sciendo
DOI 10.2478/jazcas-2022-0014

BASE DE DONNÉES NUMÉRIQUE DES CORPUS KABYLES ET EXPLOITATION. ESSAI D'ANALYSE LEXICOMÉTRIQUE DE LA DIMENSION IDENTITAIRE DANS LE DISCOURS ROMANESQUE

AREZKI IKHERBANE[1] – RAMDANE BOUKHERROUF[2] – NOURA TIGZIRI[3]
[123]Laboratoire d'Aménagement et d'Enseignement de la Langue Amazighe
[123] Université Mouloud Mammeri de Tizi-Ouzu, Tizi-Ouzu, Algérie

IKHERBANE, Arezki – BOUKHERROUF, Ramdane – TIGZIRI, Noura: Kabyle corpus digital database and exploitation. Test of lexicometric analysis of the identity dimension in the romanesque discourse. Jazykovedný časopis (Journal of Linguistics), 2021, Vol. 72, No 4, pp. 894 – 905.

Abstract: The purpose of this contribution is to show, through a preliminary analysis of a corpus sample composed of the first five kabyle novels (1963-1990), the contribution of lexicometry as a new method based on statistics, in the treatment of large corpora and the establishment of databases. The aim is to describe all the phases intrinsic to the preliminary processing of a corpus (transcription, tagging and lemmatization) before submitting them to the various stages of its exploitation. Thus, in our corpus, we have opted to deal with the theme of identity induced by the five works by highlighting both the overused vocabulary and the singularity of each work in relation to the corpus as a whole. But before moving on to the quantitative analysis of the vocabulary, a work of data preparation is necessary. We intend to focus on the orthographic choices to be adopted by removing all ambiguities, the marking out and the lemmatization of the corpus. In order to do this, we have resorted to Lexico5 computer tool.

Key words: corpus, kabyle, identity, novel, lexicometry, databases

0. DONNÉES INTRODUCTIVES

La langue amazighe (berbère), était une langue minorée, réduite à l'usage quotidien, parlée uniquement dans les régions amazighophones. Elle fait partie des langues chamito-sémitiques. Elle couvre une vaste aire géographique : de l'Afrique du Nord comprenant le Maroc jusqu'en Egypte en passant par l'Algérie, la Tunisie et la Lybie sans oublier le Sahara et une partie du Sahel Ouest-africain avec de nombreux locuteurs au Mali et au Niger (Chaker, 1983).

La réalité linguistique du domaine berbère (amazigh) montre que la langue connaît une fragmentation et une dialectalisation parfois importante d'un point à un autre. À l'intérieur du dialecte kabyle, des divergences se manifestent particulièrement en phonétique et dans le lexique. L'intercompréhension est parfois difficile avec les autres variantes de l'amazigh.

Dans le présent article, il est question de présenter quelques hypothèses préliminaires en phase expérimentale, d'un grand projet qui s'inscrit dans le cadre d'un

axe de recherche[1] portant sur les technologies linguistiques. Il s'agit d'un projet qui a pour objectif de mettre en place une grande base de données numérique des corpus amazighs[2] (oraux et écrits), accessible en ligne, dans la perspective de les exploiter dans le cadre du traitement automatique de l'amazighe (dictionnaires, atlas linguistiques, analyse du discours, grammaires, syntaxe, traduction automatique, enseignement, etc.) afin de les mettre à la disposition des chercheurs berbérisants qui veulent les exploiter.

Par ailleurs, loin de faire une analyse quantitative d'un grand corpus, l'objectif central de notre première exploitation consiste en une tentative d'application lexicométrique[3] sur un échantillon de corpus en mettant en exergue les traitements préalables de préparation adoptés : la transcription, le balisage et la lemmatisation.

Ainsi, nous nous appuyons sur un corpus des cinq premiers romans kabyles, formé d'une panoplie d'œuvres qui s'étale sur une période de trois décennies : Belaid Ait Ali (1963), Rachid Aliche (1981, 1986), Said Sadi (1983) et Amar Mezdad (1990).

