameer ki*).
i) Usage of glossary i.e. if a glossary item is found in the source text, it should only be translated as its standard glossary translation. For example, a glossary word "mission" having a standard glossary translation "مقصد"(*maqsad*), should not be translated as "مہم"(*mohim*).

iv. Word Counts Computation:

In order to manage the translation assignment, an estimate about the amount of work required is very critical for the timely completion of task. The knowledge about the number of words in the source text is also very important for the budget allocation. For this purpose, the total number of words present in the source data are automatically computed, by segmenting the text on white spaces and counting the number of segments for each source file.

v. Division of Source Sentences into Work Packages:

The source text is segmented into sentences, on the basis of carriage return and punctuation marks, such as ".","?","!" etc. Sentence segmentation is important because we want to ensure that all sentences in the source.

Sentence ID	Sentence	Source Folder Name
S1.	Page Title:	Profile
S2.	About EPG	Profile
S30.	Banner Title:	Services

After sentence segmentation, the source data is automatically divided into work packages. Each work package may contain a single or multiple text files. The number of text files inside each work package, as well as the number of words contained inside each text file are variable entities, that can be set according to the project plan. Each text file inside a work package is assigned sentences according to the word count limit. For example, for a project having a word count of 20,000 words, 5 work packages are generated. Each work package contains 8 text files, and each text file contains around 500 words.

We have developed a naming convention in order to keep a record of the work packages and the text files contained in them. The naming convention for the work packages is given as follows:

PKG<Package Number>_F<Starting File Number>-<Ending File Number>_WC<Package Word Count>_T<Translator Identification Number>

For example, work package number 2, having 9 to 16 files, and a word count of 1000 words is to be assigned to translator number 3. The name of the package will be:

PKG2_F9-16_WC1000_T3

The naming convention for text files is given as follows:

F<File Number>_S<Starting Sentence Number>-<Ending Sentence Number>_WC<File Word Count>_T<Translator Identification Number>

For example, file number 3, having 20 to 43 sentences, and a word count of 500 words is to be assigned to translator number 1. The name of the file will be:

F3_S20-43_WC500_T1

A management log file is maintained for keeping a record of the dates on which the work packages are "assigned to" and "received from" each translator.

vi. Translation:

The work packages are assigned to a team of expert translators, who use Omega T[9] for translation. Omega T is a translation memory tool that is used for ensuring overall consistency in the translation. Each sentence is individually translated, and the context of the sentence is also taken into consideration while translating. During the translation process, the translators also give their opinion about the inclusion of words in the glossary. The glossary is updated after the consultation of language experts accordingly.

vii. Automatic Evaluation of Work Packages:

After the translation of a work package is completed, it is automatically evaluated to ensure that the translated package received from a translator contains the complete translation of the assigned source package.

The automatic work package evaluation system checks for the following errors:

a) Missing files inside work packages.
b) Extra files inside work packages.
c) Wrong files inside work packages.
d) Missing sentences in files.
e) Extra sentences in files.
f) Missing sentence identifiers.
g) Duplicated sentence identifiers.
h) Wrong sentence identifiers.
i) Missing translations.

An error log file is generated by the work package evaluation system, and the indentified errors are fixed accordingly. In case the translation of a sentence is missing, the package is returned to the translator along with feedback, until all sentences are translated.

viii. Automatic Evaluation of Glossary Usage in Translated Text:

In order to ensure the usage of glossary by the translators, an automatic evaluation system has been employed. Each source sentence is checked for the presence of glossary items. If any glossary items are found in a source sentence, the corresponding translated sentence is checked for the presence of the glossary item translation, as mentioned in the glossary.

It has been observed that, sometimes, a glossary item is not exactly translated as its glossary translation, but as an inflected form of the root word of glossary translation, depending upon the context. For example, we have a glossary item "stories", with a glossary translation "کہانیاں"(*kahanian*), which is an inflected form of the root word "کہانی"(*kahani*). Consider the following case of English to Urdu translation:

English:

[9] http://www.omegat.org/

"It is a collection of stories."
Urdu:
"یہ کہانیوں کا ایک مجموعہ ہے۔" (*ye kahanioun ka aik majmua hay*).

The translated Urdu sentence contains the word "کہانیوں"(*kahanioun*) instead of "کہانیاں"(*kahanian*) as a translation of "stories". The word "کہانیوں"(*kahanioun*) is also an inflected form of the root word "کہانی"(*kahani*). Therefore, the automatic evaluation system should consider the presence of an inflected form of the root word of a glossary translation as correct glossary usage.

