L'intérêt du corpus et une idée de sa constitution en lexicographie amazighe

Abdelaziz BERKAÏ
Université de Béjaïa

Introduction

L'objet de notre proposition de contribution est de montrer l'intérêt du corpus en lexicographie et de donner une idée de sa constitution en lexicographie amazighe, faite jusqu'à présent essentiellement hors corpus.

De grands dictionnaires, comme le *Trésor de la Langue Française* (*TLF*) ou le *Collins COBUILD English Language Dictionary*, sont élaborés *in extenso* à partir de corpus qui assurent l'objectivation et la précision de l'information lexicographique garanties par des contextes d'usage. Et si un corpus d'une dimension modeste permet de recueillir de façon satisfaisante l'information relative au phonétisme, à la morphologie ou à la syntaxe d'une langue, celui qui doit servir au lexique, de par sa nature ouverte, doit être à la fois important et ouvert, c'est-à-dire continuellement enrichi, pour pouvoir l'adapter aux changements qui touchent ce niveau de langue. L'avènement de l'informatique et de l'Internet ont grandement facilité le recueil et le traitement des données et considérablement agrandi les corpus qui passent de quelques dizaines ou centaines de milliers de mots, dont ils étaient constitués avant l'avènement de ces nouveaux moyens, à des centaines de millions, voire à des milliards de mots constitués notamment à partir du Web (Geyken, 2008 : 84). L'avènement de la mico-informatique notablement à partir des années 1980, qui a grandement facilité la constitution et surtout la manipulation des corpus,a donné naissance à un nouveau courant en linguistique, dénommé «linguistique de corpus» ou, à l'origine, «corpus linguistics», puisque c'est en Grande-Bretagne que le concept est d'abord conçu. «C'est une discipline qui est très liée à l'utilisation de l'informatique, mais qui reste une discipline des sciences humaines

et non de l'informatique. Les maîtres mots sont linguistique et corpus», écrit Williams, cité par Cori et David (2008 : 112). La micro-informatique qui a considérablement amélioré tant au plan qualitatif que quantitatif le corpus, en a fait non seulement un objet de validation d'hypothèses, conçues plus ou moins introspectivement auparavant, mais en plus un objet heuristique qui permet à partir d'un ensemble de données représentatives d'un type de discours d'accéder à des déductions «étonnantes, que la simple intuition n'aurait pas pu atteindre» (Blanche-Benveniste, 2000 : 15). C'est en celaque la «nouveauté» de cette «discipline» de «linguistique de corpus» se justifie. C'est la plus grande malléabilité et extensibilité de ce matériau qui en fait aujourd'hui l'indispensable matière première de toute étude sérieuse de linguistique. Elle se justifie aussi par le travail d'objectivation et de «formalisation» de cet objet qu'est le corpus qui n'a pas été fait avant. Mais, nommer une nouvelle discipline n'est pas suffisant pour la créer. La «linguistique de corpus» n'est en fait rien d'autre que la simple linguistique qui utilise les moyens de son époque en recourant, grâce à l'informatique, plus massivement et plus avantageusement au corpus[1]. En fait, il s'agit plutôt d'une méthodologie de recherche qui concerne quasiment l'ensemble des secteurs de la linguistique que d'une discipline ayant son objet et sa théorie[2].

Le corpus : quel intérêt pour la lexicographie ?

De grands corpus textuels informatisés sont constitués dans certaines langues, notamment européennes, et servent de matière première pour des études diverses, en particulier lexicographiques. Leurs avantages sont multiples :

- Ils permettent de constituer, selon la nature du dictionnaire, la nomenclature désirée et de connaître en outre la fréquence et la répartition de chaque mot. Cette information est d'une utilité évidente dans le traitement microstructurel de ces unités ;

- Ils permettent à partir des multiples contextes d'usage des unités de définir précisément leur sémantisme et de connaître leur combinatoire, ce qui facilite leur analyse morphosyntaxique et le

relevé des locutions et collocations dont elles sont des éléments constitutifs ;

- Ils permettent enfin par la comparaison de leurs états successifs de déterminer certaines unités particulières et leurs statuts (néologisme, xénisme, pérégrinisme, archaïsme, mot rare, etc.).