L'intérêt de ce corpus réside dans le fait qu'il a suscité l'intérêt non seulement des études littéraires, mais aussi des études de statistique textuelle. C'est à ce titre que les travaux de la critique littéraire berbérisante montrent que la thématique identitaire est portée dans l'émergence de l'écriture romanesque berbère (Salhi – Sadi, 2016) puisque ce nouveau genre est né dans un contexte particulier, caractérisé par un mouvement de revendication pour la reconnaissance du berbère en tant que langue et identité. Ainsi l'étude statistique (Loikkanen, 1998) portant sur le vocabulaire du roman kabyle, montre que le thème le plus récurent est celui de la revendication culturelle et identitaire. C'est pourquoi nous proposons d'inscrire notre recherche dans ce sillage, mais avec une démarche qui diffère quelque peu, en vue d'aborder cet aspect d'identité tant de point de vue quantitatif que qualitatif.

Par ailleurs, en plus du statut relevant du sème **identité** caractérisant les conditions de production et de l'émergence du discours romanesque kabyle, notre choix en faveur de la lexicométrie, comme domaine issu de la statistique lexicale et des derniers développements en traitement automatique des langues naturelles (TALN), est motivé par le fait que l'application de ces méthodes sur un corpus littéraire, aussi vaste soit-il, permet d'obtenir des résultats d'une manière rapide et quasi systématique.

---

[1] Le Projet mené au niveau du laboratoire d'Aménagement et Enseignement de la Langue Amazighe de Université Mouloud Mammeri de Tizi-Ouzou, Algérie.

[2] Le projet s'inscrit dans le cadre des projets du comité d'évaluation de la recherche universitaire de l'enseignement supérieur et de la recherche scientifique dont il a fait l'objet. Il comprend trois axes : Base de données kabyle (2010), La transcription synchronisée des corpus oraux (2015) et les écrits anciens en tamazight : recueil, numérisation, récriture et géolocalisation (2018). Ainsi, ces derniers, ont fait l'objet de plusieurs communications et publications : Tigziri (2014), Boukherrouf et Tigziri (2015, 2016) et Tigziri, Jolivet et Boukherrouf (2017).

[3] Deux analyses lexicométriques ont été réalisées à l'aide du logiciel AntConc des corpus kabyles Tigziri (2016, 2017).

Le projet auquel nous nous attelons, compte tenu du corpus cité plus haut, sera axé sur la problématique du discours identitaire et la manière dont il est organisé à travers les œuvres précédemment mentionnées. Il s'agit d'examiner la dimension identitaire en tant qu'entité concrète, inhérente à l'ensemble des œuvres du corpus choisi, afin de dégager ses modalités de fonctionnement ainsi que sa structure interne, à partir de la qualification du mot et le sens du texte. Pour ce faire, nous ferons appel aux travaux de Labbé et Labbé (2013 ; 2019) qui ont bien décrit l'objet et la méthode de la lexicométrie:

> Elle permet de traiter de vastes ensembles de textes (corpus), d'établir leur vocabulaire, de classer les vocables en fonction de leur fréquence, de leur répartition, de leurs catégories grammaticales. Elle établit les contextes d'emploi d'un vocable et les combinaisons les plus fréquentes dans lesquelles il entre, ce qui permet de déterminer le ou les sens de ce vocable. Elle retrouve les principaux thèmes présents dans un corpus, son genre et son style. Elle segmente ce corpus en fonction des ruptures thématiques ou stylistiques. Pour obtenir ces résultats, des traitements préalables sont nécessaires : balisage des textes, correction et standardisation orthographiques, étiquetage des mots. Le texte peut alors entrer dans une bibliothèque électronique à la disposition des chercheurs. (2013, p.1).

Ceci étant, nous allons associer à l'étude du contexte qui tient compte des mots et de leurs dispositions, une analyse sémantique, qui cherche à identifier à l'intérieur de ces mots un sens au-delà des termes stricto sensu (Ada, 2004). Pour ce faire, nous tenterons d'abord de faire le relevé de tous les éléments de notre corpus (Maingueneau, 1991) avant de les exploiter pour caractériser l'œuvre dans son ensemble ou pour traiter les données lexicales par rapport à l'ensemble du vocabulaire de l'œuvre (Muller, 1982).

Le présent essai est composé de deux parties. Après une phase préparatoire et l'établissement des données, qui présentent le corpus, sa transcription, son balisage, sa lemmatisation et les différentes fonctionnalités prises en charge par notre outil d'analyse Lexico 5, nous passerons à la description progressivement les différentes phases de l'exploitation des données préparées.

