Amazigh ConCorde: an appropriate concordance for Amazigh

Siham Boulaknadel

Institut Royal de la Culture Amazighe
Avenue Allal El Fassi, Madinat Al Irfane, Rabat

boulaknadel@ircam.ac.ma

Résumé – Abstract

L'ampleur remarquable qu'a connue la constitution des corpus en langue amazighe, ces dernières années, a nécessité l'élaboration d'outils robustes permettant le traitement des données, que ce soit pour la recherche ou pour l'enseignement.
Un de ces outils est le concordancier, qui offre plusieurs fonctionnalités tels que l'affichage des mots définis dans leurs contextes et la fréquence des mots. L'obtention d'un outil pareil, répondant aux caractéristiques de la langue amazighe, s'est avérée d'une difficulté extrême. Par conséquent, *Amazigh Concorde* a été créé pour que la communauté dispose d'un tel outil.

Amazigh linguistics corpora are currently enjoying a surge activity. As the growth in the number of available Amazigh corpora continues, there is an increased need for robust tools that can process this data, whether it is for research or teaching. One such tool that is useful for both groups is the concordancer — a simple tool for displaying a specified target word in its context. However, obtaining one that can reliably cope with the Amazigh language has proved an extreme difficulty. Therefore, Amazigh ConCorde was created to provide such a tool to the community.

Keywords – Mots Clés

Langue Amazighe, concordance, corpus.
Amazigh language, concordance, corpus.

1 Introduction

A concordancer is a simple tool for analysing the corpora contents, based on words of the user interest. Otherwise, manually navigating through (often very large) corpora would be a tedious task. Concordancers are not only extremely useful as a time-saving tool, but also much more. Lexicographers and linguists can find insights into grammatical structure by studying concordance output. By isolating keywords in their contexts, linguists can use concordance output to understand the behaviour of interesting words. Lexicographers can use the evidence to find if a word has multiple senses, and also towards defining their meaning. There is also much research and discussion about how concordance tools can be beneficial for data-driven language learning (Johns, 1990). A number of studies have demonstrated that providing access to corpora through concordancers is beneficial to learn foreign languages (Dodd, 1997). Cobb et al. (2001) detailed the improved rates of vocabulary acquisition when software tools based on concordancers are used. The concordancer's simplicity and usefulness make it a valuable tool.

The general trend within corpora is that research priorities have only been focused on European languages. Which make the less commonly taught languages, as the Amazigh one, know a deep lack of resources. Effectively, when this project began, to the best of our knowledge, there were no available concordancers that support the Amazigh language. Considering Amazigh is spoken by approximately 50 % of Moroccan people, it was felt that Amazigh language deserves such useful tool. Thus, we have developed a concordancer that we have called "Amazigh ConCorde".

2 Amazigh characteristics

The Amazigh language, known as Berber or Tamazight, covers North Africa, Sahara and parts of the West African Sahel. In Morocco, it is divided into three big regional varieties with Tarifite in North, Tamazight in Central Morocco and south-east, and Tachelhite in the South-west and the High Atlas. This language was exclusively reserved for familial and informal domains, even 50% of the Moroccan population are Amazigh speakers (Boukous, 1995). But in last decade, the Amazigh language has become an institutional language in Morocco.

2.1 Amazigh writing

Since the ancient time, the Amazigh language has its own writing, Tifinaghe, which is still used up to day in Sahara areas (Tuaregs).

Nevertheless, throughout history, this writing has assimilated some changes in the aim to provide to Amazigh language an adequate and usable standard alphabetic system. Thus in 2003, the Royal Institute for Amazigh Culture (IRCAM) has developed a new alphabet system called Tifinaghe-IRCAM.

This alphabet is based on a graphic system towards phonological tendency. This system does not retain all the phonetic realizations produced, but only those that are functional (For more details see (Ameur et al., 2004)).