This has been done by incorporating an Urdu stemmer[10] to get the root word of a glossary translation. If an inflected form of the root word is found in the translated sentence, its glossary usage is marked as correct. Otherwise, the glossary usage is marked as incorrect.

ix. Manual Review of Translated Text:

After ensuring the glossary usage in the translated text, a manual review pass is carried out by a team of language experts, who critically evaluate the translation of each sentence individually. The language experts finalize the translation by making the required changes in the translated text. In order to deliver high quality translation, a detailed feedback is given to the translators by the language experts, based on their performance.

x. Reformatting of the Translated Text according to Source Text:

The finalized translated text needs to be formatted according to the source text received from the client. An automatic reformatting tool has been developed, that takes the sentence log file generated in step (v) as input. As the formatting information of each source sentence is recorded in the sentence log file, it is utilized to format the corresponding translated text accordingly. The reformatting tool also generates output folders in accordance with the source data, and places the reformatted translated text files, having the same names as their corresponding source files, inside their respective folders.

After this step, the translated Urdu data is ready for shipment to the client.

A set of guidelines have been developed, covering all the issues related to the consistency and completeness of translation. With the help of automatic counts computation utility, it has been found that 2164 sentences contain a total of 30334 words. The sentences have been automatically divided into 8 work packages. Each work package contains 8 text files, and each text file contains around 500 words. These work packages have been assigned to a team of expert translators. The translated work packages have been automatically evaluated for completeness and consistency of the translated text, followed by a detailed manual evaluation by language experts.

The language experts have reported that the translation guidelines have a very positive impact on the quality of the translated text. The following issues have been correctly addressed by the translators, due to the clarification provided in the translation guidelines:

i. The date number format has been found consistently correct in the translated text.
ii. The numerical representations are consistent throughout the translated text.
iii. The abbreviations, web addresses and e-mail ids have not been translated, as per project requirement.
iv. The special symbols are treated correctly in the translated text.
v. The sentence structures of the translated sentences are in accordance with the source sentences.

After translation, the following issues have been detected in the work packages received from translators by the automatic evaluation system:

Error Type	Error Count
Missing sentence identifiers	26
Wrong sentence identifiers	12
Duplicate sentence identifiers	2

These errors have been fixed in the work package files, and the translated data is automatically evaluated for glossary usage.

For analyzing the output of automatic glossary usage evaluation in the translated text, we have taken a subset 540 sentences. The following table shows the data counts that have been obtained after the automatic evaluation for glossary usage:

10 http://www.cle.org.pk/software/langproc/UrduStemmer.htm

Glossary Items in Source Data	Correct Glossary Translations in Translated Data	Exact Glossary Translations in Translated Data	Inflected forms of the Glossary Translation root words in Translated Data	Glossary Usage Errors
612	429	351	78	183

We have carried out an error analysis of the 183 glossary usage errors, and found out that these errors can be categorized as following:

1. Transliteration of English words into Urdu instead of translation. For example, the word "vision" is translated as "وژن"(*vision*) instead of its glossary translation "تصور"(*tasawur*).
2. Translation of English words into Urdu instead of transliteration. For example, the word "network" is translated as "جال"(jaal) instead of its glossary translation "ورک نیٹ"(network).
3. Similar but different translation of English words. For example, the word "landscapes" is translated as "مناظر"(*manazir*) instead of its glossary translation "قدرتی مناظر"(*qudrati manazir*).
4. Different spelling of Urdu translation. For example, the word "license" is translated as "لائسنس"(*license*) instead of its glossary translation "لائسینس" (*license*, containing an additional character "ی" in Urdu spelling).
5. Addition of diacritics in translation. For example, the word "responsible" is translated as "ذمّہ دار" (zimmadar, containing a diacritic "ّ") instead of its glossary translation "ذمہ دار"(zimmadar).
6. Different translation of English words. For example, the word "joint" is translated as "درمیان"(*darmiyan*) instead of its glossary translation "مشترکہ"(*mushtarka*).
7. English word used in a different sense than its glossary item translation. For example, the word "cover" has been translated as "پورا کرنا"(*poora kerna*) instead of its glossary translation "لفافہ"(*lifafa*).

