# Developing Tools for Arabic Corpus for Researchers

**Bassam Hammo[1], Faisal Al-Shargi[2]**
**Sane Yagi[3], Nadim Obeid[4]**
The University of Jordan

{b.hammo[1],obein[4]}@ju.edu.jo
f.alsharji@hotmail.com[2]
saneyagi@yahoo.com[3]

## Abstract

This paper presents an ongoing research that aims to construct a sizable and reliable text corpus along with a set of tools to experiment with natural language applications for Arabic. The corpus is used by graduate students at the University of Jordan (UJ) to conduct experiments on many useful applications. Earlier, we were not able to verify these experiments because of the lack of reliable data. We are working on annotating and tagging the corpus texts and making it available for researchers in XML format.

## Introduction

Most researchers working in the field of Arabic natural language processing (ANLP) opt to construct their own manually collected datasets to run their experiments. Most of the times, the datasets are small and therefore their experimental findings may neither be convincing nor clear as how to scale up the results. Linguistic resources, which are required to advance research of Arabic language processing, have to be built from scratch and then they should be shared with researchers in the field of ANLP to expedite the development of Arabic natural language processing applications.

Existing Arabic corpora suffer from many issues (Maamouri et al., 2004; Al-Sulaiti & Atwell, 2005; Duh & Kirchhoff, 2005; Abdelali et al., 2005). In addition, the commercial ones are very expensive.

In this work we give a brief description of an Arabic text corpus, together with a set of useful tools to experiment with the dataset. We aimed to help our graduate students at UJ to experiment with ANLP applications in areas such as: Information Retrieval, Question Answering, Text Categorization, Text Summarization, Machine Translation and many other applications. We plan to make this system available for academic researchers through http://www.nlp.ju.edu.jo. In the following sections, we describe the different sources and genres of the text corpus, the set of tools which we developed to experiment with this corpus and the plans for the future work.

## Sources and Genres Included in the Corpus

The constructed corpus contains a diverse range of sources. At first, we compiled the editorial articles of the best written daily newspapers in the Arab world (15 newspapers). In addition, many topic areas were collected from different fields such as: Arabic literature, constitutions of Arab countries (19 countries), technology news, political news and sports news. The diversity of the topics are intended to enrich the content of the modern standard Arabic (MSA), which is the language used and spoken by many Arabs in the Arab world.

For classical Arabic, we have compiled and included the Quranic corpus for its richness of Arabic words and to experiment with the current usage of these words in the modern standard Arabic. Finally, we have used two Arabic dictionaries: a modern one (Al-Mojam Alwaseet) and an old dictionary (Lisan Al-Arab). The size of the collected data is close to 41 MB of text data assembled in 61 thousand files. The corpus has a total of 7.5 million words of which 707,483 words were distinct across all genres (cf. Table1).

| Genres | Size MB | # of Files | # of Words | # Distinct Words |
|---|---|---|---|---|
| Politics | 3.56 | 1697 | 604882 | 65187 |
| Editorials | 2.81 | 737 | 488111 | 66571 |
| Sports | 5.82 | 2977 | 994032 | 70514 |
| Quran | 0.701 | 114 | 78370 | 14648 |
| Literature | 3.02 | 731 | 552679 | 90296 |
| Constitutions | 0.678 | 19 | 106791 | 14139 |
| Dictionaries | 24.29 | 54638 | 4587594 | 365857 |
| Technology | 0.646 | 124 | 110482 | 20271 |
| Total Size | 41.525 | 61037 | 7522941 | 707385 |

Table 1: UJAC sources and genres

## Corpus Design

Starting with a raw corpus collected from different sources and genres, we attempted to

build a clean structured corpus. A set of tools were developed to convert the HTML files into UTF-8 plain text files. Other tools were needed to parse the text and split it into sentences. In addition, we used a stemmer to extract the words' roots (Figure 1). An ongoing work is planned to annotate and POS tag the text corpus, and to convert it into an XML structure for various kind of studying and use on Arabic language (Figure 2).