The Tifinaghe-IRCAM alphabet contains:

- 27 consonants including: the labials (ⴼ, ⴱ, ⵎ), dentals (ⵜ, ⴷ, ⴹ, ⵟ, ⵏ, ⵔ, ⵕ, ⵍ), the alveolars (ⵙ, ⵣ, ⵚ, ⵥ), palatals (ⵛ, ⵊ), the velar (ⴽ, ⴳ), the labiovelars (ⴽⵯ, ⴳⵯ), the uvulars (ⵇ, ⵅ, ⵖ), the pharyngeals (ⵃ, ⵄ) and the laryngeal (ⵀ);
- 2 semi-consonants: ⵢ and ⵡ;
- 4 vowels: three full vowels ⴰ, ⵉ, ⵓ and neutral vowel (or schwa) ⴻ which has a rather special status in Amazigh phonology.

2.2 Punctuation

No particular punctuation is known for Tifinaghe. IRCAM has recommend the use of Latin symbols: « » (space), «. », «, », «; », «: », «? », «! », « … ».

2.3 Numeral

IRCAM has used the Arabic numeral (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) for Tifinaghe writing.

2.4 Directionality

Historically, in ancient inscriptions, Amazigh language was written horizontally from left to right, from right to left, vertically upwards, downwards or in boustrophedon. However, the orientation most often adopted in Amazigh language script is horizontal and from left to right. IRCAM has taken the horizontal direction from left to right in Tifinaghe writing.

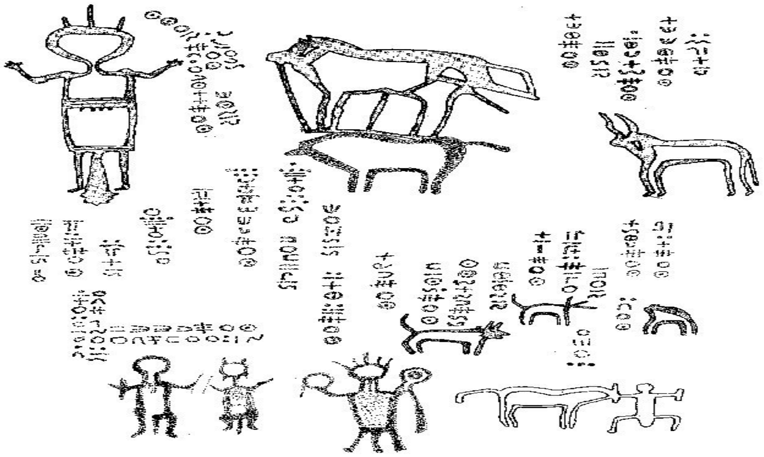

Figure 1: Plate 9, Anou Elias, Mammanet Valley (Niger)[1]

3 Amazigh ConCorde

Generally, all concordances are relatively basic. However, up to day no one of them can support the Amazigh language. In the aim to use a concordance in Amazigh lingual

[1] Henri Lhote, Les gravures de l'Oued Mammanet. Les Nouvelles Editions Africaines. 1979

applications, we have implemented the Amazigh ConCorde in a way that it supports the Amazigh language alphabet, and provides a convivial interface for the amazigh speakers. Moreover this ConCorde is relatively simple to get up and running. Its interface ensures a corpus selection, a query keyboarding or word selection from the word frequency panel, and the display of results.

