For standardised Amazigh linguistic resources

Youssef Aït Ouguengay & Aïcha Bouhjar
Institut Royal de la Culture Amazighe (IRCAM)
B.P. 2055 – Rabat (Morocco)
E-mail : ouguengay@ircam.ma, bouhjar@ircam.ma

Abstract

Amazigh language and culture may well be viewed to have known an unprecedented booming in Morocco : more than a hundred-which are published by the Royal Institute of Amazigh Culture (IRCAM), an institution created in 2001 to preserve, promote and endorse Amazigh culture in all its dimensions. Crucially, publications in the Amazigh language would not have seen light without the valiant attempts to upgrade the language on the linguistic and technological levels. The central thrust of this contribution is to provide a vista about the whole range of actions carried out by IRCAM. Of prime utility to this presentation is what was accomplished to supply Amazigh with the necessary tools and corpora without which the Amazigh language would emphatically fail to have a place in the world of NITCs. After a brief description of the prime specificities that characterise the standardisation of Amazigh in Morocco, a retrospective on the basic computer tools now available for the processing of Amazigh will be set out. It is concluded that the homogenisation of a considerable number of corpora should, by right, be viewed as a strategic move and an incontrovertible prerequisite to the computerisation of Amazigh,

1. Introduction

Conscious of the importance and challenges of NITC in the preservation, diffusion and promotion of Amazigh culture on all grounds, IRCAM, ever since the elaboration of its first action plan in 2003, embarked on a composite of various projects meant to achieve the linguistic and technological standardisation of Amazigh. This paper tries, thereby, to record the actions undertaken by IRCAM in this domain, along with a presentation of the achieved results and the problems that befell the accomplishment of such results. The paper is also meant to evaluate the technical and linguistic quality of the resources which were set up to meet Amazigh language processing ends.

2. The problematic issue of Amazigh linguistic resources

In Bouhjar (2008), the author provides a description of the nature and quality of the linguistic corpora made avail of to achieve the standardisation of Amazigh. Of most concern to the goal meant to be achieved, and which we describe in this section, is that variation, is accommodating in a “multidimensional” fashion. This “multidimensionality” is apparent in the outcomes of recording and exploiting linguistic corpora. Crucially, the process of recording and exploiting linguistic corpora ends up in highly heterogeneous corpora. The prominence of the corpora’s heterogeneity is not only apparent at the linguistic level proper (written form, spelling, morpho-syntax and lexicon) but also at the nature and content of the resources. Such diversity in form and content essentially necessitates homogenous corpora rewriting and not only transliteration, in line with the scriptural norms set by IRCAM. This rewriting stage, which is prior to the process of Amazigh language processing, could only be undertaken by Amazigh linguists who are well aware of the theoretical and methodological underpinnings underlying the generation of such linguistic resources (existing essentially in paper format). We can already see the enormity of the task to be carried out to provide the Amazigh language with digitalised linguistic resources which will be exploited for the development of efficient and competitive computer tools. Such tools will, no wonder, satisfy the Amazigh continuously growing needs given the progressive extension of the domains in which Amazigh is used in Morocco both public wise and most notably teaching and media wise.

With the above in mind, it is fair to say that Morocco has at its disposal a standard graphic system consisting of 33 Tifinagh alphabet letters made avail of in writing Amazigh. This graphic system has become official in Morocco since 10 February 2003 and has been approved by ISO under Unicode/ISO 10646 in June 2004. Amazigh has at its disposal stabilized spelling rules as well which have clearly been defined since 2003. Crucially, it is owing to these first steps of writing normalisation that the appearance of the first Amazigh textbooks and the integration of Amazigh in the educational system were possible in September 2003. Another move which deserves mention is the creation of an Amazigh grammar and a composite of specialised lexicons and vocabularies. “The polynomic approach that IRCAM adopts in the standardisation of Amazigh accounts for geolectal variation. Put in another way, the polynomic approach favours the neutralisation of variation when such a move is possible and preferable to increase communication (mainly at the phonic level); however, the approach foils the attempt of creating neutralisation when variation leads to the enrichment of the language (mainly at the lexical and morpho-syntactic levels). When a terminological gap holds, we resort to word creation to resolve the situation deriving, thereby, the lexical content from the whole range of national varieties. Neology may well be viewed as the final resort after all the language resources are exhausted.” (Bouhjar, to appear)

With the presentation above as background, we may well say that many efforts have been invested into developing tools to accommodate both the Tifinagh graphic system along with Amazigh morpho-syntactic annotation; the two of which we shall set out below with a special focus on the various stages of their development.

