

Corpus informatisé : du balisage des genres pour L'analyse des romans kabyles Computerized corpus: markup of genres for Analysis of Kabyle novels

Arezki Ikherbane¹, Amirouche Amatoui²

¹²Laboratoire d'Aménagement et d'Enseignement de la Langue Amazighe

¹²Université mouloud Mammeri de Tizi-Ouzou, Algérie

¹arezki.ikherbane@ummto.dz, ²amirouche.amatoui@gmail.com

Article information

History of the article- Historique de l'article

Received: 20/03/2022

Accepted : 29/09/2022

Published : 31/12/2022

Abstract

With the emergence of new software, lexocometric analysis can systematically, flexibly and dynamically approach large computerized corpus in a record time. However, preparing this great digital stream for good human expertise requires adequate markup. Our communication aims to illustrate the implementation of this method, which consists in explaining the segmentation norms, starting from a Kabyle novelistic corpus. *Lexico 5*, *Treecloud* software are used to conduct this study.

Keywords: Tags; genres, lexicometry; corpus; text processing software.

Résumé

Avec l'émergence de nouveaux logiciels, l'analyse lexicométrique peut approcher de manière systématique, flexible et dynamique de grands corpus informatisés en un temps record. Cependant, préparer ce grand flux numérique à une bonne expertise humaine nécessite un balisage adéquat. Notre communication se propose d'illustrer la mise en œuvre de cette méthode, qui consiste à expliciter les normes de segmentation, en partant d'un corpus romanesque kabyle. Les logiciels *Lexico 5*, *Treecloud* sont utilisés pour mener cette étude.

Mots clés : balises ; genres ; lexicométrie ; corpus ; logiciels de traitements textuels.



Introduction

Par corpus informatisé on entend toute méthode qui, tend à approcher des quantités de données mises sur ligne de manière rapide et instantanée. Mais conçu dans le cadre de passage à l'écrit électronique, ce flux échappant à la perspicacité de l'œil humain nécessite un recours au pouvoir de la machine pour permettre une rapidité dans le traitement. Et tandis que l'accès libre, la prise en charge du caractère unicode, soient deux facteurs déterminants dans le choix des logiciels que nous expliciterons tout au long de ce papier, le souci reste de situer le corpus dans le genre. La solution réside alors dans un ensemble de métadonnées (par exemple, le titre, l'auteur, la date ou toute autre caractéristique typologique du texte) permettant de mener des analyses contrastives à l'intérieur de l'unité textuelle (page) ; des structures intermédiaires (phrases, citations, segments répétés) ; et au niveau de granularité inférieur, le (mot-forme) comme unité lexicale de base.

Pour ce faire, nous avons opté le logiciel du traitement automatique **Lexico 5** (Téléchargeable gratuitement sur le lien suivant : <http://www.lexico.com/>) développé depuis le milieu des années 2010. Il est selon J.-M. Leblanc (Leblanc, 2017 : 35-40), un des outils longitudinaux qui travaillent sur la base d'un tableau lexical issu d'une segmentation de la séquence textuelle en unités minimales et graphiques. Ainsi, il introduit la notion de la partition contrastive et des mesures de ventilation du stock lexical entre les différentes-parties du corpus. Pour réunir en une seule base textuelle les différentes parties qui compose notre corpus – ((*Lwali n Udrar*-1963) de Belaid Ath-Ali, (*Asfel*-1981) et (*Faffa*-1986) de Rachid Aliche, (*Askuti*-1983) de Said Sadi et (*Id d wass*-1990) d'Amer Mezdad.) - certains codages sont absolument nécessaires.

En complément de cette étude, on propose un petit aperçu sur l'outil **Treecloud** (<http://www2.lirmm.fr/~gambette/treecloud/index.htm>), un logiciel de visualisation de texte sous la forme de nuages arborés, qui a été développé par Jean Véronis, Philippe Gambette, et Jean-Charles Bontemps (2009). Sa force d'analyse réside dans sa capacité à repérer au sein d'un texte, de façon coordonnée, simultanée et attractive la saillance des mots et l'emboîtement des sens, calculés respectivement à partir de l'occurrence et de la cooccurrence. Par ailleurs, notre objectif n'est pas de cheminer à travers ses diverses fonctionnalités, étrangement absentes de la plupart des outils actuels de textométrie (Amstutz & Gambette, 2010 : 2), mais de montrer son utilité à servir de complément d'analyse et de jonction à d'autres outils informatiques.

I. Aperçu général sur des logiciels de traitement automatiques

Selon Jenny (1997) et Pincemin (2018), on peut citer quelques logiciels, gratuits ou payants, pour effectuer une analyse du discours plus épais notamment : **Alceste** (de M. Reinert de laboratoire de psychologie de Toulouse), effectue des analyses lexicales par contexte d'un ensemble de segments textuels ; **Hyperbase** de (E. Brunet de l'INaLF.), effectue des traitements documentaires et statistiques des corpus textuels. **Sato** (système d'analyse de textes par ordinateur) développé par F. Daoust, au centre

A.T.O. de l'université de Québec à Montréal. **Trope** développé par Pierre Molette, ingénieur-système, directeur de la société ACETIC (Action Etudes Information Communication). Il s'inspire des recherches cognitive-discursives du GRP (Groupe de Recherche sur la Parole). **Iramuteq** conçu (par P. Marchand et P. Ratinaud) se veut une interface qui s'appuie sur (l'environnement statistique R et le langage **python**) pour les analyses multidimensionnelles de textes et de questionnaires.

