Iles d Imesli, 7, 2015, pp. 193-205

Base de données kabyles : collectes de données et applications. Synchronisation texte / son

Ramdane Boukherrouf[1] & Nora Tigziri[2]
1,2. Laboratoire d'Aménagement et d'Enseignement de la langue Amazighe
1,2. Département de Langue et Culture Amazighes
1,2. Université Mouloud Mammeri de Tizi-Ouzou
1. ramdaneboukherrouf@gmail.com
2. nora.tigziri@gmail.com

Agzul

Deg uḍris-a, ad d-nmel kra n yiberdan n usnas s way s yezmer ad d-yeglu usenfar-nneɣ yebnan ɣef uskar n wammud ara yilin d taɣayemt n tnefka yerzan tutlayt tamaziɣt. Tinefka-a, kkant-d seg wammud i d-nessekles, syin ɣer-s nura-t nerna nga-as tizmilin n usegzi. S wakka, ad izmiren yiselmaden d yimnuda n tmaziɣt ad t-zerwen deg taɣult n trukalt tutlayant akked tezrawin timserwasin. Γur-neɣ ddeqs n wammuden, ttwasxedmen akka tura, maca la nettkemmil asegrew d ugmar n wammuden niḍen ara d-yefken udem iwatan ɣef tutlayt. Ammuden-a, ad d-qqimen d aɣbalu utlayan. Nra daɣen ad nger afus deg usileɣ n yimnuda ilmeẓyen yettnadin deg taɣult-agi n tesnilest n wammuden. Asenfar-agi-nneɣ, d agni n umsenfel n tmussni gar yimnuda, ad lemden deg-s amek ara d-gemren d wamek ara snezwin tinefka tisenselkimin.

Abstract

We present in this article, some of the applications of our project, which involves the establishment of a corpus of oral database, digitized, transcribed and annotated for the Amazigh language, usable for scientific purposes and addressing mainly to teachers and researchers in Berber languages, particularly in the field of linguistic geography and comparative studies. We have a number of corpus exploited for various applications but we continue to harvest a large enough corpus to be representative of the language and which can be saved as a linguistic resource. We also want to contribute to the training of young researchers working in the field of corpus linguistics by providing a space for sharing their fieldworks and becoming familiar with the various computer tools necessary for processing and dissemination of data collected .

0. Introduction

Nous présentons dans cet article, une partie des applications de notre projet[1] qui consiste à la mise en place d'une banque de données de corpus oraux, numérisés, transcrits et annotés pour la langue amazighe qui soit exploitable à des fins scientifiques s'adressant principalement aux enseignants chercheurs

[1] Nous l'avons réalisé dans le cadre du projet CNEPRU intitulé : Transcription synchronisée des corpus oraux, agréé en janvier 2015 sous la direction du professeure Noura Tigziri.

berbérisants, notamment dans le domaine de la géographie linguistique et des études comparatives. Nous avons un certain nombre de corpus qu'on exploite pour différentes applications mais nous continuons à récolter un corpus suffisamment large pour qu'il soit représentatif de la langue, et afin qu'il permette sa sauvegarde sous forme de ressource linguistique. Nous voulons aussi contribuer à la formation des jeunes chercheurs qui travaillent dans le domaine de la linguistique de corpus en leur offrant un espace de partage de leurs travaux de terrain et les familiariser sur les différents outils informatiques nécessaires au traitement et la diffusion des données recueillies.