3. Tools developed for processing the Tifinagh writing system

3.1. Encoding, transcoding and classification

The written texts which were written prior to the integration of the Tifinagh writing system in the Unicode-based multilingual plan were recorded in ASCII encoding. A phonetic mapping was set between Latin, Arabic and Tifinagh (Ameur, Bouhjar & al., 2006), which has facilitated the passage from a graphic system to another, of most concern here between Tifinagh and Latin where directionality is the same (heading towards the right). Bijectivity presented no real problems. Since 2004, the recommended encoding for Tifinagh is the one set up by ISO 10646/UNICODE imposing, thereby, the transcoding of former products to the Unicode format with an eye to unifying the corpora produced and edited by IRCAM.

3.2. Tifinagh-IRCAM transcoder

Intended basically for researchers, this utility may well be viewed as a tool that transcodes Word documents (.doc) which are written in Tifinagh ASCII encoding-based fonts into a Tifinagh Unicode encoding. The accommodated texts are the ones typed by means of Tifinagh-IRCAM fonts (Web page : www.ircam.ma). In fact, the transcoder enables us to move from one encoding to another in both directions (from ASCII to Unicode and from Unicode to ASCII). This utility is of prime importance for the transcoding of web documents and on line publications.

Amazigh has also benefited from another normalisation, namely the norm ISO 14651 whose ultimate goal is to classify character strings. Significantly, a tool to undertake the sorting of character strings has become essentially necessitated owing to the mergence of a composite of publications such as new Amazigh lexicons/vocabularies, primary school textbooks and didactic supplementary material. These publications happen to be of prime utility for the insertion of Amazigh in the educational system.

3.3. Amazigh character strings sorting tool

Not unlike encoding, the utility of classifying Amazigh character strings starts by contending with documents written in ASCII and later with documents written under a Unicode format. Tifinagh was forged on the basis of Latin because both have the same directionality. It is because of this that we have initially resorted to a whole range of ASCII characters pre-processing and reordering operations. As for Unicode, a direct application of the norm ISO 14651 was entertained under a new version of the sorting tool (Outahajala, 2008)

3.4. Generalisation of Unicode use

Over the last few years, Tifinagh has gained a lot of interest from the editors of free operating systems in particular. Such interest is reminiscent of the increasing emergence of encoding systems (Unicode) coupled with the world's inclination towards internationalisation and globalisation. (118n). Tifinagh is, thereby, natively set up in many operating systems (Linux distributions for example) and is easily integrated in text editors (Openoffice, MS Office, etc.). On national grounds, Amazigh writing system edition tools are distributed for free on IRCAM's official internet site (www.ircam.ma); the latter is in charge of developing such tools and providing technical assistance for users.

4. Towards the generation of annotated corpus

Once the basic tools for Amazigh computer typing have been set up, a plan is underway to give a handle on the fundamental mechanisms of natural language analysis and processing. Sorely needed now is a morpho-syntactic classification of words which makes the purpose-built processing of Amazigh possible.

4.1. Amazigh corpus tagging tool

The use of a manual tagging tool can in no way be ignored at a first stage. Such a tool will enable linguists to morpho-syntactically tag a sufficiently big corpus making it, thereby, possible to carry out future automatic tagging. It goes without saying that on the basis of this first work that the correctness and reliability of the final tagging depends. This tool relies on a number of other contributing criteria such as the fineness of chosen tags' interplay and the technological performance and effectiveness (such as speed, portability, etc.) of the automatic tagger itself [Paroubek & Rajma, 2000].