II. Des travaux sur le balisage des genres

Pour aborder les quelques travaux portant sur la délimitation des genres, on peut citer succinctement le système riche et varié, sous TXM qui, à l'aide de 297 471 balises TEI de 64 types différents (Lavrentiev et Heiden, 2014), peut identifier des blocs de textes (<div>, <ab>, <p>, <lg>, <l>, etc.), des formes particulières (<seg>, <hi>, etc.), des corrections éditoriales ou des résolutions d'abréviation (<corr>, <reg> et <axpan> regroupé avec <sic>, <orig>, <abbr> dans <choice>), et indexer des mots (<terme>) de notre choix, etc. Leblanc (2015), présente une réflexion méthodologique comparative des approches et des méthodes en matière de traitement des textes automatisés, issue du croisement des outils logiciels, voire de leur permutation, et défend une démarche expérimentale en lexicométrie. Dans une vision plus littéraire, Labée Dominique (2018) mène, - à travers 64 entretiens de l'enquête pilotée par Etienne Schweisguth - (grâce à un balisage adéquat, à partir du tableau lexical, des index hiérarchiques, des concordances et syntagmes répétés), toute une série de tests sur la longueur des textes et principales caractéristiques de leurs vocabulaires. Une autre étude, plus ambitieuse, dont il est question d'analyser un corpus Politweets (Longhi, 2020) de 34721 tweet venant de 205 politiques influents, à l'issue de l'élection municipale 2014, recontextualisé par les moyens technologiques appelés (techno-contextualisations) grâce aux métadonnées fournies par tweeter). S'inscrivant dans le vaste projet CoMeRe composé de: bogues, tweets, SMS/textos, courriels, clavardages, forums, etc., ces corpus et leurs métadonnées on été structurés suivant des formats : TEI, CLARIN, OLAC.

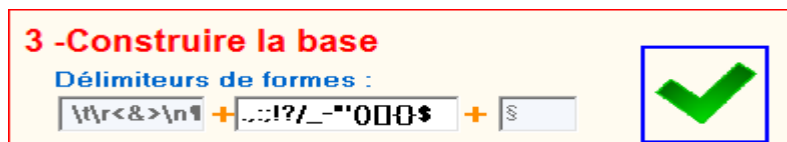
III. Methode

Notre méthode de repereage automatique et semi-automatique de differentes forme du discours - sans en constituer, dans un premier temps, une fenêtre vers le sens du mot -, consiste à proposer, à l'aide de certaines clés informatives implémentés dans *Lexico 5*, quelques suggestions pour cerner le genre romanesque. C'est pour cela que dans un second temps, nous avons tentés d'en faire appel à *Treecloud* afin d'atteindre la qualification du mot et sa saillance sémantique.

III.1. Phase préparatoire

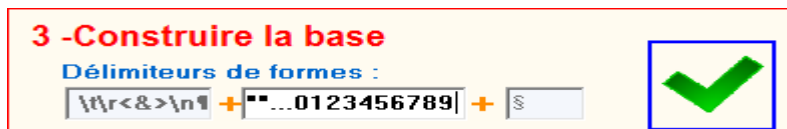
Avant de procéder à des dépouillements statistiques sur le texte, il est primordial de définir les normes (Labbé, 2010 : 35) lui permettant d'être segmenté en unités graphiques, reconnaissables par **Lexico 5**. Ainsi, pour que la machine puisse le traduire correctement, il suffit de l'enregistrer sous format brute en format (.txt). Pour ce qui est de la séparation des mots, le caractère est soit délimiteur (il sépare entre deux mots), c'est le cas de l'espace et des signes de ponctuations, soit il ne l'est pas. En définitif, une suite de caractères non délimiteurs, encadrée entre deux caractères délimiteurs (/_-'() [] {} \$.=«»“”‘’ ”’...) est donc une occurrence analysable (fig.1).

Fig. 1 : caractères délimiteurs par défaut.



Pour que des nombres, par exemples, ne soient pas pris en compte, il suffit d'ajouter les caractères suivants (0123456789) ou de les retrancher si l'on veut que ceux-ci soient comptés comme des occurrences d'un seul vocable Fig.2).

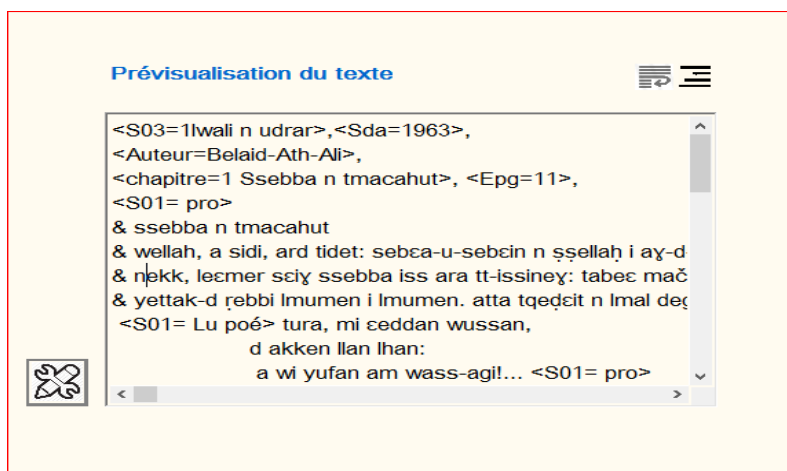
Fig. 2 : caractères délimiteurs ajoutés.



III.2. Phase du décodage

L'autre étape importante réside dans la délimitation du texte et, subséquemment, de la méthode adéquate à suivre. La définition de ces clés informatives codées, appelées également balises : <type=contenu>, constituera donc le premier point. L'exemple suivant : <S03= 1 Lwali n Udrar>, <Sda = 1963>, <Auteur= Belaid Ath-Ali>, <Epg= 11>, <S01= pro>, donné comme focus sur le fonctionnement d'ensemble (Lafon, Lefèvre, Salem et Tournier, 1985), constituera le second point. Quant au troisième point, il sera développé autour de la manière de générer un nuage-arboré, à partir d'une concordance, appliquée sur un corpus kabyle. En résumé, l'intérêt de ces codages réside donc dans la récupération des métadonnées du texte immorté afin de prendre en compte les éléments du contexte d'un texte (Blanc, Peton et Garcias, 2019) (Fig.3)

Fig. 3 : quelques balises utilisées dans notre corpus.