Le lexicographe dispose donc à travers le corpus, pour peu qu'il soit correctement constitué, d'un matériau qui lui permet d'étudier les unités lexicales dans leur milieu naturel et de connaître précisément leur fonctionnement sémantique, morphologique et syntagmatique et même sociolinguistique en sachant le contexte, la situation d'usage et le statut du locuteur (âge, origine géographique, profession, etc.). Cependant, même si le corpus a incontestablement amélioré le texte dictionnairique, il n'a pas pour autant «simplifié» le travail du lexicographe. «Là où il travaillait naguère avec beaucoup d'intuition et de bonnes facultés d'analyse, il a désormais besoin de puissance de déduction et de pouvoir de synthèse», écrit Béjoint (2007 : 20). Un bon corpus avec un mauvais travail d'inférence donnerait en effet un mauvais dictionnaire. De même qu'un mauvais corpus, même avec un bon travail d'inférence aboutirait au même résultat. Mais qu'est-ce qu'un bon et un mauvais corpus ? Mais d'abord, qu'est-ce qu'un corpus ? François Rastier en donne une définition intéressante, mais critiquable (v. notes 4 et 5) et qui concerne surtout les corpus textuels[3] : «Un corpus est un regroupement structuré de textes intégraux[4], documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive[5] en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications» (Rastier, 2004 : 2). La définition de V. Giouli et S. Piperidis nous semble plus générale et mieux adaptée pour définir le corpus lexicographique en parlant de fragments de discours : «Le mot corpus est utilisé pour renvoyer à des fragments de discours sous forme soit écrite soit orale, notamment au format électronique, réunis d'une manière systématique pour pouvoir en tirer des conclusions sur l'usage linguistique» (Duchet, 2008 : 129). La nature des ces «conclusions» à tirer détermine un choix qualitatif et quantitatif de discours et de genres à même d'optimiser les résultats.

Un bon corpus lexicographique doit satisfaire, dans une certaine mesure, à deux principes :

1. Le principe d'extensivité : le corpus doit inclure le maximum de domaines et de registres possibles impliqués par les objectifs et la nature du dictionnaire[6] ;

2. Le principe d'exhaustivité : il doit représenter de façon exhaustive le lexique d'un domaine particulier, concerné bien entendu par la nomenclature (v. Lafage, 1997 : 88).

Le respect de ces deux principes conduit à l'élaboration d'un corpus dit «représentatif», où l'essentiel du lexique visé serait représenté. Mais, «l'essentiel» ne veut jamais dire «la totalité» qui est la somme des lexiques de tous les idéolectes d'une communauté linguistique qu'il n'est pas possible de recueillir en entier. Nous ne sommes pas d'accord en l'occurrence avec Damon Mayaffre lorsqu'il affirme que «les corpus lexicographiques peuvent donc non seulement être des corpus clos mais des corpus finis» (2005 : 5). Il n'existe pas de»clos» ni de «fini» en fait de mots. Même dans la somme des matériaux de ce corpus dit «représentatif», le dictionnaire ne présente pas tous les éléments. Il opère toujours un tri où beaucoup de mots jugés non conformes aux critères de sélection de la nomenclature sont écartés (certains hapax, mots vieillis, xénismes, pérégrinismes…). En ajoutant à cela le fait que beaucoup d'autres mots, à l'exception des mots isolés, ne figurent pas dans le dictionnaire et qu'ils sont représentés par des sortes de «représentants» qu'on appelle «lemmes»[7] ou à juste titre «adresses», où ils sont censés être domiciliés.On peut s'interroger de ce fait sur le statut du dictionnaire comme source d'attestation des mots. J.-C. Corbeil (1971 : 136) invite justement à le «démystifier» en insistant sur sa qualité relative : «C'est un outil d'un certain type et d'une certaine qualité, rien de plus» et rétablit le rapport de dépendance entre mot et dictionnaire en faveur du premier : «l'existence d'un mot ne tient pas au dictionnaire, c'est l'existence du dictionnaire qui tient aux mots : pas de mots, pas de dictionnaire».

La lexicographie amazighe : Quelle place pour le corpus ?