The following table shows the number of occurrences of the each of the above mentioned glossary usage errors in the translated text:

Error Number	Number of Occurrences
1	78
2	1
3	13
4	27
5	3
6	54
7	7

From the table, it can be observed that the most glossary usage errors have occurred by the transliteration of English words instead of translating them into Urdu.

From the manual review of the translated text, it has been observed that the transliteration of English words instead of translation can be further categorized as follows:

i. Correct transliteration
ii. Correct transliteration with inconsistent spellings
iii. Incorrect transliteration

The detail of these categories is given as follows:

i. Correct Transliteration:

At some instances, the transliteration has been used instead of translation, depending upon the context of the source text. For example, the glossary translation of the word "philately" was "ٹکٹ"(ticket), but in the context, it was used as a proper noun, "Philately Club". According to the translation guidelines, all proper nouns are to be transliterated. Therefore, the transliteration of "Philately Club" as "فلاٹیلی کلب"(*philately club*) has been considered as correct translation by the language experts, although it is an incorrect glossary usage.

ii. Correct Transliteration with Inconsistent Spelling:

The translated text contains such instances of incorrect glossary usage, where the transliteration is correct based on the context of the source text, but the spellings of the transliterated words are not consistent throughout the translation. For example, the word, "exchange" has a standard glossary translation "تبادلہ"(*tabaadla*). In the context, it has been used as "Wall Street Exchange", which is a proper noun. The translators have transliterated it as "اکسچینج"(*exchange*) at some places, and "ایکسچینج"(*exchange*, with an additional "ی" character

spellings as correct translation, and all other spellings as errors. In case of the proper noun "Exchange", "ایکسچینج"(*exchange*, with an additional "ی" character in spelling) has been marked as correct translation, whereas "اکسچینج"(*exchange*) has been marked as an error.

iii. Incorrect Transliteration:

There are certain glossary usage errors where the English words have been incorrectly transliterated instead of using their standard glossary translations. For example,"foreign exchange" has been transliterated as "فارن ایکسچینج"(*foreign exchange*), instead of using its glossary translation, "زرمبادلہ"(*zar-e-mubadla*). Since it is not a proper noun, the language experts have marked all such cases as translation errors.

During manual review, It has been observed that there are certain English words that occur in a different sense in context than their glossary translations. For such cases, the glossary has been updated by the addition of glossary translations of all senses.

The language experts have suggested 156 additions in the glossary. The updated glossary consists of 474 words.

After the completion of manual review process, the translated text has been automatically formatted according to source text. The reformatted translated text files have been manually evaluated for formatting errors. The manual evaluation showed that the automatic reformatting system does accurate formatting of the translated text, and does not produce any additional errors.

4. Conclusion:

The presented translation management and evaluation framework can be implemented for the translation projects involving large volumes of data. The automatic evaluation of glossary usage has a significant impact on ensuring the consistency of the translated text. The development of glossary for a translation project is an evolutionary process that requires incremental input from the translators and feedback from the reviewers. A lot of manual effort can be saved by implementing the automatic reformatting system that gives an accurate output.

5. Future Work:

In future, the developed framework can be implemented for the translation of English content into other local languages of Pakistan such as Punjabi, Sindhi etc. An automatic translation framework can be developed for managing and evaluating translations of multiple languages. The presented framework can be extended for different file formats, such as xml, html, php. etc.

6. References:

[1] N. Bhatia, *The Oxford Companion to the English Language*, Oxford University Press, 1992.

[2] K. Sere, "*Risk Management in Translation Projects: Study ans Survey Results*", Institut Supérieur de Traducteurs et Interprètes, Brussels, Belguim, 2015.

[3] M. Ballard, "*Histoire de la traduction. Reperes historiques et culturels*", De Boeck Supérieur, Brussels, Belgium, 2013.

[4] C. R. Pérez, "Translation and Project Management," *Translation Journal,* [Online]. vol. 6, no. 4, 2002, Retrieved (May, 15, 2016).

[5] S. A. Khan, W. Anwar, U. I. Bajwa and X. Wang, "A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language," in proc. *24th International Conference on Computational Linguistics*, Mumbai, India, 2012.

[6] B. F. Grimes, "*"Pakistan". Ethnologue: Languages of the World*", Summer Institute of Linguistics, Dellas, Texas, USA, 2000.

[7] T. Rehman, "Language policy, multilingualism and language vitality in Pakistan," in *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology (Trends in Linguistics. Studies and Monographs 175)*, De Gruyter Mouton, 2006, pp. 73-104.