IV. Résultats et discussion

IV.1. Clés

La balise, dans le langage informatique, est une suite syntaxique qui se compose de cinq éléments juxtaposés ; elle permet de délimiter une séquence de caractères ou de marquer une position précise à l'intérieur d'un fichier texte. Elle s'écrit de la manière suivante :

Coder une clé.

Une clé (ex : <Auteur=Belaïd Ath-Ali>) se compose de 5 éléments :

1	<	un chevron ouvrant
2	Auteurs	le type de la clé
3	=	le signe qui sépare le type du contenu
4	Belaïd Ath-Ali	le contenu de la clé (valeur variable)
5	>	un chevron fermant

Exemple <année=1983>, <Auteur = Rachid Aliche>

L'insertion de ces *clés* constitue une phase importante dans la préparation du texte. Grâce à elles, l'utilisateur peut effectuer différentes comparaisons et types d'analyses à partir du corpus qu'elles découpent (Lamalle, Martinez, Fleury, Salem, Fracchiolla, Kuncova et Maisondieu, 2003). La syntaxe d'une balise (Mouelhi 2008 : 65) désigne donc en tout cinq types de clés prédéfinies. Trois d'entre elles ne nécessitent pas de suffixe et deux exigent l'utilisation. Il s'agit d'abord de la balise « *F* » qui

sert à l'archivage et à l'identification du fichier au sein de la base textuelle ; elle est toujours placée en tête de l'édition électronique. Puis de la balise « **C** » qui sert d'aide mémoire à l'intension des usagers ; sa particularité, en plus d'être automatiquement enjambée par le logiciel, est qu'elle s'insère à n'importe quel endroit du texte. Ensuite, vient la balise « **L** » qui indique les modalités et changements typographiques survenus lors de la réédition du texte de référence. Mais puisque nous ne possédons pas encore une base textuelle assez conséquente, les balises à suffixes suffisent largement aux besoins de notre recherche.

Par ailleurs, chaque corpus requiert une hypothèse et chaque hypothèse une approche. Ainsi, pour pouvoir s'appuyer sur une division du texte en volume de : nombre de pages, paragraphes, lignes, phrases, etc., i.e. en partie "naturelle" du texte, on peut assigner au caractère (<Epg=xx>, « & », <Elg=xx>, <Eph=xx>, ...) le rôle de délimiteur. Le fait de coder ainsi le texte¹ nous permet d'étudier la répartition des formes à l'intérieure des sections ainsi constituées (par ex. on n'aura aucun mal (cf. *infra* & fig.4 et 6) à trouver que le connecteur temporel (*Lħaşun* « en tout cas ») s'utilise systématiquement en début de phrase). Le logiciel nous le donne sous forme d'un carré en couleur qui, dans l'exemple ci-après, se situe au : 17ème paragraphe, 23ème segment phrastique et plus exactement à la page 13 (Fig. 4)

Fig. 4 : repérage d'un mot via la balise « Epg »

Formes (ordre lexicographiq...)	Fréquence
ma	36
nekk	34
t	31
lw	29
tt	28
am	27
dya	27
lyl	27
lakin	27
agl	24
ayen	24
ccix	23
k	23
yef	21
ylwen	21
armi	20
rebba	20
mi	19
win	19
a	18
netta	18
ħaşun	17

Sections
Délimiteurs

Partition: Eng

Seuillage

Effacer Exporter

Sections

Occurrences

¶ & ħaşun , ass - ma , yerlem rebba ar s tider - iv , ula d netta yerfel - <Eph=33>deg akken yerfel , yerfa - tt d urfan , di syen yerfa ttar s toumni , la yi - d - yeskad kan , itett sekou d <Epg=13>wadi - nni , la yi - d - yeqqe deg wadi - is : a tebera tin akken ara yaru rebba deg uqerru - iv ur ak - tt - maly d tamawt almanaa tcedbañ - tt ! <Eph=34>

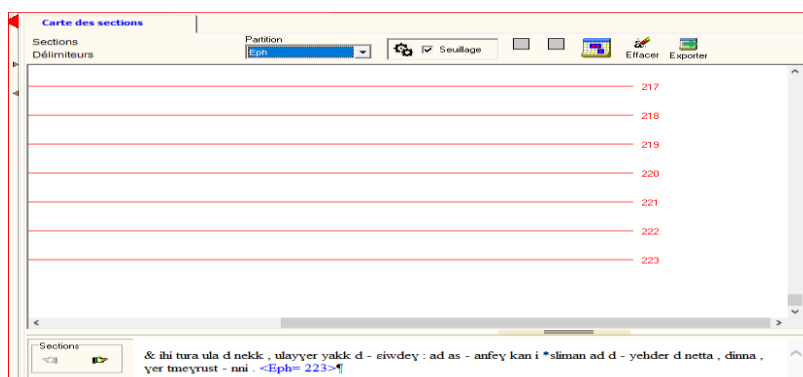
La balise « **Eph** » (cf. & Fig.5) peut également apporter à notre connaissance le phénomène de la longueur phrastique comme c'est le cas chez Belaid Ath-Ali. Pour 4456 occurrences, par exemple, on peut calculer facilement la moyenne par phrase qui est de 20 occurrences, si on le rapporte

¹ Nous n'avons considéré qu'une seule partie à savoir : le premier chapitre « ssebba n tmachut » du premier roman « *Lwali n Udrar* » « *Saint de la montagne* » tandis que ce phénomène d'utilisation de connecteurs en suremploi est une caractéristique stylistique qu'on puisse généraliser dans l'écriture de Belaid Ath-Ali. Ce codage nous dira également combien de phrases, de pages, de paragraphes qu'une section peut en contenir ; il permet également de repérer facilement les différentes parties du discours à l'aide de l'outil carte de sections.

au nombre total qui s'y trouve dans le premier chapitre, à savoir 223 phrases. Ce qui est déjà beaucoup si on s'amuse à le comparer à l'écrit de Proust dont Degezelle² disait que Belaid en avait lu beaucoup ; et savait particulièrement apprécier le style et rendait compte de ses lectures avec une grande précision.