Les «bons» dictionnaires amazighs, comme le dictionnaire kabyle-français de J.-M. Dallet, le touareg-français de Charles de Foucauld ou encore le tamazight-français de Miloud Taïfi, reposent, mais pas exclusivement, sur des corpus[8], même s'ils ne sont pas toujours signalés. Taïfi dont le dictionnaire est le résultat d'une thèse de doctorat d'Etat est le seul des trois auteurs cités ci-dessus à en parler explicitement (1991 : II). Mais son «corpus littéraire» (constitué de *timedyazin* et d'*ihellilen*), précise-t-il, lui a servi seulement à «compléter» ses «enquêtes lexicologiques» (*ibid.*)[9]. Il est donc loin de satisfaire aux critères d'extensivité et d'exhaustivité évoqués ci-dessus et qui sont nécessaires à la représentativité d'un corpus. Il ne s'agit donc pas en l'occurrence d'une lexicographie «de corpus», où celui-ci est utilisé comme objet heuristique servant à l'élaboration d'un savoir, mais à peine d'une lexicographie «sur corpus», où il sert surtout de support à la validation d'hypothèses adoptées a priori, pour reprendre les termes de l'opposition de Mayaffre : corpus comme *apport* vs corpus comme *support,* de l'anglais *corpus-based* vs *corpus-driven* de Tognini-Bonelli (2001) (2005 : 8).

Taïfi et les éditeurs du Dallet ont eu l'honnêteté de reconnaître que leurs dictionnaires sont incomplets. «Malgré la richesse de la nomenclature qui y est recensée, ce dictionnaire reste incomplet», écrit Taïfi (1991 : III). «Nous ne pouvons prétendre, bien entendu, avoir tout dit et n'avoir rien omis des richesses du parler des At Mangellat, qui en dépit de notre patiente recherche, déborde encore par sa vigueur de vie ce que nous en avons noté. Nous devons (…) reconnaître franchement les faiblesses trop évidentes, les lacunes pour une part inévitables de ce travail difficile», écrivent modestement les éditeurs du Dallet (1982 : XX). Nous pensons, pour notre part, que pour élaborer un dictionnaire qui réponde de façon satisfaisante aux attentes de son public, par ailleurs bien ciblé, le travail sur corpus est désormais nécessaire pour pallier les nombreuses lacunes et approximations qui caractérisent la lexicographie amazighe. La constitution de corpus textuels informatisés est nécessaire pour chaque dialecte, et dans la mesure du possible pour chaque parler. Ils peuvent

servir à diverses études linguistiques, littéraires et autres, et être complétés, pour tendre à la représentativité, en lexicographie, par des recueils thématiques constitués par une approche onomasiologique, c'est-à-dire en partant des notions concernant un domaine particulier pour atteindre leurs dénominations. Ces nomenclatures thématiques peuvent être constituées globalement à partir de celles qui existent dans d'autres langues en les complétant, le cas échéant, par les données lexicales spécifiques au parler concerné obtenues par des enquêtes ciblées. Car quelle que soit la dimension d'un corpus textuel, des mots courants peuvent échapper à ses mailles. Ce sont des mots usuels «qui se dérobent à la statistique», écrit J. Picoche, citée par C. Frey (1997 : 259). Des mots très connus mais pas nécessairement sollicités, sinon dans certaines circonstances ou situations où ils sont impliqués. Ce sont les mots qu'on appelle «disponibles» et qui complètent les «fréquents» dans l'ensemble des mots usuels. Le mot «fourchette», par exemple, et malgré la richesse du français dans le domaine culinaire n'a pas été recueilli dans un corpus de 312135 mots ayant servi à l'élaboration du français fondamental. Beaucoup de domaines peuvent être concernés par ces recueils «onomasiologiques» : la cuisine (ustensiles et recettes), la maison (construction, literie et objets divers), les arbres et arbustes, les plantes, les maladies (humaines, animales et végétales), le corps humain, les animaux,le sport et les jeux, les titres et fonctions, le temps et le climat, la mer, l'agriculture,les vêtements et les parures, etc. On peut établir ces recueils par des enquêtes ciblées en utilisant des sous-corpus sources constitués de référents (objet, image, description…) ou/et d'équivalents dans d'autres parlers amazighs proches ou/et dans d'autres langues (arabe, français…) des dénominations recherchées. Une fois obtenues, leur «mise en corpus» pourrait comporter les informations suivantes :