Les études de dialectologie et les travaux comparatifs amazighes ont été entreprises par René Basset (1887), suivies et développées par André Basset avec ses diverses contributions, notamment les deux travaux primordiaux : Géographie *linguistique de la Kabylie* (1929) et *Atlas linguistiques des parlers berbères (Algérie du nord)* (1936/1939), les différents travaux menés au Sahara et touareg (1933,1948) et le sud du Maroc (1942) sans oublier l'essentiel des publications dans les Articles de dialectologie berbères (1959). Cette perspective comparative a été suivie par Lionel Galand dans ses différents travaux (1950-2010) en l'explicitant clairement dans son travail plaidoyer pour la comparaison (2001). Cette démarche a eu beaucoup d'adhésion de la part des berbérisants : Karl G. Prasse, (1969) *l'origine du « h » touareg*, Salem Chaker (1972, 1997) avec ses travaux *la langue berbère au Sahara et* quelques *faits de grammaticalisation en berbère*, Maarten Kossmann (1989-2003), Miloud Taïfi (1994) avec son travail *unité et diversité du berbère. Détermination des lieux linguistiques d'intercompréhension*, Abdelaziz Allati (2002) *Diachronie tamazighte ou berbère*, Vermondo Brugnatelli (1986-2006), le travail de Mena Lafkioui (2007) avec l'*Atlas linguistique des variétés berbères du Rif.*

Notre projet vient comme un prolongement à cette perspective géolinguistique et comparative avec la mise à la disposition des chercheurs d'un certain nombre de corpus représentant pour le moment les quatre coins de la Kabylie.

1. Le travail sur le terrain :

Pour atteindre notre but nous enregistrons des corpus de locuteurs monolingues.

Ceci a un double objectif :

- cibler toutes les régions de la Kabylie grâce à eux qui proviennent des quatre coins de notre terrain d'enquête.
- compléter la formation de nos étudiants. Des consignes strictes sont

données aux enquêteurs : Faire transcrire le même corpus par deux étudiants, indépendamment l'un de l'autre. Un membre de l'équipe comparera ensuite ces deux transcriptions pour repérer d'éventuels écarts récurrents qui peuvent être l'indice de difficultés. Contrôler toutes les transcriptions faites par les étudiants indépendamment par deux membres de l'équipe.

Nous avons établi pour chaque locuteur une fiche de collecte où doivent apparaître les métadonnées préalablement définies. Pour compléter ces données, nous avons établi des listes de mots en fonction de plusieurs paramètres dont les différents champs sémantiques que nous soumettons dans les divers points d'enquête.

2. Représentation spatiale des données recueillies

Nous utilisons, actuellement QGIS[2]pour la représentation spatiale de ces points d'enquête et de la variation phonétique ; La définition des coordonnées de ces points (longitude et latitude) n'a pas été une tâche facile. En effet, les toponymes présentent une grande variation dans le temps et dans l'espace. Il nous arrive de ne pas pouvoir situer exactement un point d'enquête sur la carte parce le nom a changé ou a été transformé. En effet, les diverses sources (cartes topographiques, enquêtes de Basset, documents administratifs fournis par la Wilaya) présentent parfois, des variations importantes dans les toponymes et ceci est une difficulté supplémentaire à surmonter quand on passe à une représentation cartographique.

Ainsi par exemple, nous avons pour un même toponyme les écritures suivantes :

- AitIraten, At Iraten, Ait-Iraten, At-Iraten, At Irathen, AitIrathen, At-Irathen…
- Ait Mellal, At Mellal, Ait-Mellal, At-Mellal, Ait Mellel, AtMelel…

- Iguersafene, Igarsafen

[2]Un logiciel de cartographie numérique accessible gratuitement sur le site : http://www.qgis.org/fr/site/

Une fois les données de géolocalisation (latitude et longitude) définies pour les points d'enquête ciblés, nous les représentons sur une carte à l'aide de QGIS. Nous donnons ci-dessous un tableau comprenant des exemples de points d'enquêtes avec les coordonnées de localisation.

Latitude	Longitude	Name
36,819	4,213	Abizar
36,823	4,292	Adrar n Takdhiâa
36,811	4,172	Afir
36,614	4,042	Aglagal
36,809	4,322	Aghrib
36,71	3,768	Ain-Sakoura

La carte représentant ces points est donnée ci-dessous (Figure 01) :