[8] "Supreme Court of Pakistan (Original Jurisdiction)," [Online]. Retrieved (May, 15, 2016). Available: http://www.supremecourt.gov.pk/web/user_files/File/Const.P._56_2003_E_dt_3-9-15.pdf.

[9] K. P. Scannell, "Applications of parallel corpora to the development of monolingual language technologies," Preprint, http://borel. slu. edu/pub/ccgb. pdf. 2005.

[10] J. Hallebeek and V. Spaans, "English Parallel Corpora and Applications," *Cuadernos de Filología Inglesa,* vol. 9, no. 1, pp. 111-123, 2000.

[11] E. Dunne, "Project as a Learning Environment," in *Translation and Localization Project Management*, Springer, 2013.

[12] S. Lauffer, "The Translation Process: An Analysis of Observational Methodology, "*Cadernos de Tradução",* 2002.

[13] R. C.Sprung, "Translating into Success. Cutting Edge Strategies for Going Multilingual in a Global Age," *American Translators Association Scholarly Monograph Series,* vol. 11, 2001.

[14] J. Makoushina, "Translation Quality Assurance Tools:

Current State and Future Approaches," in proc. *29th International Conference on Translating and the Computer*, London, UK, 2007.

[15] A. Zafar, A. Mahmood, S. Shams and S. Hussain, "Structural Analysis of Linking Urdu WordNet to PWN 2.1," in *Conference on Language and Technology 2014 (CLT 14)*, Karachi, Pakistan, 2014.

[16] "http://www.oxforddictionaries.com/definition/english/approve," [Online]. Retrieved (May, 15, 2016).

[17] A. H. Homiedan, "Machnie Translation," *Journal of the King Saud University,* vol. 10, 1998, pp. 1-21.

Handcrafted Semantic Hierarchy to Develop Urdu WordNet

Palwasha Jogezai, Ayesha Zafar
Kinnaird College for Women, Lahore
palwashajogezai@gmail.com, ayeshazafarsultan@gmail.com

Abstract

This research paper highlights the issues and challenges faced during the development of semantic hierarchy of Urdu WordNet. 118 developed hierarchies are analyzed by using database World Atlas of Language Structure alongside the resources used for previous researches on Urdu WordNet [7] [8]. Problematic hierarchies are discussed with solutions supplied.

1. Introduction

WordNet is an online lexical database which is developed on the basis of contemporary theories of psycholinguistics of human lexical memory, at Princeton University. Its basic concepts and purpose of development are shared in the Five Papers on WordNet [1]. The construction of Princeton WordNet (PWN) [2] inspired the development of lexical databases for other languages [3] [4] [5] [6]. Urdu WordNet [7] was also developed following its pattern in order to align with linguistic, cultural and other contexts in Pakistan. Further, the developed senses of Urdu WordNet (UWN) were aligned to PWN 2.1 [8].

This research focuses mainly on building hypernym-labeled noun hierarchies of Urdu WordNet because the detail about hypernyms and hyponyms semantic relation provides additional disambiguating information. The lexical resource was manually constructed and it is helpful for natural language processing tasks such as machine translation and information retrieval system.

The paper is organized in the following sections. Section 2 reviews the current literature regarding WordNet hierarchies based on semantic relations. Section 3 describes the approach of developing Urdu WordNet semantic hierarchies. Sections 4 presents the challenges and solutions faced during the process. Finally, Section 5 concludes the paper by reporting the future work required in this direction.

2. Literature Review

The lexicon division of WordNet has been done into five classes, i.e. nouns, verbs, adjective, adverbs and function words which makes it different from a standard dictionary. Similarly, it incorporates the details of semantic relations [9]. For example, hyponymy/hypernymy semantic relations are represented in the creation of WordNet. Semantic relation that occurs between word meanings is hypernymy/hyponymy: e.g., {tree} is hypernym of {plant}, {plant} is a hypernym of {maple}. It's not like the lexical relations that occur between word forms are synonymy and antonymy. It is variously named subordination/superordination, subset/superset, or the ISA relation [9]. Hyponymy generates a hierarchical semantic organization that is duplicated in the noun files by the use of labeled pointers between sets of synonyms (synsets) [25].