- Une transcription phonétique partielle qui prend en charge les sons objets d'une variation (spirantisation/occlusion, emphase ou son absence dans un contexte emphatique…). Il est inutile de donner par exemple une transcription phonétique pour des mots comme *ifelfel* «piment, poivron», *amellal* «blanc», *afermac* «édenté», etc. La transcription usuelle suffit largement dans ces cas[10] ;

- La catégorie grammaticale qui situe l'item considéré parmi les neuf[11] parties du discours connues (nom, pronom, verbe, déterminant/article, adjectif, adverbe, conjonction, préposition et interjection) ;

- La flexion des mots variables. Pour le nom on donne (après les avoir recueillies) les flexions du genre, du nombre et de l'état d'annexion, lorsqu'elles existent. Pour le verbe on donne les formes de l'aoriste, du prétérit, du prétérit négatif et de l'aoriste intensif.Pour les formes dérivées, on donnera lorsqu'elles sont attestées les formes du factitif, du passif, du réciproque, mixte (combinaison des précédentes) ;celles du nom d'action, du déverbatif concret, du nom d'agent, d'instrument, de l'adjectif et même de la forme verbale *potentielle* exprimant la faisabilité d'un procès (action ou état), rarement données dans les dictionnaires amazighs : *ttwačča*y (être mangeable, comestible), *ttwasway* (être buvable, potable), *ttwaxdam* (être faisable ; réparable)... Toutes ces formes pourtant très vivantes en kabyle ne sont pas attestées dans le Dallet (Berkaï, 2011 : 33) ;

- L'information sémantique où il est utile de donner, en plus du sémantisme de l'item, l'information analogique nécessaire au travail d'encodage pour un dictionnaire de thème : le synonyme, l'antonyme, l'homonyme, la variante phonique et/ou morphologique, l'hyperonyme, l'hyponyme, etc. ;

- L'information pragmatique, c'est-à-dire toute information utile à une actualisation adéquate d'un item : registre de la langue (familier, vulgaire, grossier, enfantin, féminin…), archaïsme, néologisme, pérégrinisme, etc. On peut ajouter à ces informations lorsque c'est possible et c'est utile, en particulier pour les verbes et leurs problèmes de valence et de combinatoire, un ou des exemples d'usage des items concernés. Cette information est très utile pour l'élaboration d'un dictionnaire d'encodage ;

- L'information iconique lorsqu'elle est nécessaire à la description précise d'un référent a fortiori lorsqu'il est spécifique à la culture amazighe.

Beaucoup de mots relevant des domaines cités ci-dessus peuvent être évidemment recueillis dans un corpus textuel important. Ce corpus est d'autant plus représentatif qu'il est vaste et varié, c'est-à-dire satisfaisant aux deux principes d'extensivité et d'exhaustivité. Les mots disponibles concernent surtout les substantifs concrets qui ne sont pas fréquents dans le discours, contrairement aux mots grammaticaux et aux verbes qui y sont relativement bien représentés. Ceci d'une part. D'autre part, le sémantisme et la valence d'un mot fonctionnel ou d'un verbe et même des noms polysémiques, qui constituent la majeure partie des unités de cette catégorie discursive, ne peuvent s'établir véritablement qu'à partir d'un contexte d'usage, c'est-à-dire d'un corpus. C'est que les mots, pour reprendre Humboldt cité par Meschonnic (2008 : 11), «ne précèdent pas le discours, mais ils procèdent du discours».