De la sorte, deux lettres préétablies (« **pg** » et « **lg** ») - tout en ayant la possibilité d'ajouter autant de préfixes que l'on veut « **Eph** » – doivent être juxtaposés à la balise principale « **E** » pour permettre une segmentation en pages, lignes et même en nombre de phrases contenues dans une section et par conséquent, dans tout le corpus (Fig.5).

Fig. 5 : repérage du segment phrastique par la balise « Eph »



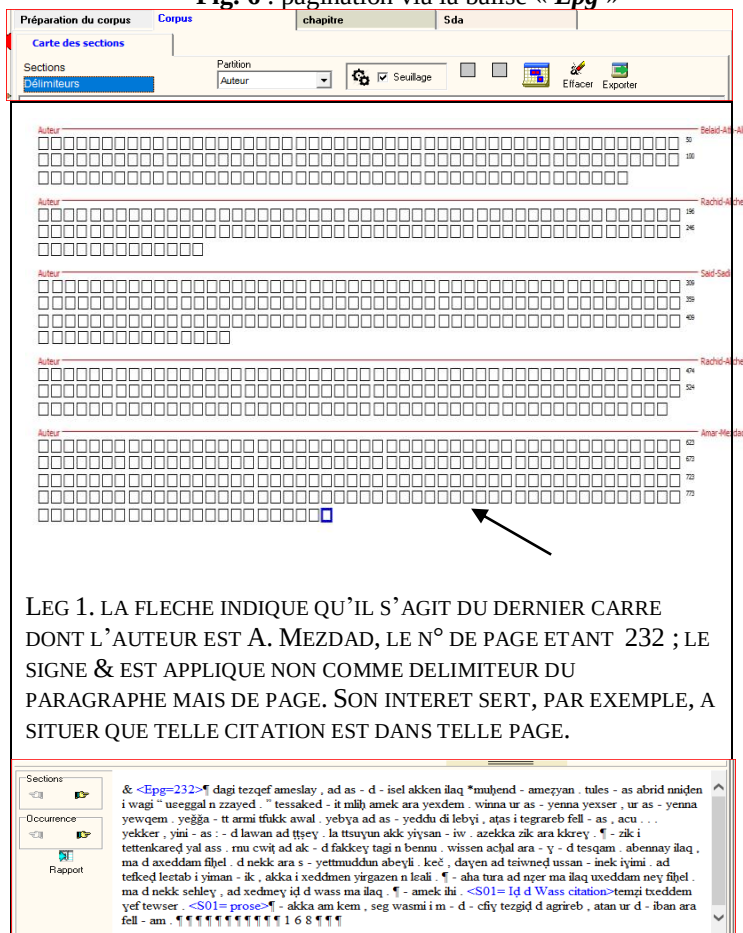
Pour le paragraphe, quoique nécessaire pour une mise en reproduction, à l'idéal, au sein de l'édition électronique du texte source, son découpage en paragraphes, qui consiste à numéroter séquentiellement ceux-ci, n'est pas d'un très grand intérêt pour rechercher une forme occurrence dans le texte. En effet, essayer, par exemple, de « situer une occurrence dans le 158^{ème} paragraphe d'un texte n'est pas très commode pour le retrouver ! » (Lafon, Lefèvre, Salem et Tournier, 1985 : 22). Par ailleurs, pour marquer le début de chaque paragraphe, il suffit d'ajouter le signe paragraphe « & » devant chacun de ceux-ci.

Est il important de souligner ici que c'est l'outil carte de section qui s'en charge pour reproduire la mise en forme de chaque paragraphe en carré, mais qu'il soit entendu que la définition du paragraphe ce n'est pas uniquement une séquence de textes comprise entre le début d'une page et un retour chariot ou entre un retour chariot et la fin d'une page. Opération qui peut être effectuée de manière générique par le logiciel en remplaçant le caractère « retour chariot » par la séquence « retour chariot » suivi de « & » et, à défaut par Word, via la commande : chercher : ^p Remplacer par : ^p &. Pour le cas des vers de poésie, des passages citationnels, ceux-ci sont rattachés aux paragraphes précédents, sauf quant ils commencent une page

² Cf. à ce propos Dallet Jean-Marie & Degezelle Jules (édit), les cahiers de Belaid ou la Kabylie d'antan (T : 1 : textes), p. XI. Et à propos du style de Proust Cf. Labbé Cyril. & Labbé Dominique. (2018). « *Les phrases de Marcel Proust.* » In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data. Roma: UniversItalia, 2018, p. 400 - 410.

nouvelle. Sur ce point, on aurait aimé d'ailleurs que le système de balisage fourni en XML³ soit d'apport pour compléter, de manière plus juste et adéquate, le balisage concernant les passages poétiques. Rien n'empêche, si c'est uniquement pour décomposer le corpus en sous corpus (poésie et prose), on peut n'utiliser que la balise spécifiquement conçue pour le logiciel (Fig. 6).