4.2. A dictionary for the Amazigh language

Another tool which is, par excellence, made avail of in the lexical analysis of corpora is the electronic dictionary. We may well say that an automatic processing can in no way hold without the aid of the dictionaries and the grammar specific to the language under study.

The core point here is a dictionary which, as construed in NLP, is designed for direct computer use. Available Amazigh electronic dictionaries may well be deployed either through a strict conversion to a standard electronic format by manual typing or after an OCR procedure (Optical Character Recognition) followed by a through manual correction phase.

This electronic dictionary project, which is underway in IRCAM, is a promising tool both for researchers and experts in NLP. The central thrust of this project is to build a data basis encompassing lexical, morphological and syntactic structures of the standardised Amazigh language. A standard platform for the Amazigh language will, accordingly, be available easing, thereby, whatever future exploitation such as the generation of conventional lexicons and dictionaries meant to be published.

To the tools and corpora conceived and produced by IRCAM, we must presumably add the tools that have been set up by LDC/ELDA under a relationship of partnership. These tools were presented during the LREC conference held in 2008 in Marrakech (Cieri & Liberman, 2008). For convenience we shall repeat them in what follows:

- Encoding converter
- Word and sentence segmenter
- Tagset and tagger

- Named-entity tagger and tagged text

5. Tools and operating systems to be developed in short term

At this stage, we may well say that if some tools are available and gradually enable the Amazigh language to find its place in the world of NTICs, there remains a whole range of other tools to be generated before the language could achieve an influential degree of development and diffusion. The core goal to be achieved, thus far, is partly to undertake the task of collecting sufficient Amazigh corpora for the prospective work of NLP and partly to conceive and develop an automatic POS tagger for the Amazigh language. Other applications are to be contended with in the future such as spelling mistake correctors and speech synthesis tools among others.

6. Conclusion

To wind up, it is fair to say that the processing of Amazigh language, notwithstanding the modest number of tools developed both internally and through international collaboration is still in its rudiments. The development of competitive and efficient tools centers on the quantitative aspect of corpora which are normalised along the linguistic standards applicable in Morocco. In fact, it appears that the criteria of quantity and quality of corpora are at the heart of the computer processing of Amazigh. Put in another way, it is only through affording homogeneous and quantitatively sufficient corpora that the reliability of the produced tools may be guaranteed. An adjustment of existing corpora with an eye to making them in fine accord with the norms applicable in Morocco is thereby essentially necessitated. There is no doubt that the contribution of national and international partners in the promotion of a not-so-much privileged language like Amazigh will bring about a complex assortment of desirable effects to this language both quantitatively and qualitatively under relatively short periods of time..

Acknowledgements

Thanks are due to Mr. Khalid Ansar for having accepted to translate this contribution into English.

References

Ameur M., Bouhjar A. & al. (2006), *Graphie et orthographe de l'amazighe*, Rabat : IRCAM, pp. 47-48.

Bouhjar, A. (2008), *Amazigh Language Terminology in Morocco or Management of a 'Multidimensional' Variation*, LREC 2008, can be consulted at the site: http://www.lrec-conf.org/proceedings/lrec2008/authors.html#Bouhjar_Aicha

Bouhjar, A. (to appear), *La variation géolectale : un atout dans l'émergence d'un amazighe standard* .

Cieri Chr. and Liberman M.(2008), *15 Years of Language Resource Creation and Sharing: a Progress Report on LDC Activities*, LREC 2008, can be consulted at the site : http://www.lrec-conf.org/proceedings/lrec2008/summaries/861.html

http://www.ircam.ma/

Outahajala M. (2008), "Les normes de tri, du clavier et Unicode"in *La typographie entre le domaine de l'art et de l'informatique*, Rabat : IRCAM.

Paroubek P. & Rajma M. (2000), *Etiquetage morphosyntaxique*, *Ingénierie des langues.*

Web page of Tifinaghe-IRCAM fonts, http://www.ircam.ma/fr/index.php?soc=telec&rd=3