1. PRÉPARATION ET ÉTABLISSEMENT DES DONNÉES

1.1 Présentation des données

Le corpus que nous soumettons à l'étude est constitué du discours romanesque produit en kabyle. Ces œuvres s'étendent sur une période allant de (1963 à 1990). Le premier roman intitulé *Lwali n udrar « Le saint de la montagne »* a été publié en 1963 par Dallet dans le Fichier de Documentation Berbère (FDB). L'ensemble de ces textes qui constituent notre corpus représentent, comme déjà cité, cinq romans, à savoir : *Lwali n udrar « Le saint de la montagne »* de Belaid Ait Ali, *Asfel « le sacrifice »* et

Faffa « diminutif de la France» de Rachid Aliche, *Askuti « Le scout»* de Said Sadi et *Iḍ d wass « nuit et jour »* de Amar Mezdad. Mis à part Rachid Aliche qui compte deux œuvres, ces textes qui représentent une œuvre de chaque auteur, ont été classés selon l'ordre chronologique des dates de publication pour pouvoir étudier l'évolution du vocabulaire, abstraction faite de la date de rédaction car dans le contenu des ouvrages, il y a presque aucune symétrie entre la date de publication et celle de sa rédaction ; l'écart est souvent conséquent. C'est le cas de « *Lwali n udrar* » de Belaid Ait Ali, écrit entre 1940 et 1945, mais qui n'est publié qu'en 1963 à titre posthume ; idem pour *« iḍ d wass »* de Amar Mezdad, écrit à partir des années 80 et non publié qu'en 1990.

1.2 L'outil d'analyse : Aperçu sur ses fonctionnalités

L'analyse des données est menée à l'aide du logiciel **Lexico 5**[4], développé depuis le milieu des années 2010. Il est le prolongement de Lexico 3 développé à partir de 2003. Lexico 3 est lui-même une suite logique de ses deux versions précédentes (Lexico 1 puis Lexico 2). L'une des nouveautés de la version Lexico 5 par rapport à la précédente est la prise en charge de l'encodage des corpus en Unicode, offrant une possibilité à un plus grand nombre de langues, entre autre le berbère, de bénéficier de ce traitement lexicométrique. Les fonctions documentaires (concordances, contextes), statistiques (spécificités), analyses multidimensionnelles (analyses factorielles des correspondances, arborées), sont l'essentiel de ce que renferme cet outil en termes de fonctionnalités.

Le dictionnaire est une donnée que le logiciel génère automatiquement sous forme de deux colonnes. Cette liste des formes présente dans le texte s'affiche toujours dans la colonne de gauche du logiciel, sous l'onglet dictionnaire, à côté des onglets navigation et rapport. Dans le dictionnaire de gauche (cf. Figure 1) les occurrences sont affichées dans un ordre lexicométrique (classement décroissant). Si on clique sur l'onglet Formes (ordre lexicographique), le classement des formes du texte s'affiche alors par ordre lexicographique (ordre alphabétique).

1.3 Etablissement et standardisation orthographique des textes

En dépit des disproportions constatées entre les romans proposés à l'étude, en termes de volume, qui diffèrent tant par le nombre de pages que de mots, nous les avons établis sous forme d'un seul et unique corpus. Ainsi, de tout le texte narratif des romans, nous avons supprimé uniquement les éléments qui renvoient au péritexte, autrement dit, les numéros et rubriques éventuels des chapitres, les dédicaces, les numéros des pages, les titres, les notes en annexe qui expliquent les mots, les lexiques, les chiffres, les dates, les présentations biographiques, les préfaces ainsi que les passages écrits en entier dans la langue étrangère.