Fig. 6 : pagination via la balise « Epg »



Le cas des lignes peut s'avérer très encombrant pour l'édition électronique, ce pourquoi on peut ne pas introduire de balises pour le

³ La technologie XML permet d'analyser non seulement les fréquences des occurrences, mais également leur rôle et position au sein de chacun de ces sous-ensembles tels que la section, la strophe ou encore le vers. Ainsi, chaque texte peut être abordé, dans une perspective d'ensemble ou, individuellement grâce à la précision donnée par les différentes annotations. Cf. Clémence Jacquot. (2012), « *continue et discontinue dans l'œuvre poétique d'Apollinaire : la phrase Apollinarienne au croisement de l'analyse stylistique et des statistiques textuelles* ». In congrès mondial de la linguistique française–CMLF, SH Web of conférences. EDP Sciences. p.1104. https://www.shs-conferences.org/articles/shsconf/pdf/2012/01/shsconf_cmlf12_000131.pdf

marquage des lignes. De ce fait, il suffit juste de les remplacer par une barre oblique « / » à l'extrémité de chaque ligne. « **Lexico 5** se charge alors d'incrémenter le compteur de lignes chaque fois qu'il rencontre une barre oblique et le réinitialiser à chaque page, c'est-à-dire à chaque fois qu'il rencontre une nouvelle balise <Epg> » (Mouelhi, 2008 : 68).

Par ailleurs, si notre objectif est sériel, discursif, voir stylistique, l'organisation du texte requerra alors une division en parties : (auteur, édition, titre et sous-parties) : chapitres, (actes, scènes, etc.). Intervient alors la balise « **S** » dont le rôle sert plus à la réactivation du lien hypertextuel que du texte lui-même.

Dans l'exemple ci-après, (cf. *infra* & fig.7) la figure nous donne à voir un corpus divisé en sous-parties, dont chacune d'elles renvoie à un titre du roman, suivant donc leur ordre d'appariation dans le temps. Si les titres sont compositionnels dont l'ordre est synaptique et nominal (par apposition), on est en droit de supposer que, le premier roman, traite d'un seul mode d'être et, le dernier, d'univers antithétiques. Quant aux trois autres, articulés sous forme d'un mot unique (indécomposable), leur niveau d'abstraction est tel qu'on ne peut rien présupposer d'avance. Cet exemple donne pour voir comment le corpus, au lieu d'être corseté par des hypothèses trop contraignantes, est susceptible de nous interroger *bottom up*, ce qui implique la méthodologie du *corpus driven* et non *corpus based* (Mayaffre, 2010 :38).

S'agissant de la balise « **S** », elle sert de manière générale, à introduire, dans l'édition électronique, « des informations extrêmes, des connaissances détenues par le chercheur, qui caractérisent tout ou partie de textes. Chaque type de clef définit une variable de situation. L'ensemble des variables constitue un système, qui permet de diviser le corpus en fragment homogène ou d'extraire du corpus des sous ensembles textuels correspondants à des valeurs déterminées de certaines variables » (Lafon, Lefèvre, Salem et Tournier, 1985 : 24).

Le chercheur a la latitude d'introduire autant de balises de ce type (cf. *infra* & fig.7) qu'il souhaite, sans aucune restriction, car extensible et de nature variable, la balise « **S** » peut représenter, outre la caractéristique (Auteur, date, titre, chapitre), l'ensemble de l'appareil énonciatif et interactionnel de n'importe quel discours. De ce fait, ce sont « les objectifs d'analyse et les traitements prévus sur le texte qui définissent le système des variables extrêmes retenues pour figurer dans l'enregistrement » (Lafon, Lefèvre, Salem et Tournier, 1985 : 24). Dans ce qui va suivre, nous traitons, cas par cas, de toutes les variables situationnelles possibles notamment celles applicables sur notre corpus (Ikherbane, Tiziri et Boukherrouf, 2020) (Fig. 7).

Fig. 7 : balise " paratextuelle " ⁴ par la Clé S03

S03							
	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme
✓	1	1lwalli n udrar	33678	4210	2212	2570	d
✓	2	2 asfel	14711	3714	2263	1195	d
✓	3	3 askuti	32439	5027	2820	2347	d
✓	4	4 taffa	23511	5639	3410	1418	d
✓	5	5 id dwass	47776	6428	3262	3461	d
	T	Corpus	152115	14706	7167	10991	d

La balise *Auteur*, (cf. *infra* & fig.8) somme toute paratextuelle, laisse émerger des taxonomies stylométriques. S'est on posé la question, d'ailleurs, pourquoi certains écrivent en phrases courtes, tandis que d'autres ont le penchant indélébile pour des phrases complexes comportant plusieurs propositions ? Pour François Dominic Laramée (2019), Il est inconcevable, d'imaginer deux auteurs utilisant, à la fois, les points-virgules, les tirets et autres signes de ponctuation exactement à l'identique (Fig. 8)

Fig. 8 : la balise stylistique ⁵ *Auteur*

Auteur							
	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme
✓	1	Amar-Mezdad	47776	6428	3262	3461	d
✓	2	Belaid-Ath-Ali	33678	4210	2212	2570	d
✓	3	Rachid-Aliche1	14711	3714	2263	1195	d
✓	4	Rachid-Aliche2	23511	5639	3410	1418	d
✓	5	Said-Sadi	32439	5027	2820	2347	d
	T	Corpus	152115	14706	7167	10991	d

⁴ Le paratexte de l'œuvre est « ce par quoi un texte se fait livre et se propose comme tel à ses lecteurs, et plus généralement au public. Plus que d'une limite ou d'une frontière étanche, il s'agit ici d'un seuil [...] qui offre à tout un chacun la possibilité d'entrer, ou de rebrousser chemin. » Genette Gérard (2002), *Seuils*, Editions du Seuil, pp, 7-8. Dans notre cas, il s'agit du titre, perçu comme premier élément, jeté comme fleur aux yeux du lecteur.

