

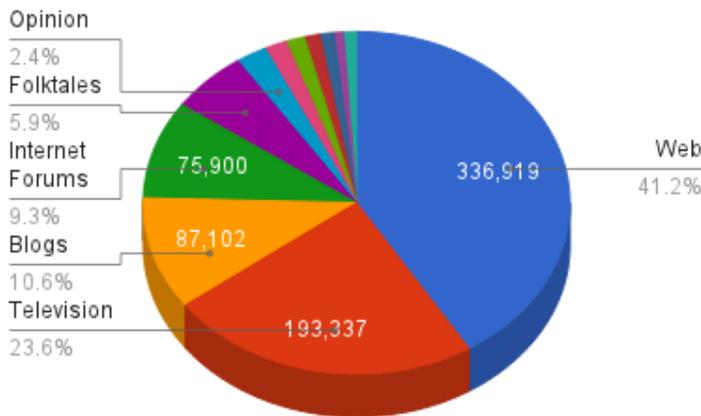
# Tunisian Arabic Corpus

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

**Tunisian Arabic Corpus**

Home

### ! A la Une: Hram... Tout est Hram

– المذيع: الشروق عملت دورة باهية ع البلدان العربية اللي شعوبها ثارت وحلفت ما تريض كان ما تلقع الأنظمة الفاسدة اللي تحكم فيها بعروقتها، في ليبيا الوضع تعرفوه، الغرب يضرب والغدافي يتوعد، يقولهم "عدوانكم إلى مزبلة التاريخ"، الغدافي دقدقو لو سلاح الطيران متاعو الكل و شطر الجيش، وهو مازال يبق في الساعات و يدور من زنفة لزنفة و يقول راني باش نتتمصر ونفقد الثورة العالمية، واللي يحاربوا فيا حكام ظالمين و شعوبهم ثابرة عليهم، في اليمن الثوار يحضرون لجمعة الزحف، أكثرية الجيش و برشه نواب برلمانيين انضموا للثورة الشعبية وصوتو ضد حالة الطوارئ، علي صالح قالهم نتنحي في آخر العام، قالو لو لا، تو تو، و كالعادة للعبة باش تتلعب نهار الجمعة، في سورية موتى و مجاريح على بعضو في درعا، البوليسية فرقوا اعتصام في جامع بالعنف والكزوتوش وقالوا عصابات مسلحة و زادوا بعثوا القناصة، ما تحكيوش عالعبرين عاده خاطر قوات درع الجزيرة الي جاية باش تقمع الشعب ما هيش تدخل أجنبي ولا احتلال خارجي، أيا خويا ازبولنا رواحكم، كملوا حرروا بلدانكم و ابعثوا رؤسائكم للسعودية و وخلي تصدموا لكل لغادي، نكملوا نحرروا البلاد هانكيا، و فرد مرة الي عندو شكون حاجتو بيه غادي يكمل يجيبوا و نعدوا إلى البيت فرحين مسرورين نعملوا أمة عربية حرة موحدة بالحق (LAUGHING) شنيته شبيك تضحك إنت؟ ماش عجبك الحكاية؟

PAUSE

الصريح اليوم عندها خبر باهي، مصادرة أموال و ممتلكات بن ضياء و عبد الوهاب عبد الله والقلال والودرني اللي خلطوا على الرئيس السابق و مرتو و فاميلتو و نسابو، نلاحظوا لهنا اللي قالمقالات sérieux ما يقولوش المخلوخ وهذا بفضل إحدات لجنة المصادرة المكلفة بأمالك الدولة والشؤون العقارية في وزارة أملاك الدولة للشؤون العقارية، أيها هيا اللجنات ولا لؤح، الزعيم وسيم الحريسي الأمين العام لحزب الحي الوطني التقدمي ألقى خطاب بالمناسبة هادي نفلو لكم منو بعض المعتطفات في الإرسال التالي:

**Text Metadata**

- Title: A la Une: Hram... Tout est Hram
- Date: Mar 24, 2011
- Category: Radio
- Transcriber: Adnan el Bakkali (A3CYGGER5HVCV6)
- Channel/Station: Mosaïque FM
- File: corpus/Hram...Tout\_est\_hram\_1

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 will return all present-tense singular forms of لوج

Searches can also be limited by category.

The figure shows three screenshots of the search interface for the Tunisian Arabic Corpus. Each screenshot has a light blue background and a white search box. The first screenshot shows an exact search for 'لوج' with the 'Exact' radio button selected. The second screenshot shows a stem search for 'لوج' with the 'Stem' radio button selected. The third screenshot shows a regex search for '[nty]lwj' with the 'Regex' radio button selected. Each screenshot includes a 'Submit' button and a 'Category' dropdown menu set to 'All'.

Figure 2. Three different searches for لوج

The screenshot shows the Tunisian Arabic Corpus search results page. The header features a blue and white decorative border with a central image of a blue door. The title 'التونسية' is written in white Arabic script, and 'Tunisian Arabic Corpus' is written in white text below it. The navigation menu includes 'Home', 'Our Team', 'Using the Corpus', and 'Admin'. The search results section is titled 'Search Results for لوج' and includes a 'Download Results' link. The text below the search results indicates that the search for 'لوج' returned 117 results in 70 texts. The results are presented in two columns of Arabic text, showing various inflected forms of the word 'لوج' used in different contexts.

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