Conclusion

Notre idée de la constitution du corpus lexicographique en tamazight est donc de compléter le corpus textuel traditionnel constitué de la plus grande variété possible de discours (contes, poésie, proverbes, devinettes, textes en prose, discours informels…) par un corpus «thématique» constitué par des enquêtes ciblées concernant le maximum de champs lexicaux (lexique des animaux, des plantes, des maladies, des titres et fonctions…) dont une bonne partie des termes, des nominaux notamment, a très peu de chance de se retrouver dans le premier corpus. Ce corpus «complémentaire» est constitué essentiellement par une approche onomasiologique, c'est-à-dire en partant des notions ou/et référents concernant un champ lexical particulier pour aller à la recherche de leurs dénominations. Une démarche inverse de celle concernant le corpus textuel où nous avons des dénominations ou signifiants dont il convient de chercher les signifiés. C'est l'approche sémasiologique dominante en lexicographie. Ces deux corpus, textuel et thématique ou sémasio et onomasio, sont évidemment ici les deux parties (ou sous-corpus) d'un même corpus lexicographique. L'absence de grands corpus textuels en tamazight pouvant, comme c'est le cas de ceux du *Collins COBUILD Dictionary* et du *TLF*, contenir l'essentiel du lexique de cette langue

exige donc, en plus de la sémasiologique, une autre démarche pour aller chercher des mots, les «disponibles» en particulier, qui ne risquent pas de se faire prendre dans les filets à très grosses mailles d'un modeste corpus textuel.

Bibliographie

Bacelar do Nascimento M.-F., 2000, «Corpus de référence du portugais contemporain», dans Bilger M. (éd.), *Corpus : Méthodologie et applications linguistiques*, Paris, Honoré Champion, p. 25-29.

Basset A., 1952, *La langue berbère*, Col. *Handbook of african languages*, Oxford University Press For International African Institute.

Béjoint H., 2007, «Informatique et lexicographie de corpus : les nouveaux dictionnaires», Vol. XII-1 : *Corpus : état des lieux et perspectives*, Editions De Werelt, Amsterdam, p. 7-24.

Berkaï A., 2010,»Lexicographie amazighe : inventaire et propositions», dans Dourari, A. (dir.),*La dictionnairique des langues de moindre diffusion : le cas de tamazight*, Edition du Centre National Pédagogique et Linguistique pour l'Enseignement de Tamazight (CNPLET), p. 118-131.

Berkaï A., 2011, «Quel programme microstructurel en lexicographie berbère ?», dans Naït-Zerrad K. (éd.), *La standardisation du berbère à la lumière des évolutions récentes en Europe et dans le Nord de l'Afrique*, Actes du colloque organisé à l'INALCO (Paris) 6-7 octobre 2008, *Revue des Etudes Berbère*, Vol. 5, p. 25-45.

Blache Ph., 2000, «A quoi sert l'annotation syntaxique de corpus ?», dans Bilger M. (éd.), *Corpus : Méthodologie et applications linguistiques*, Paris, Honoré Champion, p. 82-94.

Blanche-Benveniste C., 2000, «Type de corpus», dans Bilger M. (éd.), *Corpus : Méthodologie et applications linguistiques*, Paris, Honoré Champion, p. 11-15.

Corbeil J.-C., 1971, «Aspects du problème néologique», dans *La Banque des mots* n°2.

Cori M. et David S., 2008, «Les corpus fondent-ils une nouvelle linguistique ?», *Langages*, n° 171, p. 111-129.

Dallet J.-M., 1982, *Dictionnaire kabyle-français. Parler des Ait-Manguellat (Algérie)*, Paris, SELAF.

De Foucauld Ch., 1951, *Dictionnaire touareg-français*, T. I, II, III, IV, Paris, Imprimerie nationale de France.

Duchet J.-L. et *al.*, 2008, «Corpus massifs et corpus alignés : leur impact sur la recherche linguistique», dans *Bulletin de la Société de Linguistique de Paris*, T. CIII-2008, Fasc. 1, p. 129-150.

Frey C., 1997, «Corpus et information», dans C. Frey et D. Latin, *Le corpus lexicographique : Méthodes de constitution et de gestion*, Actes des troisièmes journées scientifiques du réseau thématique de recherche «Etude du français en francophonie», Paris, Editions Duculot.

Gueyken A., 2008, «Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus», *inConstruction de faits en linguistique : la place des corpus*, Langages 171, Larousse, p. 77-94.

Lafage S., 1997, «De quelques principes apparemment contradictoires dans la constitution d'un corpus lexicographique différentiel», dans C. Frey et D. Latin, *Le corpus lexicographique : Méthodes de constitution et de gestion*, Actes des troisièmes journées scientifiques du réseau thématique de recherche «Etude du français en francophonie», Paris, Editions Duculot, p. 87-100.