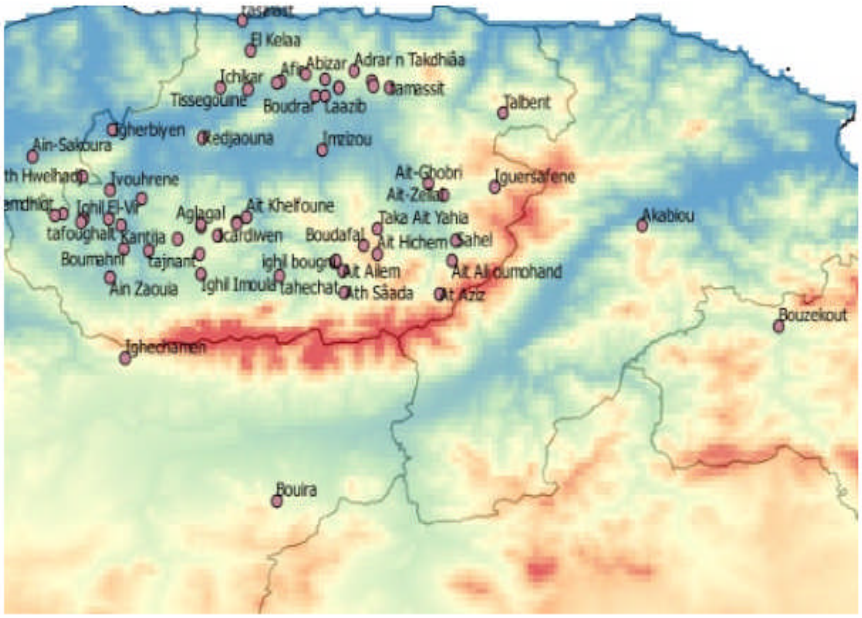

Figure 01

3. Synchronisation texte / son

Pour chaque point d'enquête représenté dans QGIS, nous avons associé un certain nombre d'actions qui nous sont utiles à savoir : le corpus son (fichier son) accompagné d'une transcription en notation usuelle, d'une transcription phonétique et d'une fiche de collecte associée à ce point.

Nous avons jugé plus intéressant pour l'exploitation de ces corpus oraux – que nous prévoyons de mettre sur un site - de prévoir une autre application, l'alignement son/texte afin de faciliter l'apprentissage de la langue amazighe à des non-amazighophones.

Cette application sera réalisée par l'intermédiaire d'un autre logiciel : le WINPITCH développé par le Professeur Philippe Martin qui a eu la gentillesse de le mettre à notre disposition (figure 02).

Figure 02

Winpitch comporte un certain nombre d'applications dont celles qui nous intéressent pour le moment : l'analyse acoustique, l'alignement son/texte. En lançant, Winpitch, nous voyons apparaitre la fenêtre suivante :

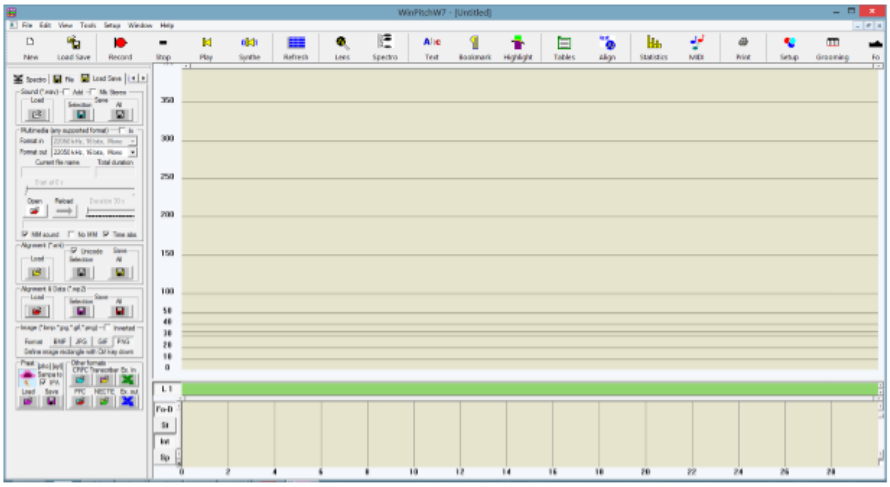

Figure 03

En cliquant sur la touche Alignement, une fenêtre s'active à gauche :

4. Niveaux de transcriptions adoptés

Dans l'objectif d'uniformiser la présentation de nos données linguistiques nous avons adopté une seule méthode de transcription et de présentation des données recueillies.