In conventional dictionaries the definition of tree, for example, doesn't include information that trees possess roots. There is no information regarding its coordinate term. Similarly, the information related to its kinds and features are also not available in common dictionaries. All such information is missing due to economic pressure to minimize redundancy [9]. The knowledge about such semantic relations is useful in NLP applications: e.g., reformulation of query [10] [27], addressing query [11], summarization of written text [12], classification [13]. There are two categories in which the task on concept hierarchy learning [14] is done: Harris [15] distributional hypothesis & Hearst's [16] lexical patterns. The basic order imposed on nouns semantic memory is a tree, not circular form, meaning that lexicographers give tree represented graphically. So the construction of lexical tree can be done by adopting superordinate terms: oak – tree – plant – organism, and according to Miller et al., this can be read as "is a" or "is a kind of." This shows that a hierarchy is going from upward to limited generic terms from huge amount of specific terms at the bottom [9]. Hierarchies are said to be providing the "conceptual skeletons" for nouns; say for example the knowledge about some specific nouns is found to

be hung onto this tree like structure e.g. Christmas tree [9].

In the field of computer science such hierarchies are known by the term inheritance systems [9]. Psycholinguists assume that that comparison of superordinates cannot be done with hyponyms [17]. However, Quillian claimed explicitly that inheritance system is formed through the lexical memory of individuals for nouns [18] [19]. Collins and Quillian reported experimental tests, assuming hierarchical levels number can be identified when the two meanings are taken apart [9] [20]. Cognitive scientists and linguists came to conclusion that Quillian was not right, inheritance system is not how semantic memory is organized [21] [22] [23]. Miller and Charles indicated that there is no difference in meaning of word, but it's in the usage of word or that the distance is established semantically, not pragmatically [24].

The hierarchical principle can be construed with assumption that the existence of all nouns is in one hierarchy [9]. Theoretically, {entity} is placed at the bottom and {object, thing} are placed as nearest subordinates and then continuing towards specific meanings. However, practically, it fails to convince. The solution lies in partitioning the nouns due to various advantages one being less size occupied and the possibility of having the different files assigned to lexicographers for writing and editing. But, again, the problem is to choose the primitive semantic components. This issue of having the possibility of combination of adjective-noun happening was resolved by Philip N. Johnson-Laird, being discussed in the revised edition of Miller et al. with rationale [9]. This is one of the major challenges being faced even while development of Urdu WordNet, where levels didn't go any deeper.

3. Research Methodology

The manual constructions of 118 hypernym-labeled noun hierarchies of Urdu WordNet were done by incorporating both expand and merge approaches. The hierarchies of hypernym-labeled nouns in PWN 2.1 were kept as a model [25]. Moreover, World Atlas Language Structures was also used as a supporting resource. Urdu WordNet 1.0 Wordlist [26] was used as corpus and high frequency nouns from Urdu news websites were selected. The following steps were followed during the process.

1. Urdu Nouns were chosen from the corpus.
2. The selected nouns were mapped with PWN 2.1 Sense ID.
3. Urdu noun hierarchies were constructed following the PWN 2.1 hypernyms.
4. Missing Urdu senses were added to complete the hierarchies.
5. Shallow Senses (of the hierarchies) are translated from PWN 2.1 to complete the Urdu hierarchies.
6. Urdu language resources Qamos-e-Mutradifat and Urdu Lughat were used in order to check the accuracy of the developed hierarchies.
7. Typological information (to confirm the hypernyms hierarchy interruption) was done from World Atlas Language Structures dataset.

The example of a complete hierarchy of Urdu word کام has been shown in **Figure 1**.