⁵ Vu comme (écart par rapport à la norme), entre les œuvres d'un même auteur, par rapport à son prédécesseur /contemporaine, appartenant à une école ou à un idiome différent, la stylistique est « *essentiellement synchronique... mais elle est susceptible de point de vue historique et diachronique et peut étudier le style dans son évolution.* » Guiraud Pierre. (1961), *La stylistique*, Paris, presses universitaires de France, p. 104.

La tendance sérielle (*cf. infra & fig.9*) dans le corpus s'annonce dès lors qu'une division par date saute aux yeux du chercheur ; c'est ce que la statistique textuelle appelle une série chronologique, autrement dit, un ensemble de textes possédant cette particularité sérielle étendus sur un axe du temps chronologique, dont les conditions de production du discours donnent une certaine cohérence hypothétique (De Sousa, 2009 : 72). A l'évidence, nous ne pouvons passer sous silence toute une série de traitements possibles qui, peuvent varier, selon la nature du corpus et le choix de l'hypothèse scientifique, et motiver, aussi bien, l'évolution des concepts, l'émergence du sens, l'édition du texte et sa réception auprès du lectorat, etc.

Pour cela, il suffit donc de faire accorder à chaque variable un suffixe particulier, sans omettre que les deux suffixes alphabétiques préétablis doivent obligatoirement correspondre aux deux variables suivantes :

- La date de livraison du texte : cette variable est représentée par les suffixes et les contenus suivants : **<Sda=1963>** : le suffixe « **da** » représentant (date-année) ayant pour contenu l'année d'écriture réelle ou supposé du texte ou, à défaut, un quelconque indice référencié. Et s'il ya un différent sur la datation d'un ouvrage, ou l'anonymat de son auteur, le chercheur peut dépister la faille en recourant soit par recoupement du document avec ceux de son époque, soit en utilisant d'autres méthodes – plus sophistiquées - telles que des ressources dictionnairiques informatisées ou bien à des chaînes de caractères programmables en python (Laramée, 2019) (Fig.9).

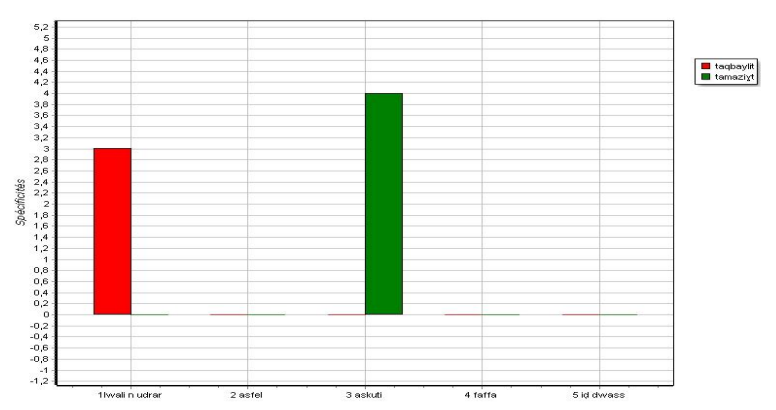
Fig. 9 : la baisse sérielle par la Clé Sda

Sda								
	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme	
✓	1	1963	33678	4210	2212	2570	d	
✓	2	1981	14711	3714	2263	1195	d	
✓	3	1983	32439	5027	2820	2347	d	
✓	4	1986	23511	5639	3410	1418	d	
✓	5	1990	47776	6428	3262	3461	d	
	T	Corpus	152115	14706	7167	10991	d	

- L'attribution du texte : puisque le texte peut être aussi appréhendé de point de vue de son émetteur, la variable qui sert à étudier son système énonciatif se représente alors par le suffixe « **at** » et peut prendre les contenus suivants :
<Sat=0> : tous les passages où l'auteur prend la parole (monologue de base).
<Sat=1> : tous les passages où l'auteur cite ou reformule (autocitation ou dialogue).
<Sat=2> : passage cité par l'auteur au style direct et qui n'est pas de lui.
<Sat=3, 4, ..., n> : autres valeurs, laissées au libre choix du chercheur.

Par contre, si (*les clés* « Sat ») ne se trouvent sciemment explicitées, ce n'est pas tant par défaut du codage mais juste parce qu'elles n'intéressent pas le sujet de notre thème. Mais, dans une autre perspective, il serait tout à fait intéressant de coder la parole dans le texte selon la partition locuteur, par exemple, mais puisque les ressources du logiciel ne permettent de synthétiser (cf. & *infra* fig. 9) tout le stock lexical de celui-ci, il est préférable d'essayer d'autres ressources, à moins qu'elles ne rejettent elles aussi, le caractère « Unicode-Amazigh », comme c'est le cas de *Treecloud* qu'on verra plus loin. Dans le même sillage, le logiciel *TexTobsever* (Le Blanc, 2017 : 48), un des rares qui acceptent d'autres codages (ANSI ou UTF8), sans dictionnaires ni étiqueteurs, dont l'utilisation reste strictement lexicaliste, ne dépassant pas le cadre de surface (Fig.10).