4.1. Les métadonnées :

> Chaque corpus est précédé d'une fiche de collecte qui est réservée à la présentation de l'ensemble des informations nécessaires du corpus. En effet, elle nous renseigne sur le lieu exact, la date, l'informateur, le thème et les conditions de l'enregistrement.

1. divers		
date de collecte :		
lieu :		
support de l'enregistrement :		
durée de l'enregistrement :		
lieu de l'enregistrement :		
sujet de l'enregistrement :		
Y avait-il un public ?		
Référence		
2. enquêté		
(Nom :)		
Date de naissance :		

Sexe :		
Village d'origine :		
Tribu :		
Domicile actuel (village, région, Commune, Daïra):		
Caractéristiques du point d'enquête	Longitude : Latitude :	
Dialecte parlé, (nom donné par le locuteur à son parler)		
Autre (s) langue (s) parlée (s) :		
(Au travail :)		
(À la maison :)		
Séjour (s) à l'étranger		
Durée du/des séjour(s)		
Scolarité et formation		
Langue(s) de l'enseignement reçu :		
Profession :		
Personne(s) ayant joué un rôle dans l'apprentissage linguistique (par exemple son père, sa mère, personne avec qui le locuteur a passé son enfance)		
- lien de parenté, relation avec la personne :		
- lieu d'origine :		
- scolarité (et langues d'enseignement) :		
situation familiale (mariage(s), enfants) :		
langue (s) parlée (s) par le conjoint :		
attitude du locuteur par rapport à sa langue et à sa façon de parler :		
3. Collecteur		
nom, prénom :		
langue (s) parlée (s) :		
origine :		
relation enquêteur-enquêté :		

4. Debriefing		
conscience du micro :		
attitude du locuteur par rapport à l'enregistrement :		
attitude du locuteur par rapport à l'entretien, aux questions posées…		
5. Autres infos		

4.2. Les données

La transcription[3] du corpus, comporte trois niveaux. Les deux premiers niveaux consistent en la transcription phonétique et la notation usuelle du corpus, le troisième est réservé à la traduction juxtalinéaire. Pour rester le plus fidèle possible au texte oral, nous avons donc opté pour une transcription phonétique et une notation usuelle. Pour reproduire fidèlement les réalisations orales du corpus, nous avons pris en considération les faux départs, les hésitations, les pauses, représentées par (/). Le choix des deux transcriptions est justifié par la prise en charge des réalisations (phonétiques) orales du corpus, et le décodage du corpus par les chercheurs linguistes non berbérophones pour la transcription phonétique et la fluidité de son décodage par les praticiens pour la notation usuelle.

Dans la transcription phonétique, nous avons reproduit globalement les graphèmes de l'Alphabet Phonétique International (API).

- Le graphème barré indique l'emphase [ɾ, s̶, t̶]
- Les affriquées sont transcrites avec deux graphèmes associés par une ligature [t͡s, d͡ʒ]
- Les labiovélaires sont accompagnées d'un graphème [w] en exposant [kʷ]
- La tension est transcrite avec deux graphèmes identiques [mm, ss, ll]

[3] C'est la police Unicode (Doulos Sil) qui est adoptée pour la transcription des corpus.

- A côté des sons emphatiques les voyelles [a, i, u] sont réalisées, respectivement [ɑ, e, o].

Afin de mettre en exergue les des différentes unités linguistiques et de les aligner avec les autres niveaux de transcription, nous avons adopté une transcription segmentée en syntagmes morphosyntaxiques.

Pour la notation usuelle, nous avons adopté essentiellement les recommandations[4] (1996) du Centre de recherche berbère de l'Institut National des Langues et Civilisations Orientales (INALCO).

