Base de données kabyles : collectes de données et applications

Noura TIGZIRI
Département de Langue et Culture amazighes
Université Mouloud Mammeri de Tizi-Ouzou

Cet article porte sur un projet «corpus oraux» que je dirige et auquel ont participé : Remi Jolivet, Boukherouf Ramdane, Chalah Seidh, Merzouki Samia, Hassani Said

Notre projet consiste en 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 linguistes. Nous souhaitons 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. Cette recherche fait intervenir deux institutions : le département de langue et culture de Tizi-Ouzou et la section linguistique de la Faculté de lettres de l'université de Lausanne. Ce projet a été intégré dans le laboratoire de recherche «Aménagement et enseignement de la langue amazighe» agrée en 2009.

Le travail sur le terrain :

Pour atteindre notre but nous enregistrons des corpus de locuteurs monolingues. Ces corpus sont recueillis par nos étudiants de licence de notre département.

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 (Annexe 1) où doivent apparaître les métadonnées préalablement définies. Pour compléter ces données, nous avons établi des listes de mots (Annexe 2) en fonction de plusieurs paramètres dont les différents champs sémantiques que nous soumettons dans les divers points d'enquête.

Nous avons, pour le moment utilisé Google Earth 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.

Google Earth présente un certain nombre d'avantages mais aussi des inconvénients.

Avantages

La cartographie sur Google Earth est assez bien faite. Nous arrivons à situer les villages les plus reculés de la Kabylie.

Nous avons accès à tous les éléments qui peuvent nous aider dans l'interprétation de nos résultats (montagnes, rivières, oueds, …)

Google Earth dispose d'une base de données de toponymes ce qui facilite la recherche d'un toponyme et les coordonnées- latitude et longitude - du point d'enquête.

Inconvénients

Google Earth n'est pas statique ce qui oblige l'existence d'une connexion internet efficace pour travailler

Sa base de données n'est pas complète et pas toujours actualisée

Représentation spatiale

Pour toute représentation sur google earth, on doit créer un fichier excel des points d'enquête à représenter. Chaque de ces points doit être défini par sa longitude et sa latitude. Cette opération est assez ardue étant donné que ces coordonnées ne sont pas connues préalablement. Aussi, on exploite toutes les bases de données de toponymes dont celle de google earth pour mener à bien cette opération mais de nombreuses difficultés sont à signaler :

- non homogeinisation dans l'écriture de toponymes :

Ex : Ait Mellal peut s'écrire : Ait Mellal, At Mellal, Ait-Mellal, At-Mellal, Ait Mellel, At Melel…

Iguersafene, Igarsafen

Ces différentes écritures ne sont pas prises en considération dans les bases de données d'où la difficulté de situer le point.

- Un toponyme qui s'écrit différemment : ex Imzizou, Oumzizou d'où la question légitime de savoir : Ces deux écritures correspondent à deux variantes d'un même toponyme ou est-ce deux toponymes différents ?

- On ne retrouve pas certains points d'enquête sur les bases de données consultées.

Ex : Igherbiene qui s'écrit aussi igherviene ne se trouve pas sur la carte de googleearth. Sa représentation est donc approximative.

(Annexe3, 4, 5)

Une fois le fichier Excel créé, on crée un fichier pour Google Earth/Maps en .Kml.

En cliquant sur ce fichier en .kml, les données seront affichées dans googleearth

Application avec PRAAT :

PRAAT () est exploité en analyse acoustique. En créant de nombreuses tires, on arrive aligner le signal temporel, le sonagramme, la notation usuelle, le découpage en unités préalablement définies ou étiquetage linguistique (racines, schèmes, syntagmes…) (Annexe 8). Des scripts sont également utilisés à des fins de segmentation en énoncés par exemple. Evidemment toute la problématique de la définition de l'énoncé en ce qui concerne l'oral est difficilement maîtrisable. Pour notre part, les pauses sont prises comme indicateur de séparations d'énoncés (Annexe 9). Evidemment PRAAT a aussi la qualité d'aligner son/transcription.

Au bout de trois années de recherche et d'enquêtes, nous avons réalisé :

100 points d'enquête, et 100 enregistrements de 20mn chacun pour la plupart transcrits (Annexe 6: exemple de corpus). Nous avons établi une «carte exemple» d'un certain nombre de points d'enquête (Annexe 7).

Conclusion

Le projet continue puisque l'objectif est d'arriver à l'élaboration d'une base de données de corpus oraux avec comme application, la représentation spatiale

1- Des points d'enquêtes accompagnés de fiche de collecte, du corps oral et de sa transcription

2- Le corpus oral doit être synchronisé et aligné avec sa transcription

3- D'établir des cartes linguistiques, principalement phonétique et lexicale

Ces opérations permettront de mieux connaitre la variation, donc de mieux l'appréhender dans l'aménagement de la langue amazighe.