Mayaffre D., 2002, «Les corpus réflexifs : entre architextualité et hypertextualité», *Corpus* [en ligne] n° 1, URL : http//corpus.revues.org/11.

Mayaffre D., 2005, «Rôle et place des corpus en linguistique : réflexions introductives», dans *Texto* [en ligne], vol. X, n°4. Disponible sur: http://www.revue-texto.net/Reperes/Themes/Mayaffre_Corpus.html.

Meschonnic H., 2008, «Le dictionnaire mon trésor», préface à Dotoli G., *La construction du sens dans le dictionnaire,* Linguistica 33, Schena Editore, Hermann Editeurs, p. 11-19.

Rastier F., juin 2004, «Enjeux épistémologiques de la linguistique de corpus», dans *Texto !* [en ligne], Rubrique Dits et inédits. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.

Rey A., 2008, *De l'artisanat des dictionnaires à une science du mot. Images et modèles*, Paris, Armand Colin.

Saldanha G., 2009, «Principles of corpus linguistics and their application to translation studies research», dans *revista tradumàtica* [en ligne], n° 7. Disponible sur: http://www.fti.uab.cat/tradumatica/revista : ISSN : 1578-7559.

Taïfi M., 1991, *Dictionnaire tamazight-français (parlers du Maroc central)*, Paris, L'Harmattan-Awal.

Taïfi M. et Pognan P., 2011, «Un dictionnaire en tant que corpus : traitements informatiques du dictionnaire raisonné berbère-français de Miloud Taïfi», Actes du 4ème atelier international sur l'amazighe et les TICs des 24 et 25 février 2011, *Les ressources langagières : construction et exploitation*, IRCAM, p. 33-51 (en ligne).

1-Philipe Blache écrit à ce propos que la dénomination de «linguistique de corpus» «présuppose tout d'abord qu'il puisse exister une linguistique qui n'utiliserait pas de corpus. Cela est évidemment faux et tout linguiste, y compris ceux dont les travaux sont plus formels, s'appuient sur des corpus» (2000 : 83).

2-«Corpus linguistics is not a linguistic theory but a methodology that can be applied to a wide range of linguistic enquiries» (Saldanha, 2009 : 2).

3-Voir, par exemple, l'opposition faite par Damon Mayaffre (2005) entre *corpus lexicographiques ou sacs de mots*, *corpus phrastiques* et *corpus textuels*.