Fig. 10 : ventilation des formes (*taqbaylit/tamaziyt*)



En plus de ces deux variables à savoir (l'attribution et la date) qu'on vient de voir supra, la balise « S » permet aussi « d'introduire des indications thématiques, des genres (poésie, prose, etc.), ou encore de séparer différents niveaux du texte (titre, sous titre, etc.). Dans tous les cas, le suffixe de la clé sera choisi dans la série 01, 02, ..., 99 (en évitant bien évidemment les combinaisons déjà utilisées) ; son contenu sera un code alphanumérique. » (Lafon, Lefèvre, Salem et Tournier, 1985 : 25). A ce titre, notre travail tente de recréer une partition dans ce sens : l'objectif étant de proposer des comparaisons interdiscursive ou même intertextuelles. Dans cet exemple, si le discours poétique dans '*Asfel*', excède avec 288 occurrences et 158 formes différentes, agit en contradiction avec sa taille générale, ce n'est pas le cas du discours citationnel dans '*Id d Wass*' avec 290 occurrences et 181 formes, qui serait tout à fait dans un rapport logique avec celle-ci. De ce fait, une hypothèse peut alors émerger dans le sens que le vocabulaire dans '*Asfel*' serait plus riche, surtout que de nombreuses études (Kastberg, 2004 : 158) s'y réfèrent à la richesse d'un texte notamment lorsque celui-ci intègre d'autres types du discours (Fig.11)

Fig. 11 : balise interdiscursive⁶ (Poésie et citation/prose)

Parties		Ventilation	AFC	Sp. tot.				
S01								
	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme	
✓	1	Asfel citation	88	70	56	3	d	
✓	2	Asfel poésie	288	158	110	26	d	
✓	3	Askuti citation	167	110	88	18	d	
✓	4	Askuti poésie	175	117	90	13	d	
✓	5	Faffa citation	41	36	31	2	d	
✓	6	Faffa poésie	133	90	72	10	ad	
✓	7	Id d Wass citation	290	181	137	15	ur	
✓	8	Lwali n Udrar citation	55	45	38	4	d	
✓	9	Lwali n Udrar poésie	48	39	34	4	d	
✓	10	prose	150831	14621	7144	10902	d	
	T	Corpus	152116	14707	7168	10991	d	

IV.2. Visualisation à l'aide d'un nuage arboré

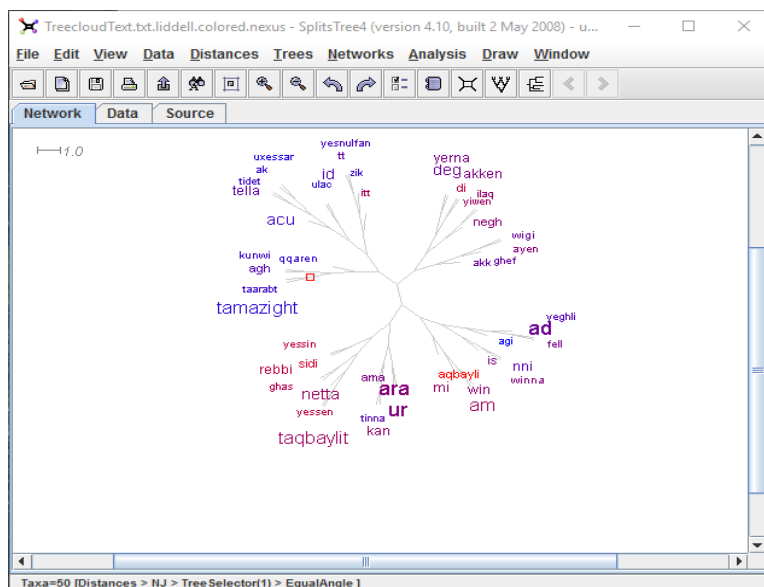
Comme il été mentionné précédemment, le maillon faible des études berbères, et à moindre degré celles portant sur l'Arabe, reste le domaine lexicométrique qui, en dépit de la rareté des ressources en « bases de données » et en « outils du traitement », « de nombreux travaux ne cessent de se déployer dans cette perspectives » (Annouz, 2019). On peut citer dans ce sens les travaux de (A. Amatoui : 2010) portant sur les patrons intonatifs de la phrase en kabyle ; de (F. Yamouni : 1994), sur l'étude de la phrase déclarative simple via Nooj ; de (S. Loikkanen : 1997, 1998) sur l'étiquetage morphosyntaxique; de (M. Mahrazi & Dj. Nahali : 2012) sur la création de générateur automatique de l'état d'annexion ; de (M. Mahrazi : 2013) sur le schéma d'annotation des passages entre guillemets ; de (H. Ainnouz : 2016) sur "*treatments of Kabyle derived Nominal Verbs with Nooj*", etc.

En tout cas, elles sont légions ces langues dont le souhait d'intégrer la TALN, en toute urgence, se fait plus que jamais ressentir. L'exemple que nous allons donner porte sur le défaut de codage d'un de ces logiciels qui, malgré sa relative ancienneté, refuse d'implémenter en son sein les caractères spéciaux et par conséquent, de les traduire correctement. Ce logiciel n'est autre que *Treecloud*⁷ qui, pour le mettre en place nécessite toute une série de logiciels annexes (Fig.12).

⁶ L'interdiscours selon Maingueneau serait plus qu'une somme des circonstances entourant le discours, mais comme une modalité du discours agissant en catimini sur l'énonciation du sujet. Maingueneau Dominique. (1994), L'énonciation en linguistique française, Éditions Hachette Livre, Paris, p. 153.

⁷ Pour plus de détail sur le fonctionnement de ce logiciel cf. TreeCloud – www.treecloud.org. Egalement le manuel d'utilisateur dans le fichier *ManualTreecloud.pdf* situé dans le même dossier que le programme *Treecloud.exe*.

**Fig. 12 : nuage arboré : application sur la concordance des mots
« Aqbayli » et « Amaziy » avec leurs formes fléchies.**



Conclusion

Nous ne pouvons sanctionner cette étude en passant outre l'intérêt du *Lexico5* pour le balisage des genres qui, en plus de se poser en bon allié lexicométrique, utilise la norme Saint-Cloud. Son apport se trouve pleinement en phase avec notre objet d'étude, notamment en raison de la convergence de son programme avec la *TEI* (*Text Encoding Initiative*) plus précisément écrit dans le langage XML. Puisque notre corpus est balisé selon cette norme précitée, son "XMLisation" ou son transfert vers le format *TEI*, ne devrait pas poser de problèmes, notamment en raison de la transportabilité de l'un vers l'autre.