La place de la traduction juxtalinéaire est justifiée par la mise en exergue des structures morphosyntaxiques de la langue source du corpus, destinée principalement aux chercheurs non natifs pour leur indiquer la position des unités linguistiques de la langue, la source du texte et leurs combinaisons. Les conventions et les abréviations utilisées sont données ci-dessous :

- Le verbe à l'infinitif suivi de la marque aspectuelle :

- P : Prétérit
- A : Aoriste
- AI : Aoriste Intensif
- PN : Prétérit Négatif

Suivi de PR. Pour la forme participiale

Et des modalités préverbales et/ou postverbales :

- NR : Modalité du non Réel (modalité aspectuelle du verbe)
- AC : Modalité d'Actuel Concomitant (Modalité aspectuelle du verbe).
- ICI : Modalité d'orientation spatiale vers le locuteur.
- LA-BAS : Modalité d'orientation spatiale (éloignement)
- NEG : Négation

[4] www.centrederechercheberbere.fr/tl_files/doc-pdf/notation.pdf

- Le nom suivi de la marque d'état :

- EL : Etat Libre
- EA : Etat d'Annexion

4.3.Combinaison des trois niveaux de transcription

Pour nous permettre de d'aligner les trois niveaux de transcription, nous avons adopté une transcription intercalée, en ce sens que chaque ligne présente un niveau de transcription.

— Ligne 1 : Transcription phonétique
— Ligne 2 : Notation usuelle (italique)
— Ligne 3 : Traduction juxtalinéaire

Exemple :

Ligne 01 : zwadʒ nziɣ ðvavas ak͡kʷ ðjəmmas igxə͡ttvən
Ligne 02 *: Zwağ n zik d baba-as akk d yemma-s i yxeṭṭben*.
Ligne 03 : Mariage-EL d'autan c'est père-son et mère-sa qui choisir-AI-PR.

5. Présentation d'un exemple de synchronisation : texte /son

Après avoir présenté précédemment la représentation spatiale des corpus, le logiciel de synchronisation et les niveaux de transcription adoptées dans la présentation des corpus, ce point est réservé à la présentation de l'application de la synchronisation texte / son avec le logiciel Winpitch. En effet, la synchronisation présente quatre niveaux principaux :

- L1 : Fichier son
- L2 : Transcription phonétique segmentée
- L3 : Notation usuelle
- L4 : Traduction juxtalinéaire.

Ci-dessous (figure 04), un exemple de figure de winpitch qui présente un exemple d'un corpus synchronisé son /texte.

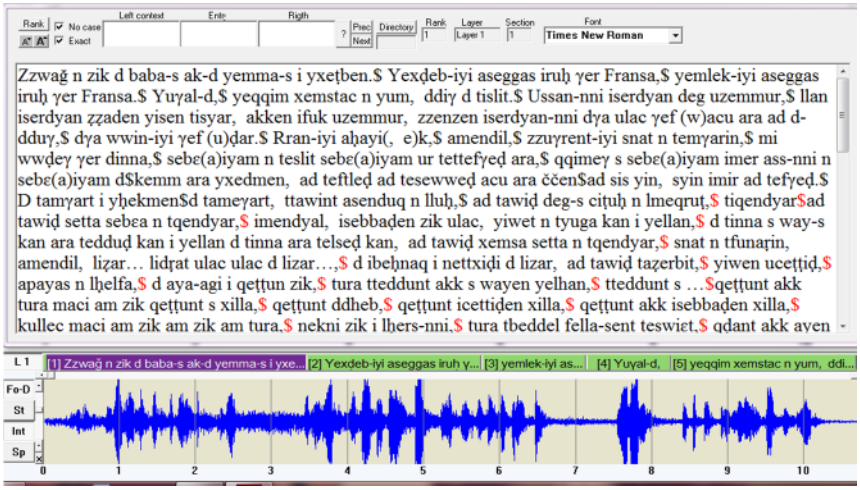

Figure 04

Le « dollar » représente les frontières de segmentation qui sont établies par l'auteur au fur et à mesure qu'il avance dans Winpitch. Evidemment, cette segmentation est réalisée au feeling, c'est-à-dire à la sensation de l'auteur dans la détection de pauses.