4-Il n'est pas toujours possible d'accéder à l'intégralité d'un texte lorsqu'il est protégé juridiquement par les droits d'auteur/éditeur. On recourt, en l'occurrence, à des extraits ou échantillons «représentatifs» de ces discours. Le résultat «ne serait pas un *corpus* textuel (en ce sens qu'il n'aurait pas forcément des textes complets) mais qu'il serait un *corpus* de référence (en ce sens qu'il aurait des échantillons de chaque œuvre ou document représenté)» (Bacelar do Nascimento, 2000 : 26).
5-Damon Mayaffre définit la «réflexivité» du corpus au sens «idéal» où «ses constituants (articles de presse, discours politiques, pièces de théâtre ; de manière plus générale sous-parties) renvoient les uns aux autres pour former un *réseau sémantique* performant dans un tout (le corpus) cohérent et auto-suffisant» (2001 : 5). Nous ne croyons pas, en ce qui nous concerne, à l'*autosuffisance* d'un corpus quelle que soit sa dimension. Il dépendrait toujours d'un "hors-corpus", sinon pour sa complétion, du moins pour son interprétation (v. note 7). Les caractères «réflexif» et «intégral» (v. note 4) ne sont pas indispensables, de notre point de vue, à la définition du corpus.
6- Il doit y avoir par ailleurs une juste représentativité des différents types de discours qui évite une sur- ou une sous-représentation d'un type particulier ou carrément l'absence d'un tel autre type. D'où l'usage de la notion de «corpus équilibré» (*balanced corpus*).
7- Paradoxalement, ces lemmes qui représentent des formes attestées peuvent parfois être des formes non usitées et qui sont de pures constructions de lexicographes. La lemmatisation de la forme verbale de l'impératif de la deuxième personne du singulier en tamazight tient exclusivement à sa simplicité. Du point de vue de l'usage, c'est sans doute la forme la moins conseillée comme lemme. A. Basset qui a beaucoup travaillé sur le verbe amazigh le souligne très bien en affirmant qu' «on est en effet amené parfois à dégager artificiellement cette deuxième personne qu'il n'est pas toujours aisé d'obtenir au cours de l'enquête» (1952 : 19). Nous aurions nous-mêmes, dans un essai d'élaboration d'un dictionnaire dans le cadre de notre thèse, pu dégager cette forme lemmatique pour un verbe qui n'est pas attesté à l'impératif (**nuḥ*»être, exister»). Nous avons préféré, en l'occurrence, donner la forme du verbe à la 3ème personne du sing. au prétérit (*inuḥ, tnuḥ*) qui est, elle, bien attestée. Et c'est précisément la forme lemmatique adoptée en lexicographie arabe (*faʿala*). Alain Rey concernant le français "regrette" «que cette forme dans les dictionnaires français, anglais, etc., soit l'infinitif. Il serait plus simple de la remplacer par la première personne de l'indicatif (…) L'adresse à la première personne du présent de l'indicatif amorce un paradigme mémorisé (…), et donc une possibilité de phrase, de discours, l'infinitif donne au verbe un caractère quasi nominal et métalinguistique» (Rey, 2008 : 26). Les formes non attestées dans l'usage lorsqu'elles sont nécessaires en lexicographie ou en grammaire montrent l'importance de l'introspection dans tout travail de constitution et d'exploitation de corpus. Celui-ci fait toujours appel à un savoir situé hors corpus.

8-Le Dallet repose en partie sur les nombreux *Fichiers de Documentation Berbère* que l'auteur a lui-même dirigés depuis leur création en 1946 jusqu'à sa mort en 1972. Ses travaux antérieurs comme *Le verbe kabyle* (1953) ainsi que certains des travaux de ses prédécesseurs comme la *Méthode de langue kabyle* (1913) de Si Saïd Boulifa et son glossaire ont aussi servi à l'élaboration de ce dictionnaire. Le Foucauld aussi s'est appuyé en partie sur des travaux antérieurs sur le touareg (parler de l'Ahaggar), dont ceux de l'auteur lui-même : *Textes touaregs en prose* et *Poésies touarègues.*

9- Dans le cadre d'un projet récent de «*Dictionnaire raisonné berbère-français*», Miloud Taïfi, avec la collaboration de Patrice Pognan, utilise le contenu de son ancien dictionnaire comme élément important d'un corpus (représentant 40% de celui-ci) servant de support à l'élaboration du «dictionnaire raisonné» (Taïfi et Pognan, 2011 : 34)

10- Le dictionnaire du français le plus populaire, *Le Petit Larousse*, ne propose la transcription phonétique que dans les rares cas où les mots en question présentent une sérieuse difficulté de prononciation qui résulte d'une différence très marquée entre l'oral et l'écrit. Généralement des mots latins ou des emprunts : *sine qua non* [sinekwanɔn], *hic et nunc* [ikɛtnɔ̃k]… Ce dictionnaire transcrit cependant entièrement les mots en question, alors que nous proposons de transcrire seulement les sons objet de la difficulté qui peut parfois se poser pour tous les sons d'un mot. Un historien français raconte dans une émission de télévision que Krim Belkacem, lors des négociations d'Evian où il était le chef de la délégation algérienne, avait prononcé [sindi] le mot *sine die* qui se prononce correctement comme [sinedje]. S'il l'avait appris dans un dictionnaire où il était accompagné de sa transcription phonétique, ce politicien, bon francophone par ailleurs, n'aurait pas commis cette erreur.

11- Mieux adaptées à l'amazighe que les trois parties de la grammaire arabe : le nom, le verbe et la particule. Tripartition que M. Mammeri a reprise dans sa définition de *mot* : *isem* (nom), *amyag* (verbe) *d tzelɣa* (et la particule) (v. *Tajerrumt n tmazight, grammaire berbère (kabyle)*).