Eng ID	Eng Word	Category	Concept	Example	Synset
00708623	job	noun.act	روزمرہ یا مقررہ وقت کا کام	وہ سارا دن گھر پر کام کرتی رہتی ہے	کام
00708412	duty	noun.act	کام جو اتفاقاً یا اسوقاً کرنا ضروری ہو	اُس کی ڈیوٹی آخر سے پانچ بجے تک ہے	ڈیوٹی، فرض، منصبی، خدمت، کام، نوکری
00570312	work	noun.act	کچھ کرنے کا عمل یا کام	تمام لوگوں نے کار خیر میں بڑھ چڑھ کے حصہ لیا	کار، کام، عمل
00403481	activity	noun.act	قوت عمل، سرگرمی، حرکیت	ہمیشہ تبت سرگرمیوں میں حصہ لینا چاہیے	فعالیت
00029085	human action/ act	noun.Tops	انسان کا کسی کام کو کرنے کا عمل	اس سارے عرصے میں کوئی انسانی سرگرمی دیکھنے میں نہیں آئی	انسانی سرگرمی
00028105	event	noun.Tops	واقعہ جو کسی وقت کسی جگہ وقوع پذیر ہو	اس واقعہ کے بعد وہ کبھی نظر نہیں آیا	واقعہ، ماجرا، ایونٹ
00022007	psychological feature	noun.Tops	کسی جاندار کی ذہنی تبدیلی کا ایک پہلو	اس کے اس برتاؤ کے پیچھے کون سی نفسیاتی خاصیت کارفرما ہے	نفسیاتی خاصیت
00002236	abstraction	noun.Tops	نظریہ یا خیال جو خاص مثالوں کے مشترکہ عوامل پر مشتمل ہو	تجرید کا عمل انسانی ترقی کے لیے ضروری ہے	تجرید
00002119	abstract entity	noun.Tops	چیز جس کا وجود صرف خیالوں میں ہو، وہ چیز جو صرف غیر مرئی شکل پر موجود ہو	غیر مرئی اشیا نظر نہیں آتیں	غیر مرئی شے
00001740	entity	noun.Tops	چیز جو اپنا وجود رکھتی ہو چاہے مادی یا خیالی	کسی شے کا وجود ہونا یا نہ ہونا خیالی دونوں صورتوں میں ہو سکتا ہے	وجود

Figure 1: Hierarchy of Urdu Word کام

For example hierarchy of moon in PWN 2.1 is as follows:

- Moon-- Star-- Celestial body-- Natural object-- Object-- Physical entity—Entity

Similarly, the hierarchy of چاند was developed by following the above hierarchy as:

Figure 2: Hierarchy of Urdu Word چاند

4. Challenges and Solutions

In this section the Interrupted Hypernym-labeled Noun Hierarchies are being shared with explanation and detailed analysis on how and why were they being interrupted.

• 4.1. Development and Mapping of Urdu WordNet Hierarchies with PWN 2.1

The biggest challenge was to align and map Urdu WordNet hierarchies with PWN 2.1. In order to complete the hierarchies many new Urdu senses were developed manually. The example of newly added senses to complete the hierarchy of Urdu word تجربہ has been shown in the following **table 1** below:

English ID	English Word	Category	Urdu Word
00633318	**experiment**	noun.act	**تجربہ**
00635582	**research project**	noun.act	**تحقیقاتی منصوبہ**
00630686	**research**	noun.act	**تحقیق**
00627860	**investigation**	noun.act	**دریافت، استفسار، تفتیش، انویسٹی گیشن، کھوج، تلاش**
00570312	**work**	noun.act	**کار، کام، عمل**
00403481	**activity**	noun.Tops	**فعالیت**
00029085	**human action**	noun.Tops	**انسانی سرگرمی**
00028105	**event**	noun.Tops	**واقعہ، ماجرا، ایونٹ**
00022007	**psychological feature**	noun.Tops	**نفسیاتی خاصیت**
00002236	**abstraction**	noun.Tops	**تجرید**
00002119	**abstract entity**	noun.Tops	**غیرمرئی شے**
00001740	**entity**	noun.Tops	**وجود**

Table 1: Example of Newly Added Senses of Urdu Word تجربہ

In order to complete this noun hierarchy, two new senses تحقیقاتی منصوبہ and تحقیق were added. The missing senses were translated from PWN 2.1 with their concepts.

• 4.2. Translation Issues

Many of the Urdu Noun hierarchies couldn't complete because of the translation issues. All the synsets of the hypernym are not present in English language because of the difference and unavailability of concepts. For example, اڈا is not found in English culture. Similarly, اردو بازار، صدر بازار are places which are loosely translated in English as market but that's a very weak translation. Although the word "Bazar" is found in English dictionaries but it gives a different sense. Moreover, its concept doesn't match with the

Urdu Bazaar and Sadar Bazar because these are limited to Urdu culture.

Moreover, it was difficult to the complete Noun hypernymy hierarchies of آئل "Oil" (ID: 7568129). Only shallow levels of its hierarchy were developed. It was problematic to find out the exact translation of the words: lipid – macromolecule – organic compound. These are scientific terms and are not found in OUD.

For translation of "lipid" words that were looked up were: حیاتی کیمیا؛ لحمیات؛ روغنیات؛ شحمیات؛ چربی but none of them were mapped with the PWN 2.1 sense for lipid. Likewise, for "macromolecule" no Urdu translation exists. There is translation for "molecule" but macromolecule is also a scientific term that needs to be added to Urdu language. As for "organic compound" It's formed by the combination of two words, no such phrase exists in OUD.