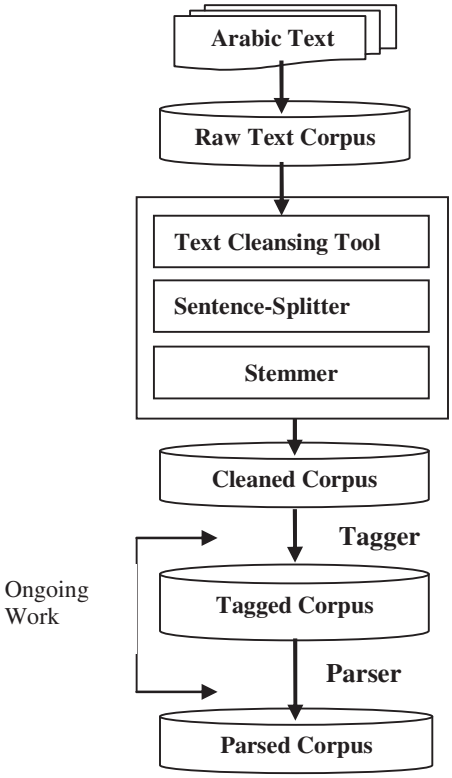


Figure 1. Schema of building the corpus (phase I)

```
<doc>
  <text>
     <body>
     <s>
     <word id="" pos="tag">WORD</word>
     </s>
     </body>
  </text>
</doc>
```

Figure 2. Annotated and POS tagged XML Structure of the text corpus (phase II)

To see how balanced is our collection, we conducted some experiments with Zipf's law, which is used in corpus analysis to draw a relationship between the frequency of a word (f) and its rank (r) in the list (Manning & Schutze, 1999). The law is expressed as follows:

$$r.f = c$$

Where r is the rank of a word, f is the frequency of occurrence of the word, and c is a constant that depends on the analyzed text. Table 2 and Figure 3 show that our corpus is highly balanced.

| Genre | R* | Genre | R* |
|---|---|---|---|
| Politics | 0.982 | Literature | 0.990 |
| Editorials | 0.981 | Constitutions | 0.989 |
| Sports | 0.979 | Technology | 0.990 |
| Quran | 0.990 | | |

Table 2: Zipf's results for UJAC, R*= Linear Regression

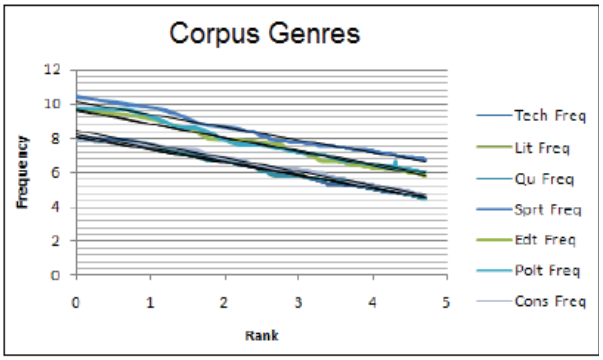


Figure 3: The balanced UJAC

## The Corpus Interface and Tools

Figure 4, depicts the corpus GUI and the set of tools which we have conducted to experiment with the corpus. The following subsections describe the tools briefly as we have a limited space.


Figure 4: The UJAC System

Lexicographers always need to investigate the meaning and the usage of words. In addition, they need to study words synonyms. In this

work, we developed a set of tools to answer questions such as:

- How common are different words?
- How common are the different senses for a given word?
- How words are associated with other words?

The UJAC system provides the following set of tools and modules:

### 1. Searching through different word forms

One of the advantages of the Word-Form tool is to show all the contexts in which a word occurs. By using a corpus, it is possible to identify the different meanings associated with a word and to resolve its ambiguity. Figure 5 depicts an example of using this tool.

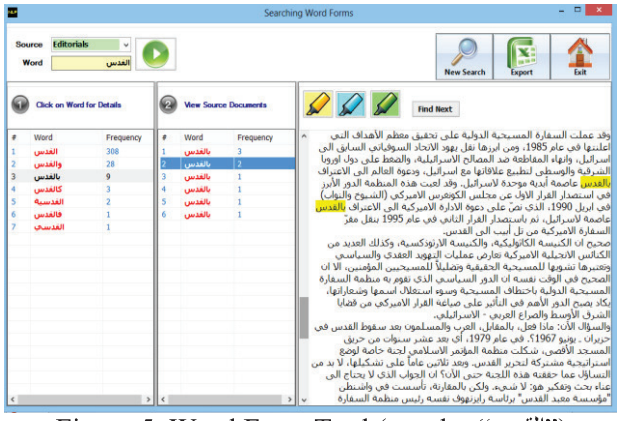


Figure 5. Word Form Tool (word = "القدس")

### 2. Searching based on a word length.

The Word-Length tool (Figure 6) helps understanding the word structure and usage across different genres. By experiments, we found that words of length equal to 4, 5, 6 and 7 letters represent about 80% of the total distinct words in the collection. These words were amongst the most frequent words in the corpus with frequencies: 14.26%, 23.98%, 24.12% and 17.68% respectively.