Conclusion et perspectives

En guise de conclusion vraiment superficielle, il convient de dire que nous sommes tout au début du projet, qui consiste à la mise en place d'une banque de données de corpus annotés qui seront mis à la disposition des chercheurs berbérisants dans le cadre de leurs recherches. Comme nous l'avons mentionné dans l'introduction, il s'agit d'un espace de formation pour nos jeunes chercheurs dans le cadre de la linguistique de corpus et de géographie linguistique et les outils informatiques à exploiter dans le cadre des nouvelles technologies de l'information et de communication.

En matière de perspectives, nous envisageons à l'avenir de classer ces données recueillies par champs thématiques, et ce pour permettre aux chercheurs de mener leur travaux de géolinguistique dans les différents domaines de la vie quotidienne.

Bibliographie

Allati, Abdelaziz, 2002, *Diachronie Tamazighte ou berbère*, Publications de l'Université Abdelmalek Essaâdi, Tanger.

Ameur, Meftaha, 1990 : « A propos de la classification des dialectes bérbères », *Etudes et Documents Berbères* 7, pp.15-27.

Basset, André, 1929 : *Etudes de géographie linguistique en Kabylie (sur quelques termes concernant le corps humain)*, Paris, Librairie Ernest Leroux.

— , 1952 : *La Langue berbère, (Handbook of African Languages, PartI)* London: Oxford University International African Institue.

Basset, René, 1887 : Manuel de langue kabyle (dialecte Zouara), Paris, Maisonneuve et Ch. Leclerc.

Brugnatelli, Vermondo, 1993 « Quelques particularités des pronoms en berbère du Nord », *A la croisée des études libyco-berbères, Mélanges offerts à Pellette GALAND-PERNET et Lionel GALAND*, Paris, pp. 229-245.

— , 1998 : «Encore à propos des pronoms berbère », *Groupe Linguistique d'Etudes Chamito-Sémitiques (GLECS)*, Volume 32, pp. 151-158.

Chaker, Salem, 1972 : « La langue berbère au Sahara», Méditerranée, Volume 11, Numéro 1, pp. 163-167.

— , 1983 : *Un parler berbère d'Algérie (Kabylie) : syntaxe*, Aix-en-Provence : Publications de l'Université de Provence.

— , 1991 : *Manuel de linguistique berbère I*, Alger, Bouchène.

— , 1996 : *Manuel de linguistique berbère II : syntaxe et diachronie*, Alger, ENAG.

— ,1997 : « Quelques faits de grammaticalisation su système verbal berbère », *Mémoires de la Société de Linguistique de Paris (Grammaticalisation et reconstruction)*, pp. 103-121.

Cohen, David, 1988 : «Le chamito-sémitique», *Les langues dans le monde ancien et moderne*, Jean Perrot (dir.), 3ème partie, *Langues chamito-sémitiques*, Paris, CNRS, pp. 9-29.

Galand, Lionel 2001 : « Plaidoyer pour la comparaison», *Etudes berbères*, Actes du 1. Bayreuth –Frankfurter Kolloquim zur Berberologie, pp. 63-71.

— , 2002. : *Etudes de linguistique berbère*, Paris, Peeters Levain.

Kossmann, Maarten, 1999 : *Essai sur la phonologie du proto-berbère*, Allemagne, RUDIGER Koppe VERLAG KOLN.

Lafkioui, Mena, 2003 : *Atlas linguistique des variétés berbères du Rif*, Berber Studies Volume 16, RUDIGER Kopee VERLAG KOLN.

Nait-Zerrad, Kamal, 2004 : *Linguistique berbère et applications*, Paris, L'Harmattan.

Prasse, Karl-G, 1969 : *A propos de l'origine du H touareg (tahaggart)*, Copenhague, Munksgaard.

Taïfi, Miloud, 1994 : « Unité et diversité du berbère : détermination des lieux linguistiques d'intercompréhension, *Etudes et Documents Berbères*, N°12, pp. 119-138.

Tigziri, Noura, 2014 : « La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique », *ASINAG* N° 9, Rabat, Maroc, IRCAM.