ANNEXES

ANNEXE 1

Fiche de collecte

1. divers		
Date de collecte :	4/11/2010	
Lieu :	Sétif	
Support de l'enregistrement :	portable	
Durée de l'enregistrement :	06:45	
Lieu de l'enregistrement :	Maison	
Sujet de l'enregistrement :	Azetta	
Y avait –il un public ?	Non	
Référence		
2. enquêté		
(nom :)	Nacel aldjia	
Date de naissance :	26-02-1952	
Sexe :	Femenin	
Village d'origine :	Bouzekout	
Tribut :	At Ouajhan	
Domicile actuel (village, région) :	Village Bouzekout. Commune Bousselam daïra Bouandas, w Sétif	

Dialecte parlé, (nom donné par le locuteur à son parlé)	Kabyle	
Autre (s) langue (s) parlée (s) :	Français, Arabe	
(Au travail :)		
(A 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 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 :	Maternité	
Lieu d'origine :		
- scolarité (et langues d'enseignement) :		
Situation familiale (mariage(s), enfant) :		
Langue (s) parlée (s) par le conjoint :	Les mêmes langues que la mère	

Attitude du locuteur par rapport à sa langue et à sa façon de parler :		
3. collecteur		
Nom, prénom	Nacel Amar	
Langue (s) parlée (s) :	Kabyle	
Origine :	Bousselam	
Relation enquêteur enquêté :	Mère fils	
4. Débriefing		
Conscience du micro :		
Attitude du locuteur par rapport à l'enregistrement :	à l'aise	
Attitude de locuteur par rapport a l'entretien, aux questions posées …		
5. Autres infos		

ANNEXE 2

Villages / Terme en français	(A) **V:**Abizar C^{ne}:Timizart	(B) V:Arour C^{ne}:Tirmitin	(C) V : Ichetouanen C^{ne}:Boudjima	(D) V : Agouni outekkous C^{ne}:Boudjima	(E) V:Tifra C^{ne}:Tigzirt	(F) V: Ait Abd Elmoumen C^{ne} : Tizi n tlatha	(G) V:Takichort C^{ne}: Ait Ouacif	(H) V:Tighzert C^{ne}:Ait Aissi	(I) V:Mekla C^{ne}:Mekla	(J) V:Mzeggen C^{ne}:Ilioula
Cuillère	F:Tayenjayt FP:Tiyenjayin	Tayenjayt Tiyenjayin	Tayeŋgayt Tiyeŋgayin	Tayeŋgayt Tiyeŋgayin	Tayeŋgayt Tiyeŋgayin	Tiflewt Tifelwin	Tijyelt Tijeylin	Tayeŋgayt Tiyeŋgayin	Tayeŋgawt Tiyeŋgawin	Tayeŋgawt Tiyeŋgawin
Louche	M: Ayenja MP: Iyenjayen	Ayenja Iyenjayen	Ayeŋga Iyeŋgayen	Ayeŋga Iyeŋgayen	Ayeŋga Iyeŋgayen	Iflew Ifelwen	Tiflut Tifelwin	Ayeŋga Iyeŋgayen	Ayenja Iyenjayen	Ayeŋga Iyeŋgayen
Marmite	F: Tasilt FP: Tasilin	Tuggict Tuggicin	Tasilt Tasilin	Tasilt Tasilin	Tasilt Tasilin	Tasilt Tasilin	Tuggi Tuggiwin	Tasilt Tasilin	Tasilt Tasilin	Tamermiḍt Timermiḍin
Anon	M: Ajhiḥ F: Tajḥiḥt MP: Ijḥiḥen FP: Tijḥiḥin	Ayyulatuṭaḥ Tayyult tatuṭaḥt Iyyal iṭuṭaḥen Tiyyal tiṭuṭaḥin	Ajhiḥ Tajḥiḥt Ijḥaḥ Tijḥaḥ	Ajhiḥ Tajḥiḥt Ijḥiḥen Tijḥiḥin	Ajhiḥ Tajḥiḥt Ijḥiḥen Tijḥiḥin	Ayyulamectuḥ Tayyult tamectuḥt Iyyal imectaḥ Tiyyal timectaḥ	Ayyulamectuḥ Tayyult tamectuḥt Iyyal imectaḥ Tiyyal timectaḥ	Ajhiḥ Tajḥiḥt Ijḥaḥ Tijḥaḥ	Ajhiḥ Tajḥiḥt Ijḥaḥ Tijḥaḥ	Ajhiḥ Tajḥiḥt Ijḥiḥen Tijḥaḥ
Lait	Ayefki	Iyefki	Ifki	Ifki	Ifki	Ayefki	Ayefk	Ifki	Ayefki	Ayefki

