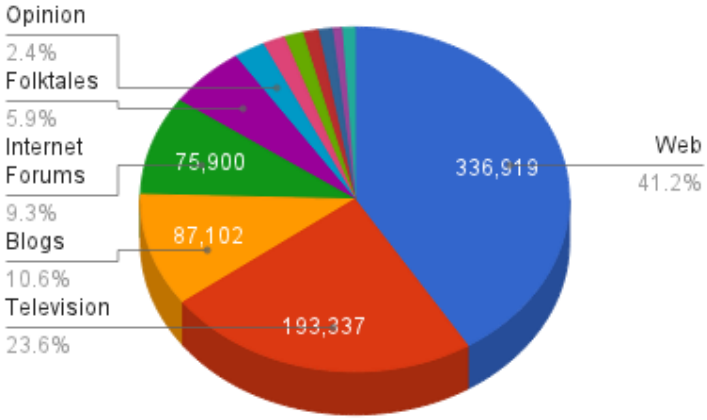
Tunisian Arabic Corpus

Karen McNeil • karen_mcneil@brown.edu • tunisiya.org

The Tunisian Arabic Corpus (TAC) is a project, led by Karen McNeil and Miled Faiza, seeking to build a four-million-word, publicly-available corpus of Tunisian Spoken Arabic. The corpus is available freely online at tunisiya.org, where users can perform complex concordance searches and view search results in context, with access to the full text. The corpus is stored in a database, which is accessed through a web application written in Python, with the web framework Django.

There are many challenges to creating Arabic corpora, and dialectal corpora in particular, including those of sources, spelling variations, balance, and parsing. The corpus currently has only about 820,000 words, and issues of balance and parsing have not been completely solved. Nonetheless, TAC has proved to be a useful resource to Arabic students and researchers all over the world, and also presents a model for others who wish to create dialectal Arabic corpora.

Corpus Composition



Total: 818,310 words

Sources

- Traditional Written Sources
  - Folklore
  - Songs / Folk Poetry
  - Proverb collections
  - Screenplays
- New Written Sources
  - Blogs
  - Email
  - Facebook
- Transcribed Audio
  - Radio

Figure 1. *Screenshot of one of the texts in the corpus, with associated metadata.*

Corpus Search and Concordancing

Users of the corpus can search for individual words three different ways:

- **Exact**: This will return words exactly as written. In the example below, an exact search for the verb لوج ('to search') will return *only* لوج
- **Stem**: The user types in the stem of the word (i.e. لوج or بيت) and will get back all inflected forms, such as البيت , يلوجوا , etc. This is accomplished using a custom-built parser which has an 88% accuracy rate. (See screenshot of results below)
- **Regex**: This allows searches using regular expression. Searches must be typed in the modified Buckwalter transcription. A search for `[nty]lwj` with return all present-tense singular forms of لوج

Searches can also be limited by category.

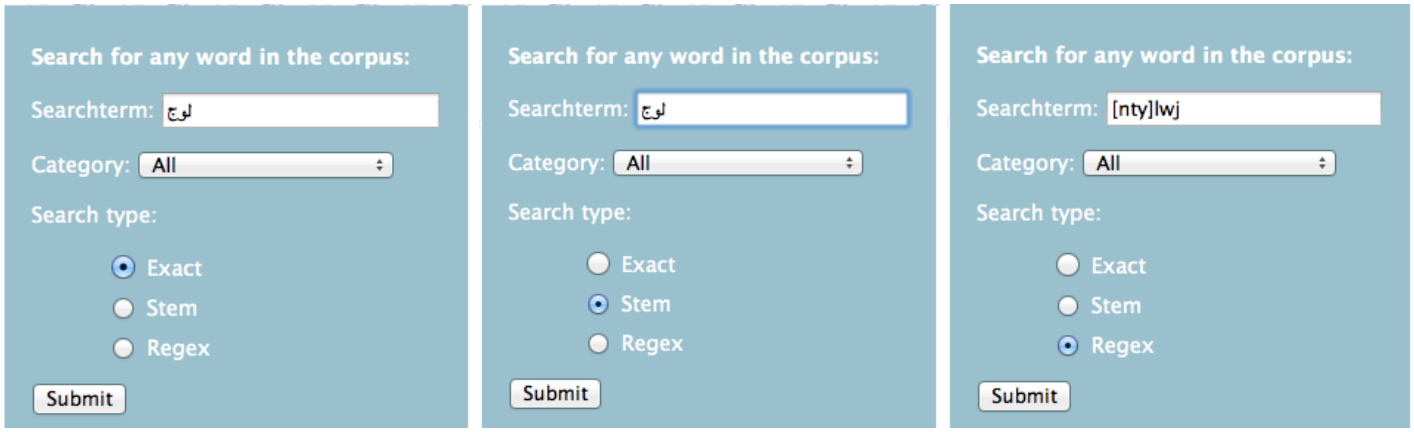


Figure 2. *Three different searches for* لوج

Figure 3. *Results from a "stem" search for* لوج *which returns all inflected forms.*