[4] Téléchargeable sur http://www.lexi-co.com/Produits.html

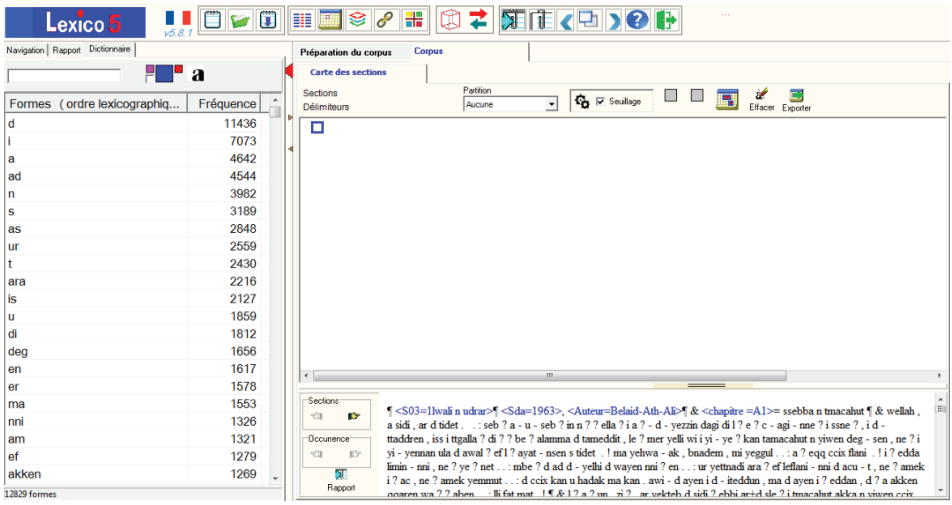


Fig. 1. Index alphabétique des données textuelles

Aussi, à cause de la notation que chaque auteur a choisi d'utiliser dans son texte, le découpage du texte en mots à base d'unités graphiques parait délicat. On a souvent écrit des éléments différents ensembles, comme c'est le cas pour *les affixes personnels* et *les prépositions.* Pour les y homogénéiser, nous avons dû procéder à la séparation entre les mots outils à fonction grammatique des autres lexèmes auxquels ils se rattachent. Mais malgré l'apparente ambiguïté de certaines classifications, nous avons dû s'en tenir, outre les choix adoptés par (Loikkanen, 1998), aux différentes propositions des manuels du kabyle pour déterminer certaines unités mal identifiées.

Par ailleurs, le mot est avant tout une unité graphique, séparé des autres par des blancs ou par des signes de ponctuation. Mais d'autres cas, ils sont constitués de deux ou plus de morphèmes composés qui rendent ambigu la définition que nous pourrions donner du mot *mulac « sinon »*, *bu qerru « grosse tête »* et *adrar ufud « le tibia »*. Ainsi, une unité graphique peut correspondre à plusieurs unités du lexique ou inversement une unité du lexique peut se composer de plusieurs unités graphiques. Pour effectuer le décompte des occurrences afin de regrouper les vocables, la méthode lexicométrique recommande d'élaborer les critères à base desquels on traite le texte et on reconstitue le lexique. Nous avons constaté, qu'hormis quelques compositions synaptiques, dont le processus de lexicalisation n'est pas encore achevé, il est rare de trouver beaucoup de mots composés écrits ensemble Dallet (1982). Ainsi, nous écrivons *ma ulac « si non », yal wa « chacun »*, au lieu de les écrire en un seul mot dont les parties sont groupées ou liées par un tiret. S'agissant du trait d'union justement qui, en kabyle, peut parfois si-

gnifier une seule unité de sens et d'autre fois deux voir plus, nous avons opté de le considérer comme un caractère délimiteur. Par ailleurs, dans le cas où l'unité est considérée comme une seule forme, nous avons introduit à la place du tiret le caractère « + » non délimiteur pour voir l'expression comptée comme une seule occurrence. Concernant le cas des particules de possession *bu « celui », u « de »*, nous les avons tantôt considérées séparément et tantôt rattachées. La particule est ainsi rattachée pour le cas des noms propres, du lexique des maladies et des plantes mais souvent détachée pour le cas des numéraux ou encore de ceux dont la formation n'est pas totalement grammaticalisée.

1.4 Balisage du texte

Avant de commencer à travailler statistiquement sur le texte, il est nécessaire de définir les normes lui permettant d'être segmenté en formes (unités textuelles) et afin qu'il soit reconnaissable par le logiciel, et ce conformément aux caractéristiques textuelles de chaque genre du discours. Ainsi, pour que la machine puisse traduire le texte dans un code qu'elle comprend, il suffit d'enregistrer le texte en format brute (.txt). Pour ce qui est de la séparation des mots, le caractère est soit délimiteur, c'est le cas de l'espace et plus généralement des signes de ponctuations, soit il ne l'est pas. En conséquence, une suite de caractères non délimiteurs contenue entre deux caractères délimiteurs est considéré comme une occurrence analysable. En raison de la spécificité de la langue amazighe, nous avons considéré le trait d'union comme délimiteur, car dans la plupart des cas celui-ci relie différents affixes aux verbes, noms ou adjectifs, etc.