Another example is English Word "terminology" (ID 6220865) which was translated as اصطلاح. It's evident that this difference between English and Urdu causes the term language unit not remains in the state of noun when it is translated. Hypernymy hierarchy of two words: "newspaper advertisement" (ID 7150204) and "advertisement" (ID 7149579) get interrupted due to word "promotion". Both "newspaper advertisement" and "advertisement" are translated as اشتہار so there is one noun for both of these nouns. Another issue is the word "promotion" which has no concept in Urdu.

For the word "discipline", انضباط, تادیب, مضمون, قاعدہ, لِٹریچَر were looked up in OUP but none of them gave English sense.

It was observed that noun hierarchy as lexical inheritance system is limited in depth and it seldom goes more than ten levels deep. The deepest examples usually contain technical levels that are not part of everyday vocabulary which have been discussed above with reference to translation issues. Moreover, shallowest levels are also considered as too vague

• 4.3. Mismatch of Part of Speech Category

Mismatch of part of speech category was another challenge. English word "abstract entity" is Noun in PWN 2.1 where as its translated word غیرمرئی is an Adjective. However, the same translated word was used in the hierarchies because it was used with the word شَے so a compound word غیرمرئی شئے (Adj+N) was used.

• 4.4. Addition of translated Words

For the alignment and completion of Urdu hierarchies translated words were adopted and used in order to complete the hierarchies. However, these translated words are not available in OUD. For example, it was difficult to find out the proper word for "humanistic discipline" because humanistic is an adjective in Urdu and no coined word as "humanistic discipline" exists. Therefore, انسان شناس ادب as a translated word is used. However, such words are made considering compound translations of the English words which further lead towards the mismatch of part of speech category.

Also, definition for "Indo-European" and "Indo-Iranian" were not found in Urdu Lughaat so they were self created by taking definition from Wikipedia.

• 4.6. Issues of Compound Words

Another important issue was of compound words. English word "written communication" in PWN 2.1 is a compound word. It couldn't be found in Urdu language in compound form. So it needs to be added to have a complete set of hierarchy formed. Another example of interruption caused by "written communication" is that of ادب، لٹریچر "literature" (ID 6279556). Yet another hierarchy being interrupted by "writing communication" is اساطیر، This hierarchy for "myth" (ID 6287133) couldn't go to deep level due to absence of coined word "writing communication". Another reason was the multiple senses for the term "myth" as اساطیر،افسانے، کہانیاں،دیوکتھا were giving the same sense for the above mentioned single word.

Moreover, the "auditory communication" wasn't in OUD and its translation as سمعی مواصلات was used in order to complete the noun hierarchy and its semantic relations. Not only that but also the word "nonstandard speech" is not available in Urdu and it would need to have the word ہوا added which would turn it into non-Noun form. Just like "auditory

communication" and "written communication" nouns, another form "visual communication" causes interruption in the construction of hypernymy hierarchy of "name" with sense ID 6248892.

• 4.5. Confusing Hyponyms with Hypernyms

Another major challenge was to discover the actual relation of hypernym-hyponym. A noun "X" is a hyponym of a noun "Y" if "X" is a subtype or instance of "Y". Thus "Ashfaq Ahmed" is a hyponym of author and conversely "author" is a hypernym of "Ashfaq Ahmed" and so on.

5. Conclusion

There are issues and challenges constructing noun hypernymy hierarchy for Urdu WordNet by aligning and mapping it with PWN 2.1. Lexical relations, specifically hyponymy-hypernymy, are important in the development of information retrieval systems. There is rapid alteration in the theme of lexical semantics, computational lexicography, and computational semantics, but due to the availability of online lexical resources the construction of noun hypernymy hierarchies of Urdu WordNet is feasible. However, for Urdu language there is still need to develop more hierarchies in order to make Urdu WordNet's coverage better. Although, manually developed Urdu hypernymy hierarchy of nouns are highly accurate It is concluded that with handcrafted hierarchies there's a need for an automatic construction of hypernymy hierarchies of Urdu WordNet. Further research needs to be conducted in partitioning of noun hierarchy into separate hierarchies with unique top hypernyms. Moreover, it can further lead towards Parts and Meronymy: part whole relations.

6. References

[1] G. A. Miller, *Five Papers on WordNet,* 1993.

