CLE Urdu Digest Corpus

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

Abstract

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

1. Introduction

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

2. Literature Review

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recommendations concerning representativeness, with general sampling frames including writing (published), writing (unpublished), speech and scripted speech.

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

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

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

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

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

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

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

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

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

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

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

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

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

3. The Process of Corpus Construction

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

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

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

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

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

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

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

3.1. Corpus acquisition

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

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

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

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

Table 1: List of Glossed Words

Original	Modified
’بیماری،	’بیماری‘
’اصول شفا،	’اصول شفا‘
’سنئروتھراپی،	’سنئروتھراپی‘
"آج میرے سر میں درد ہے۔،،	"آج میرے سر میں درد ہے۔"

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

3.2. Corpus organization

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

3.3. Corpus cleaning

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

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

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

Table 2: Genres of CLE Urdu Digest Corpus

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

Table 3: Errors of letter insertion

Original	Modified
االله	الله
التتمش	التمش
مى	میں

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

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

Table 4: Examples of Compound Words

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

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

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

Table 5: Examples Loan Words

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

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

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

Compounded Word
قواعدِ انشا
سربالیں
ردِ عمل
اجرِ آخرت

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

Table 7 (a): Single Word Abbreviations

Original	Modified
ایس ڈی او	ایس‌ڈی او
اے ایم ڈی	اے‌ایم ڈی
ایم این اے	ایم این اے
آئی ایم ایف	آئی ایم ایف

Table 7 (b): Abbreviations with Multiple Words

Original	Modified
اے کے سومار	اے کے سومار
این این مارک	این این مارک
ڈی واے لی	ڈی واے لی
این بی سی فوکس	این بی سی فوکس

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

Table 8: Words with Affixes

Original	Modified
دست مال	دست‌مال
ہموزن	ہم‌وزن
مندرجہبالا	مندرجہ بالا

4. Results

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

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

5. Discussion and Future Work

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

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

Table 9: Domain-wise Corpus Distribution

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

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

6. Conclusion

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

7. Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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