Oiseau	M: Afrux F: Tafruxt Mp: Ifrax Fp: Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax	Afrux Tafruxt Ifrax Tifrax
Ongle	M: Iccer Mp: Iccaren	Iccer Iccaren	Iccer Accaren	Iccer Accaren	Iccer Accaren	Iccer Iccaren	Iccer Accaren	Iccer Iccaren	Iccer Iccaren	Iccer Accaren
Dent	F: Tuɣmest Fp: Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas	Tuɣmest Tuɣmas
Ail	F: Ticcert	Ticcert	Ticcert	Ticcert	Ticcert	Ticcert	Ticcert	Ticcert	Ticcert	Ticcert
Rouge Gorge	M: Aεeẓẓi F: MP: Iεeẓẓiyen FP:	Aεeẓẓi Iεeẓwa	Aεeẓẓi Taεeẓẓit Iεeẓwa Tiεeẓwa	Aεeẓẓi Taεeẓẓit Iεeẓwan	Aεeẓẓi Taεeẓẓit Iεeẓwan	Aεeẓẓi Taεeẓẓit Iεeẓwa Tiεeẓwa	Aεeẓẓi Taεeẓẓit Iεeẓwan	Aεeẓẓi Taεeẓẓit	Aεeẓẓi Taεeẓẓit Iεeẓwa Tiεeẓwa	Aεeẓẓi
Oreille	M: Ameẓẓuɣ F: Tameẓẓuɣt MP : Imeẓẓuɣen FP:Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin	Ameẓẓuɣ Tameẓẓuɣt Imeẓẓuɣen Timeẓẓuɣin
Pluit	M: Ageffur MP: Igefran	Ageffur Igefran	Ageffur Igefran	Ageffur Igefran	Ageffur	Ageffur	Lehwa	Ageffur	Lehwa	Lgarra
Garçon	M: Aqcic MP: Arrac	Aqrur	Aqcic Arrac	Aqcic Arrac	Aqcic Arrac	Aqrur Iqrar	Aqcic Arrac	Aqcic Arrac	Aqcic Arrac	Aqcic Arrac

Femme	F: Tameṭṭut M: Argaz FP: Tilawin MP: Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tulawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen	Tameṭṭut Argaz Tilawin Irgazen
Dos	M: Aɛrur MP: Iɛrar	Azagur Izugar	Aɛrur Aɛruren	Aɛrur Aɛruren	Aɛrur Aɛruren	Azagur Izugar	Azagur Izugar	Aɛrur Aɛruren	Aɛrur Aɛruren	Aɛrur Iɛrar
Forêt	F: Lyaba FP: Leɣwabi	Lexla Lexlawi	Amaday Imuday	Lyaba Lyabat	Amaday Imuday	Lexla Lexlawi	Lyaba Leɣwabi	Amaday Imuday	Lyaba Leywabi	[ʲRava] [ʲRwavi]
Lentisque	F: Tidekt	Imidek	Tidekt	Tidekt	Tidekt	Imidek	Tidekt	Imidek	?	?
Grenouile	M: Amqarqur F:Tamqarqurt Mp: Imqarqar FP: Timqarqar	Amqerqer Tamqerqert Imqerqren Timqerqrin	Amqarqur Tamqarqurt Imqarqar Timqarqar	Amqarqur Tamqarqurt Imqarqar Timqarqar	Amqarqur Tamqarqurt Imqarquren Timqarqurin	Imqerqer Timqerqert Imqerqren Timqerqrin	Imqerqer Timqerqert Imqerqren Timqerqrin	Amqarqur Tamqarqurt Imqarqar Timqarqar	Amqerqer Timqerqert Imqerqren Timqerqrin	Amqarqur Tamqarqurt Imqarqar Timqarqar
Intestin	M: Ajyed MP: Ijeɣdan	Iẓrem Izerman	Aẓrem Izerman	Aẓrem Izerman	Ajyed Ijeydan	Aẓrem Izerman	Aẓrem Izerman	Aẓrem Izerman	Aẓrem Izerman	Ajeɛbub Ijeɛbuben
Tortue	F: Tafekrunt M: Afekrun FP: Tifkrunin MP: Ifekran	Tafekrunt Afekrun Tifekrunin Ifekrunen	Tafkert Afker Tifekratin Ifekran	Tifkert Ifker Tifekrin Ifekran	Tafekrunt Afekrun Tifekrunin Ifekrunen	Tifkert Ifker Tifekrunin] Ifekrunen	Tafkrunt Afekrun Tifekrunin Ifekrunen	Tifkert Ifker Tifekrin Ifekran	Tafekrunt Afekrun Tifekrunin Ifekrunen	Tifkert Ifker Tifekrar Ifekran