Par ailleurs, pour considérer une unité composite comme une seule occurrence, il suffit de les relier à l'aide du signe « + » non délimiteur. De même, pour distinguer les unités homonymiques entre elles, notamment un nom propre et un adjectif, nous avons précédé l'ensemble des noms propres du corpus d'un astérix, et ce dernier, est considéré comme un marqueur des noms propres de l'ensemble du corpus.

L'autre étape importante réside dans le codage du texte à l'aide de certaines clés informatives codées, qui caractérisent les métadonnées et les données du corpus. Ainsi, la clé <Sda=1963> désigne l'année de publication, la clé <Auteur=Belaid Ait Ali> la partition par auteur, <S03=1> les titres d'un texte qui réunissent en une seule base textuelle différents sous-corpus délimitant (une division logique, thématique ou chronologique), nous pouvons comparer plusieurs parties du corpus et également naviguer dans le corpus plus facilement. Concernant notre corpus, il était nécessaire de répartir les textes selon les titres des romans pour pouvoir les comparer entre eux (<S03=1lwali n udrar>, <S03=2 asfel>, <S03=3 askuti>, <S03=4 faffa>, <S03=5 iḍ d wass>). C'est la même chose pour obtenir une séparation par date d'édition <Sda=1963>, par auteur <Auteur=Belaid Ait Ali>, ou bien encore par chapitre <chapitre =A1>=*ssebba n tmacahut « l'argument de l'histoire »*>. Aussi, pour qu'une

seule forme ne soit pas traitée deux fois, les majuscules devront être supprimées. Ainsi, la balise <S03=titre d'un texte> par exemple, nous permet de générer les données de chaque corpus (cf. Tableau 1).

Partie	Occurrences	Formes	Hapax	Fmax	Forme
1lwali n udrar	33898	4258	2251	2606	d
2 asfel	14789	3739	2281	1195	d
3 askuti	32626	5095	2882	2355	d
4 faffa	23614	5692	3472	1426	d
5 iḍ dwass	47957	6507	3323	3460	d
T corpus	**152884**	**15071**	**7444**	**11042**	**d**

Tab. 1. Partition par la balise du titre de l'œuvre

1.5 La lemmatisation des données du corpus

Après dépouillement et comptage des différents mots, pour des besoins d'analyse, nous avons regroupé un certain nombre de mots en un seul lemme. A l'aide de ce procédé, nous avons réuni en une seule classe quelques lemmes en suremploi. Ainsi, la lemmatisation est caractérisée par le regroupement en une seule forme canonique de toutes les flexions appartenant à une seule classe et catégorie grammaticale. Ainsi, nous avons représenté sous un seul lemme toutes les déclinaisons avec lesquelles nous avons pu écrire les mots dans le corpus étudié. Le lemme *tawwurt « la porte »,* nom féminin singulier, regroupe l'ensemble de ses variantes et leur flexion comme, *tiwwura, tewwura, tabburt, tebburt, tibbura, tebbura*. Le lemme préposition *« avec »* englobe les formes *s, es, yis, yiss*, et le lemme *ad*, particule de l'aoriste, englobe les formes variantes, comme *a, an, at.* Pour la forme verbale, comme langue aspectuelle, nous avons réduit l'ensemble des flexions à la forme de l'aoriste simple.

2. EXPLOITATION DES DONNÉES ET INTERPRÉTATION

Après avoir présenté les étapes principales inhérentes à la préparation et l'établissement du corpus, nous tenterons de présenter un essai de son exploitation en mettant en exergue l'index alphabétique de quelques données lemmatisées et la présentation des résultats des données des cinq œuvres par partition AFC.

2.1 Index alphabétique des données

La première information sur notre collection de textes apparaît sous forme d'une liste (d'index lexicographique) de toutes les unités du lexique ou vocables (entrées du dictionnaire) avec leurs effectifs (c'est-à-dire le nombre de leurs occurrences). Sous cette liste, qui donne, pour chaque mot, les formes graphiques sous lesquelles il s'actualise, nous avons résumé les plus importantes (cf. Tableau 2).

ini («dire, aoriste simple »)	1418
yenna	320
nniɣ	123
yini	112
yeqqar	105
tiniḍ	95
tenna	89
iniɣ	70
nnan	62
tenniḍ	34
yinin	29
inin	28
teqqar	27
innan	25
yeqqaren	23
qqareɣ	21
teqqareḍ	18
nini	15
nenna	13
neqqar	09
qqarent	08
yenni	07
teqqarem	07
ini	04
init	03
tinim	02
nnin	01
tinem	01
(…)	
ili («être», aoriste simple)	999
yella	278
yelli	96
tella	122
llan	120
yili	88
tili	78
lliɣ	33
nella	27
llant	26
telliḍ	21
llin	04
tiliḍ	09
ttilin	06
tettili	05
llin	04
ttilint	03
uxxam	156