[2] C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[3] U. N. Singh. *Proceedings of the First Global WordNet Conference.* Central Institute for Indian Languages, Mysore, India, 2002.

[4] P. Sojka, K. Pala, P. Smrz, C. Fellbaum, and P. Vossen. *Proceedings of the Second International WordNet Conference*. Masaryk University, Brno, Czech Republic, 2004.

[5] P. Sojka, K. Choi, C. Fellbaum, and P. Vossen. *Proceedings of the Third Global WordNet Conference.* Masaryk University, Brno, Czech Republic, 2006.

[6] P. Vossen. EuroWordNet. Dordrecht, Holland: Kluwer, 1998.

[7] A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using
the Merge Approach *", in the Proceedings of Conference on
Language and Technology 2012 (CLT12)*, Lahore, Pakistan.

[8] A. Zafar, A. Mahmood, S. Shams, and S. Hussain, "Structural Analysis of Linking Urdu WordNet to PWN 2.1", *in the Proceedings of Conference on Language and Technology 2014 (CLT14)*, Karachi, Pakistan.

[9] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. Introduction to WordNet: An On-line Lexical Database. Princeton University, New Jersey, 1993.

[10] R. Jones, B. Rey, O. Madani, and W. Greiner. "Generating query substitutions." In WWW '06: *Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA. ACM, 2006, pp.387–396.

[11] V. Lopez, V. Uren, E. Motta, and M. Pasin. Aqualog: An ontology-driven question answering system for organizational semantic intranets. Web Semant., 2007, pp. 72–105.

[12] C. Dang, X. Luo, X., and H. Zhang. WordNet-based summarization of unstructured document. W. Trans. on Comp., 2008, pp. 1467–1472.

[13] J. Li, Y. Zhao, and B. Liu. "Fully automatic text categorization by exploiting wordnet." In *AIRS '09: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, Berlin, Heidelberg. Springer-Verlag, 2009, pp. 1-12.

[14] S. Afonin. "On Automated Hypernym Hierarchy Construction Using and Internet Search Engine." Moscow, Russian Foundation, 2010.

[15] Z. S. Harris. Mathematical Structures of Language. Wiley, New York, 1968.

[16] M. A. Hearst. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th International Conference on Computational Linguistics,* 1992, pp. 539–545.

[17] T. G. Bever and P. S. Rosenbaum. "Some Lexical Structures and Their Empirical Validity" in Jacobs, R. A., & Rosenbaum, P. S. (eds.). Readings in English Transformational Grammar. Waltham, Mass.: Ginn, 1970.

[18] M. R. Quillian. “Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities.” Behavioral Science, 1967, pp. 410-430.

[19] M. R. Quillian. “Semantic Memory.” In *Minsky, M. (ed.). Semantic Information Processing.* Cambridge, Mass.: MIT Press, 1968.

[20] A. M. Collins and M. R. Quillian, M. R. ”Retrieval Time From Semantic Memory.” *Journal of Verbal Behavior and Verbal Learning,* 1969, pp. 240-247.

[21] A. J. Wilkins. “Conjoint Frequency, Category Size, and Categorization Time.” *Journal of Verbal Learning and Verbal Behavior,* 1971, pp. 382-385.

[22] E. E. Smith. “Theories of Semantic Memory.” In Estes, W. K. (ed.). Handbook of Learning and Cognitive Processes, vol. 5. Hillsdale, NJ: Erlbaum. The Synonym Finder. 1978. Emmaus, Pa.: Rodale Press, 1978.

[23] C. Conrad. “Cognitive Economy in Semantic Memory.” *Journal of Experimental Psychology*, 1972, pp. 75-84.

[24] G. A. Miller, and W. Charles. “Contextual Correlates of Semantic Similarity.” Language and Cognitive Processes, 1991.

[25] A. Zafar, "Development of Urdu WordNet Noun Hierarchies and their Hypernymy Relations", Presented at 3-Day International Workshop on Corpus Linguistics, 2015.

[26] Center for Language Engineering, *Urdu WordNet 1.0 Wordlist,* 2013. Available at: http://www.cle.org.pk/Downloads/ling_resources/wordlists/Urdu%20WordNet%201.0%20Wordlist.pdf

[27] P. A. Chirita, S. C. Firan, and W. Nejdl. « Personalized query expansion for the web.” In *SIGIR '07:Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA. ACM, 2007, pp. 7-14