Annexe 3

Annexe 4

Annexe 5

- <?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
- - <CORPUS>
- **<CORPUS> <NOTATION USUELLE> : Nekkni zik, ad d-neker deg yiḍ, nṣaf n yiḍ ad nruḥ ta ad teɣɣar i ta, ad d-nagem d talla n wadda mi i d-newwed ad necɣel seksu, mi nfuk seksu-nni, ad nɛedi ad nneneḍ leɣbar mi nfuk leɣbar-nni ad nɛedi ad nerfed iqeṭṭaren tasebḥif; ad nruḥ ɣer cɣel. Nettewqam amarḍil Ad naweḍ a yelli ɣer uzemmur, ad nawi iqcer n uɣrum deg yiciwan n nay; ur nettawi ara llɛali-agi i ttawin akka medden tura, wellah ar d tidet a yelli. Ad nawi iqceran-nni n uɣrum deg yiciwan nnteɣ ad naweḍ aken nemwellah d tislatin d lxalat, deg mi ara nali ɣef lǧedra alama n fuk-itt-id deg yixef, mi ara d-nars, aɛeqqa, ad awḍeɣ ar lǧedra ad xezreɣ tazemurt ma ufiɣ aɛeqqa ar teqacuct ad qleɣ, ad t-id-yeqḍeɣ, ḥemlaɣ arrezq a yelli, maci am tura; lǧil n tura.**

.....

</NOTATION USUELLE>

- **<CORPUS> <TRANSCRIPTION PHONETIQUE>:**

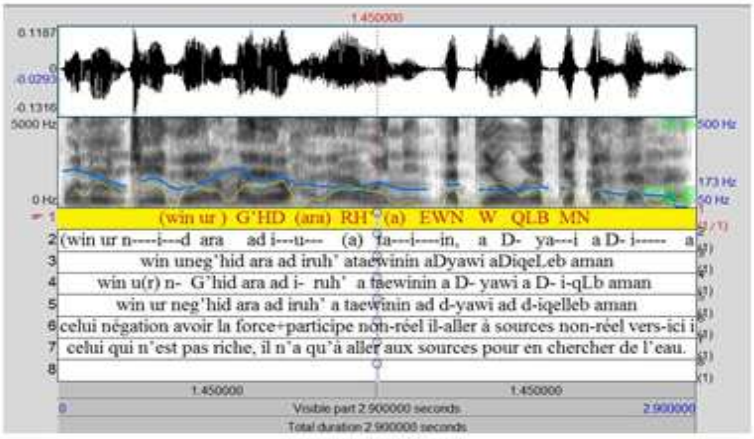
[nəkwnizi:yadnəkaʀðəgiðnssafəgiðandheaatsəʁʁaʀieəaadnay
wemðieəajjapaddamidneppəðaneʃʁ:ojsəksɔ:minfəksəksɔ:nnia
nʕədianənnəðjaʁvarminfəkjaʁvarnniaʕədiannfeðiqətaʀəneas
əvhieannrðhʁəʀʃʁojnetsəwqamamrðijanawoðajelliarɔzəmm
oranwiiqʃaʀəppaʁʁromðəgʃiwannntaʁontsawijarajʕajijagiitswi
nakkamədðanwollahartsiðətsajellianawiiqʃrannnippaʁʁromð
əgʃiwannntaʁanawəðakannamwolahtsisklðəjxajaeðəgmaranaji
afəldʒðraajamanfokitsidðəgaxəfmaradnarsaʕəqqaadawðaʁal
dʒeðraaðxəzraʁəazəmoremaofiʁaʕəqqaareqaʃɔʃəaðqəjaʁaəi
djəqðaʁhəmjaʁaʀʀəzqajəllimatʃameoraarʒʒijntoraadnarsajəl
lieamədieppasaʁinaʁjaʒʒaanətsðəgiriqʃəranmniəpwʁromeʃf
adjəmmamatsiɣrəppinəʀʀahmas

Annexe 6

POINTS D'ENQUÊTE

Annexe 7

ANALYSE AVEC PRAAT



Annexe 8

SEGMENTATION MORPHO-SYNTAXIQUE

- **Le corpus**
-
- **- < Bruit de fond> (00 : 00 s. – 00 : 03 s.)**
- **B-** ‖ Ad wen-d-hedreɣ | ɣef ***cinquante-huit***? ‖
- SPV synt. Prépos.
- **A-** ‖ Ah ? ‖
- **- < Bruit de fond> (00 : 06 s. – 00 : 08 s.)**
- **A-** ‖ **< euh : > *alors* < euh : >**
- ‖ Aqlaɣ | di Tesga-Mellul | d ssebt |
- Présentatif syntagme nominal syntagme nominal
- tnayen-uɛecrin ɣuct | ttesɛa ɣir ɣɣbeɛ ‖
- syntagme nominal syntagme nominal
- **C-** ‖ ɣir ɣɣbeɛ ‖
- syntagme nominal

Annexe 9