(…)	
tamurt « pays »	408
tmurt	127
tmura	15
timura	04
(…)	
taddart « village »	282
tddart	249
tudrin	25
tuddar	08
leqbayel	34
aqbayli « kabyle »	79
leqbayel	34
teqbaylit	22
taqbaylit	15
aqbayli	06
uqbayli	02
(…)	
ṛebbi « dieu »	376
(…)	
Ccix « prêcheur »	259
(…)	
Netta/nettat « il/elle	483/236
(…)	
iw/ik « ma /ta »	393/336
(…)	
nekk/keč « moi/toi »	341/183
(…)	
Nutni/nekkni « ils/nous »	107/82
(…)	
ameẓyan « petit »	188
(…)	
muḥend « nom d'un personnage	208
(…)	
malḥa « nom d'un personnage »	136
(…)	
axxam « maison »	360
uxxam	164
axxam	156
ixxamen	29
yixxamen	11
(…)	
ass « jour »	398
wass	224
ussan	104
wussan	70
(…)	

nelli	17	ddunit	257
yettili	17	(…)	
ylin	17	awal	230
yettilin	03	awalen	8
ilit	03		
iliɣ	05		
nili	04		
tellamt	04		
yilin	04		
yettilin	03		
ilint	01		
tilimt	01		

Tab. 2. Extraits de l'index alphabétique : lemmatisation semi - automatique

Les premiers tests montrent que le discours est caractérisé par des unités graphiques sur-employées telles que *ini « dire » et ili « être »*, utilisées successivement comme verbe et/ou auxiliaire. Le prétérit de la deuxième personne du singulier (*Yenna*, 320 fois et *yella* 278) semble être la classe grammaticale la plus dominante dans le discours. Le récit sur des événements passés, discours rapporté serait la raison principale.

Si les substantifs *ṛebbi « dieu »* 376 et *ccix « prêcheur ou vieux »* 259 occurrences, peuvent être des marqueurs discursifs, à cause d'un supposé lien entretenu avec l'existence et la religion, ce n'est peut-être pas le cas de ces mêmes lexies selon qu'elles soient mises ou non dans un rapport de contextualisation au même titre que le substantif *ddunit « vie »* cité 257 fois.

La déictisation par emploi de pronoms personnels indépendants ou affixes (*netta/nettat (lui/elle)* 483/236 fois, *iw/ik (mien/tien)* 393/336 fois, *nekk/kečč (moi/toi)* 341/183 fois, *nutni/nekkni (eux/nous)* 107/82 fois, réunis par paires oppositives, rendent compte d'une mise en contexte du discours sur soi et l'autre. Ce rapport à l'altérité qui est par conséquent un des indicateurs majeurs du discours sur l'identité, constitue un réactif essentiel grâce auquel il est possible de clarifier l'ethos de celui qui parle et de celui pour qui le message est supposé prédestiné.

Le recours à l'adjectif *amezyan « petit »,* cité 188 fois aux côtés du nom propre *Muḥend* 208 fois, dénoterait de cette volonté de produire un discours sur la jeunesse. De la même manière, *Malḥa* qui a été citée 136 fois, est un autre nom propre qui, en plus de la vivacité qui le caractérise, rend bien compte du sentiment qu'un auteur puisse éprouver à l'égard d'une femme.

L'emploi des termes comme *tamurt « pays », axxam « maison » et taddart « village »* respectivement répétés 408, 360 et 282 fois, en tant que des espaces - concrets ou abstraits - ne laisse aucun doute quant à la place que ces auteurs leur ont accordés grâce à des choix discursifs qui rendent ces espaces plus attrayants. Ainsi, l'espace local serait l'une des caractéristiques principales dans le roman identitaire.

Contrairement au suremploi d'unités comme *ass « jours »* 398 fois, *awal « parole »* 230 fois, dont on ignore le sens exact en dehors du contexte de leur mise en discours, il y a lieu de souligner le sous emploi des formes adjectivales comme *aqbayli « le kabyle »* 79 fois et *amaziɣ « l'amazigh/berbère »* 30 fois seulement.