### 3. Providing statistics on matched and mismatched words in a text.

This is a very useful tool to check the use of words in Modern Standard Arabic (MSA) articles compared with classical source of Arabic language represented through the Quranic corpus and Arabic dictionaries such as Al-Mujam Alwaseet (modern) and Lisan Al-Arab (old). The tool can check a single article or multiple articles. Table 3 shows an experiment of checking a sport article of 184 words using this tool.

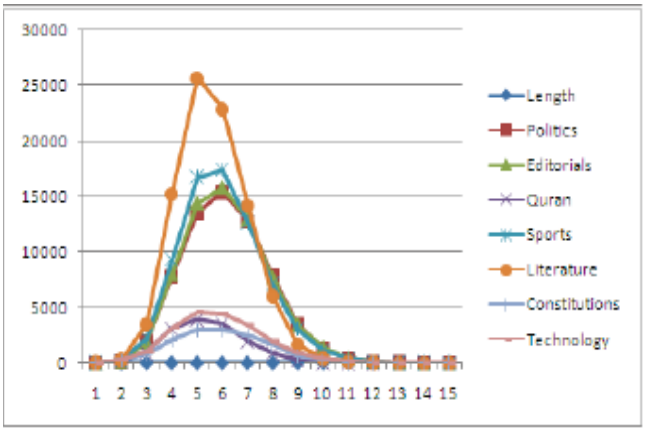


Figure 6: Distribution of words based on length

| Source | # of Words | # of Match | # of Missing | % of Match |
|---|---|---|---|---|
| Quran | 14648 | 64 | 120 | 34.8% |
| Lessan Al-Arab | 299919 | 119 | 65 | 64.7% |
| Alwaseet | 131052 | 128 | 56 | 69.6% |

Table 3: Word-Check tool

### 4. Searching for words frequencies within the corpus

The Word Frequency tool is a useful tool to answer some questions like: which words are the most common in a language and which words are uncommon?

Experiments with the Word-Frequency tool show that long words (> 10 letters) are most likely Arabic transliteration words. (cf. Table 4).

| N# | Word | Freq | Length |
|---|---|---|---|
| 1 | والبتروكيماويات | 4 | 15 |
| 2 | الهيدروجيولوجية | 3 | 15 |
| 3 | الانتركونتنتال | 2 | 15 |
| 4 | والبروتستانيين | 1 | 15 |
| 5 | الكهروميكانيكية | 1 | 15 |
| 6 | والانتروبولوجيا | 1 | 15 |
| 7 | الانجلوسلاطيني | 3 | 14 |
| 8 | والديموقراطيون | 2 | 14 |
| 9 | الجيوستراتيجية | 1 | 14 |
| 10 | الارسطوطاليسية | 1 | 14 |

Table 4. Arabic transliteration words

### 5. Searching for a word usage before/after *n* words.

This is another tool to help lexicographers to understand how a word can be employed in a sentence by examining the *nth* words before and after the tested word. This is also useful to

figure out the different meanings of a word based on its order within the context (cf. Figure 7).


Figure 7: Before & after a word

## 6. Searching for words/roots within the corpus.

This tool helps to search the corpus using a word form or through the root of a word. It helps detecting all variations of a word originated from the same root.

## 7. Tagging text passages based on the Quran tagged corpus

Tagging text based on the tagged Quran is another useful tool to experiment with text POS tagging. We plan an ongoing work to annotate and tag the corpus and then to use it in part of speech tagging.

## Conclusion

Arabic NLP is still not yet widely studied by computer scientists and computational linguists. This is mainly due to the problem of obtaining large amounts of text data (Duh & Kirchhoff, 2005). Its traditional rules such as inflectional, derivational, part of speech, sentence constituents and many other features are difficult to computerize. In this paper we focused on the design of a sizable, reliable balanced Arabic corpus collected from different sources to experiment with Arabic NLP applications. We described some of the useful tools which we have developed throughout this project. We have set a plan to annotate and POS tag the corpus using a reasonable tag set. We intended to make the corpus and the tools open to public research.

## References

Al-Sulaiti L., Atwell E. 2005. "Extending the Corpus of Contemporary Arabic". In Proceedings of Corpus Linguistics conference 2005, University of Birmingham, UK, 1-9.

Abdelali A., Cowie J. and Soliman H. 2005. "Building a Modern Standard Arabic Corpus". Workshop on Computational Modeling of Lexical Acquisition 2005. The Split Meeting. Croatia, 1-7.

Duh K., Kirchhoff K. 2005. POS tagging of dialectal Arabic: A minimally supervised approach, Association for Computational Linguistics POS Tagging 55-64.

Maamouri M., Bies A. and Buckwalter T. 2004. "The penn Arabic treebank: Building a large-scale annotated arabic corpus". In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.

Manning Christopher and Schutze Hinrich. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.