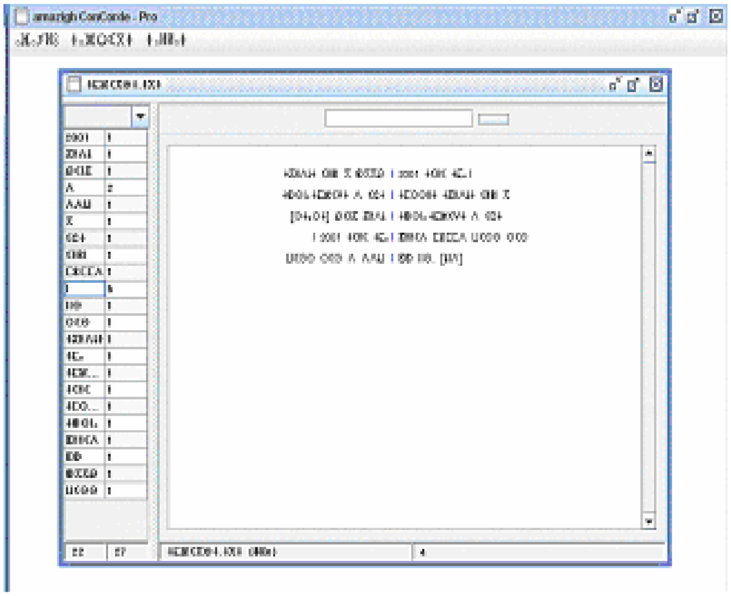

Figure 2: Amazigh ConCorde interface

4 The Amazigh ConCorde System

The Amazigh ConCorde tool was originally conceived as a system to support the Amazigh language. This system is built by using the Java programming language to allow the programmer to use it in most popular operating systems without any extra effort. Thus the Amazigh ConCorde can correctly run on Microsoft Windows, Linux or Mac OS (and others) only by installing the Java Runtime Environment.

4.1 Amazigh ConCorde features

The Amazigh ConCorde system contains a number of features that allow it to stand out from competing products. These features are:

- Full Amazigh support, which means that the system does not need to transliterate the Amazigh alphabet to Roman alphabet before applying the concordance.
- Multi-platform, which makes the system able to run on most major operating systems.
- Multi-format support that allows the user to load text files, HTML files.

- Results' storage, in way that the user can save the results either as a plain text file, or HTML file that can keep the concordance alignment.
- Word frequency analysis.
- A range of query options, which provide the user to choose between the range key-word, phrase, or wildcard.
- Freeware and open-source. The Amazigh ConCorde is released under the General Public License.

4.2 Amazigh ConCorde limitations

At the time of writing, Amazigh ConCorde is still limited in terms of some features, such as: Mark-up, Full context and the Amazigh language rules.

- Mark-up: Amazigh ConCorde is generally ignorant of mark-up annotation within a corpus. When a corpus is annotated with part-of-speech tags, Amazigh ConCorde is not recognising them, and it treats them as a 'simple word' belonging to the text itself. Once the set of part-of-speech tags will be completed by the IRCAM researchers, it will be integrated into the tool.
- Full context: Amazigh ConCorde doesn't have the functionality to allow the user to see the full context of a selected concordance item. The current display permits only to see the word within the sentence where it was found. Actually, it is the largest scope currently implemented.
- Amazigh language rules: Until few years ago, the Amazigh was known only as a spoken language; but on 2003, it was standardized. Since, researchers are focused on elaborating and developing all the Amazigh linguistic rules and resources that will be integrated progressively in Amazigh ConCorde.

5 Conclusion

This paper has summarised the features of the Amazigh concordance, and the issues regarding the difficulty and the limitations to build it. The Amazigh ConCorde was built in attempt to resolve some core issues of the researchers and foreign language learners. It displays the Amazigh text correctly. Moreover, its output is good, as expected by the Amazigh speakers. Furthermore, no transliteration was required neither for the input nor for the output of this system. Future directions are addressing the limitations confronted, especially the mark-up problem and the Amazigh language rules' lack.

References

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. M., Souifi H. (2004), Initiation à la langue amazighe, Rabat, IRCAM.

Boukous A. (1995), Société, langues et cultures au Maroc : Enjeux symboliques, Publications de la Faculté des Lettres de Rabat, Casablanca, Najah El Jadida.

Chaker S. (2003), Le berbère, Les langues de France, Paris, PUF, pp. 215-227.

Cobb, T., Greaves, C., Horst, M. (2001), Can the rate of lexical acquisition from reading be increased? an experiment in reading french with a suite of on-line resources, In Regards sur la didactique des langues secondes, P. Raymond and C. Cornaire, eds., Éditions Logique, Montréal.

Dodd, B. (1997), Exploiting a corpus of written german for advanced language learning, In Teaching and Language Corpora, Anne Wichmann, Steven Fligelstone, Gerry Knowles and Tony McEnery, eds., pp. 131–145, Longman, London.

Johns, T. (1990), From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning, Computer Assisted Language Learning, 10, pp. 14-34.