2.2 Essai d'analyse factorielle des correspondances des cinq œuvres

Pour rendre compte approximativement des grandes oppositions sous-jacentes dans le corpus de textes, deux constats sont mis en exergue, d'une part, l'œuvre de Belaid Ait Ali apparaît excentrée toute à droite avec comme caractéristique des termes religieux (*ṛebbi « Dieu » et ccix « vieux-prêcheur »).* De l'autre, les quatre œuvres restantes qui se rapprochent toutes de l'axe du milieu avec des différences parfois remarquables sur le plan du lexique. Ainsi, au moment où Rachid Aliche se rapproche des termes comme *ddunit « la vie »,* rendu par le pronom affixe *iw « mien(ne) »*, Said Sadi en porte le souci de *tamurt « pays »,* de manière subjective grâce au pronom personnel *nekk « moi »*. Un peu plus en haut de l'axe vertical, le nom propre, la forme adjectivale ainsi que la forme du personnel *nutni « eux »* sont tous ce qui font de Amar Mezdad une œuvre à part (cf. Figure 2). Nous pouvons conclure que l'écart et la singularité mis en exergue dans les cinq œuvres est d'ordre chronologique. En effet, la singularité de l'œuvre de Belaid Ait Ali est un fait marquant caractérisant la vision du discours identitaire kabyle.

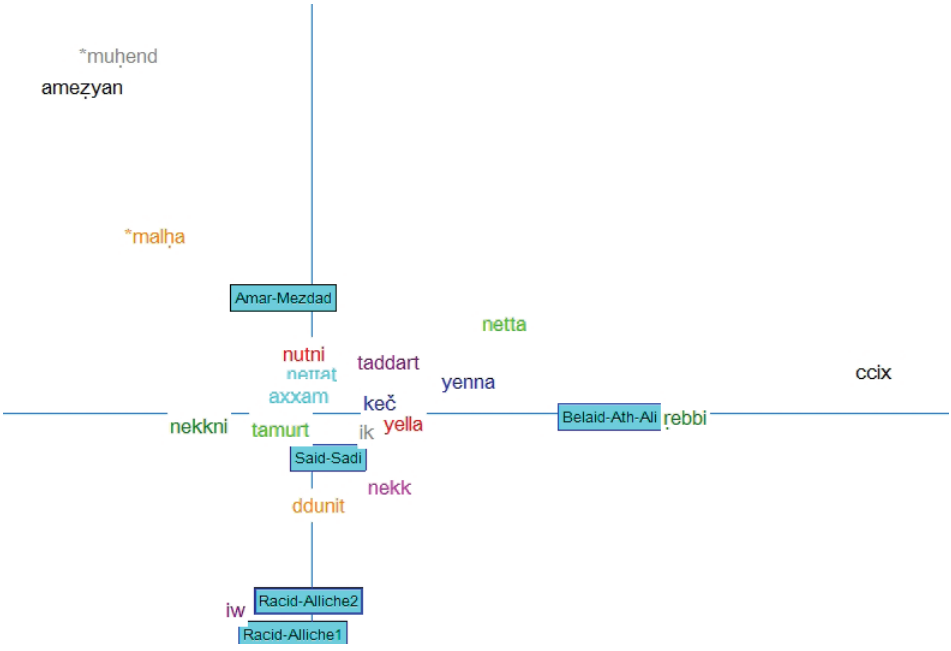


Fig. 2. Essai de présentation de partition par AFC

3. CONCLUSION ET PERSPECTIVES

Parmi les objectifs fixés dans notre projet, comme il a été signalé au début, il s'agira de mettre en place une base de données numérique des corpus amazighs, transcrits et balisés avant de les mettre à la disposition des chercheurs berbérisants pour des besoins d'analyse dans plusieurs domaines.

Ainsi, une fois toutes les données seront regroupées (orales et écrites), il conviendra de les classer en genres du discours et en domaines de traitement accompagnées de toutes les informations nécessaires à chaque corpus (les métadonnées).

Par ailleurs, avant de passer à l'étape de la transcription et de l'établissement des données, un travail de préparation des corpus en fonction des outils et des méthodologies à adopter est plus que nécessaire. Il permettra d'établir les données conformément aux perspectives théoriques et méthodologiques adoptées. Aussi, une phase d'expérimentation et de tests sur des choix de transcriptions à adopter seront d'une importance capitale à cette tâche de préparation des données à soumettre à l'exploitation.

C'est dans cette perspective, que nous avons choisi d'inscrire notre projet dans le domaine de la lexicométrie pour établir les données recueillies conformément aux choix méthodologiques de cette approche. Le choix de l'échantillon du corpus soumis aux premiers tests réside dans le fait que les cinq œuvres s'inscrivent dans un seul genre et partagent la thématique de l'identité.

Les résultats de ce test sont en phase expérimentale, d'autres étapes s'imposent afin d'établir les données avant de les généraliser pour l'ensemble des corpus.

Bibliographie

ADA, Myriam Scanu : Séminaire d'Analyse du texte fantastique : Littérature et informatique, Hyperbase logiciel pour l'analyse textuelle. Università degli Studi di Bologna 2004.

ALICHE, Rachid : Asfel. Mussidan : Fédérop 1981.

ALICHE, Rachid : Faffa. I yuɣen irgazen ur ttrun. Mussidan : Fédérop 1986.

BOUKHERROUF, Ramdane – TIGZIRI, Noura : Base de données kabyles : collectes de données et applications. Synchronisation texte / son. In : Iles d Imesli, 2015, No 7, pp. 193–206. Disponible sur : http://revue.ummto.dz/index.php/idi/issue/view/122

BOUKHERROUF, Ramdane – TIGZIRI, Noura : Etablissement des corpus et construction de l'objet pour l'analyse du discours. Transcription synchronisée à l'aide du logiciel Winpitch. TICAM'16, 28 et 29 novembre 2016, IRCAM, Rabat-Maroc.

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

DALLET, Jean-Marie : Dictionnaire kabyle-français. (Parler des Ait-Menguellat, Algérie). Paris : SELAF 1982.

DALLET, Jean-Marie – DEGEZELLE, Jules-Louis (Eds.): Les cahiers de Belaïd ou la Kabylie d'auton, T1 (Textes) et TII (traduction). Fort National, Fichier de Documentation Berbère (FDB) 1963. Disponible sur : http://dx.doi.org/10.1080/17409292.2016.1120548

LABBÉ, Cyril – LABBÉ, Dominique : Lexicométrie : quels outils pour les sciences humaines et sociales ? In : Usages de la lexicométrie en sociologie, France : Guyancourt 2013. Disponible sur : https://hal.archives-ouvertes.fr/hal-00834039

LABBÉ, Cyril – LABBÉ, Dominique : Humanités numériques. Données et méthodes. Marcel Proust. A la recherche du temps perdu. Semaine Data-SHS, 9-14 décembre, France : Université de Grenoble-Alpe 2019.

LOIKKANEN, Sinikka : Vocabulaire du roman kabyle (1981-1995). Mémoire de DEA, Paris : INALCO 1998.

MAINGUENEAU, Dominique : L'Analyse du discours. Paris : Hachette 1991.

MEZDAD, Amar : Iḍ d wass. Alger : Asalu 1990.

MULLER, Charles : Une nouvelle façon de voir le lexique : le Brunet. In : Le français moderne, 1982, No 4, Conseil International de la Langue Française, pp. 321–328.

SADI, Said : Askuti. Paris : Imedyazen 1983.

SALHI, Mohand Akli – SADI, Nabila : Le Roman Maghrébin en Berbère. In : Contemporary French and Franchophone Studies, 2016, Vol. 20, No 1, pp. 27–36.

TIGZIRI, Noura – JOLIVET, Remi – BOUKHERROUF, Ramdane : Base de données de corpus oraux : applications. The 3rd International Conference on Business Intelligence, march 29-31, 2017, Beni Mellal-Maroc. Disponible sur : https://www.cbi-bm.com/

TIGZIRI, Noura : La réalisation de grands corpus berbères normalisés et interopérables : enjeu culturel et enjeu d'ingénierie linguistique. In : Revue Asinag, 2014, Vol. 1, No 9, pp. 75–90.

TIGZIRI, Noura : Analyse textuelle à l'aide du concordancier AntConc d'une œuvre de Bélaid Ait Ali. In : Iles d Imesli, 2016, No 8, pp. 163–172.

TIGZIRI, Noura : Ait Menguellet ou l'éveilleur des consciences. In : Langue, Cultures, Communication, : L2C, 2017, Vol. 1, No 2, pp. 269–287.