Par ailleurs, un des problèmes majeur auquel nous avons fait face portait sur le codage des textes et l'incompatibilité de codage UTF-8 avec les logiciels existants en notre possession et plus précisément *Treecloud* ainsi que les versions antérieures à *lexico 5*. Nous avons attendu jusqu'à la version 5.8.1 pour que le codage UTF-8 soit en mesure de supporter une grande quantité de caractères, y compris tous les signes diacritiques disponibles sous format Amazigh-Unicode.

Dans l'espoir que des logiciels existants – actuellement sur le marché – puissent se pencher sur cette problématique, nous souhaitons inscrire notre recherche dans cette voie (de la littérature informatisée) œuvrant dans le sens d'une démarche qui consisterait à se servir de logiciels spécifiques qui soient à même de prendre en charge des corpus de la langue berbère.

Nous avons démontré la faisabilité de la création dynamique et ajustable de plus d'un outil de traitement automatique en vue de mener une étude du genre sur un texte littéraire berbère. Pour ce faire, nous avons d'abord

constitué un corpus de plus de 152000 mots ; appelé à s'étendre dans l'avenir proche, afin de constituer un matériau pour des études littéraires berbère et ensuite à effectuer déférentes repérages sur le niveau textuel, de la structure intermediaire ainsi que du (mot-forme) maintenue au niveau de sa granularité inferieure en tant que unité textuelle de base. Mais pour que les résultats soient exhaustifs, il va falloir en préalable nettoyer, désambiguïiser et le cas échéant baliser, lemmatiser et enrichir le corpus à laide d'annotation et à partir d'un système d'étiquetage et ce dans tous les endroits du texte et des niveaux de l'analyse. C'est ce que nous envisageons de faire, tout en ouvrant à enrichir notre corpus par de nouvelles bases de données, notamment avec *Lexico 5* d'autant plus que son système permet de prendre en charge un corpus aussi vaste soit-il en un temps record.

Bibliographie

- Amatoui, Amirouche, 2009, « Les patrons intonatifs de la phrase en kabyle » *Iles d Imesli*, volume 1, numéro1, pp. 89-121.
<https://www.asjp.cerist.dz/en/article/37450>
- Amstutz, Delphine et Gambette, Phillipe, 2010, « Utilisation de la visualisation en nuage arboré pour l'analyse littéraire. » *JADT'10: 10th International Conference on statistical analysis of textual data*, Rome, Italie.
- Annouz, Hamid, 2019, Traitement morphologique des unités linguistiques du kabyle à l'aide de logiciel Nooj : Construction d'une base de données, thèse de doctorat, INALCO, Paris.
- Blanc, Antoine, Peton, Hélène et Garcias, Frédéric, 2019, « L'analyse lexicométrique des macro-discours par les vocabulaires – enjeux théoriques et méthodologiques », *Finance Contrôle Stratégie* [En ligne], NS-6], mis en ligne le 23 mai 2019, consulté le 21 octobre 2021. URL : <http://journals.openedition.org/fcs/3459>
- Dallet, Jean-Marie et Degezelle, Jules, 1963, (édit), *les cahiers de Belaid ou la Kabylie d'antan (T : 1 : textes)*.
- De Sousa, Serge, 2009, « *Le discours de Fidel Castro : essai de lexicométrie politique* », in Fleury Serge et Salem André (coord.), *Lexicometrica*, n° spécial, exploration textométriques, vol. 2, séries textuelles chronologiques, p. 68-94.
- Gambette, Phillipe et Véronis Jean, 2009, « Visualising a text with a treecloud », In Locarek-Junge H. and Weihs C. editors, *Classification as a tool of research, proc. Of IFCS'09 (11th Conference of the International Federation of Classification Societies)*.
- Ikherbane, Arezki, Tigziri, Noura et Boukherrouf, Ramdane, 2020, « Base de données numérique des corpus kabyles et exploitation. Essai

d'analyse lexicométrique de la dimension identitaire dans le discours romanesque », *ARANEA* 2020, 6-7 novembre, Slovaquie (A paraître).

Jenny, Jacques, 1997, « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologiques française contemporaine. », In : *bulletin de méthodologie sociologique (B.M.S)*, n°54, Paris, pp. 64-112.

Kastberg Sjöblom, Margareta, 2004, *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*, Thèse de doctorat, sous (dir) Brunet Etienne, Mellet Sylvie, Université de Nice.

Labbé, Dominique, 2018, « Apports de la lexicométrie à l'analyse des entretiens. Le cas de l'enquête " Les Français et la politique". Avec une note d'Etienne Schweisguth et une réponse de l'auteur ». In *Colloque DIME-SHS "Des instruments au service de la recherche en sciences sociales"*, Paris, CNRS Université Grenoble-Alpes.

Labbé, Cyril et Labbé, Dominique, 2018, « *Les phrases de Marcel Proust.* » In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the: 14th International Conference on Statistical Analysis of Textual Data*. Roma : UniversItalia, p. 400 - 410.

Labbé, Dominique, 2010, *Normes de saisie et de dépouillement des textes politiques*. Cahier du CERAT.

Lafon, Pierre, Lefèvre, Josette, Salem, André et Tournier, Maurice, 1985, *Le Machinal. Principes d'enregistrement informatique des textes*, Paris.

Lamalle, Cédric, Martinez William, Fleury Serge, Salem André, Fracchiolla Béatrice, Kuncova Andrea et Maisondieu Aude, 2003, « Lexico3 - *Outils de statistique textuelle. Manuel d'utilisation.* », *SLEID-CLA2T*, université de la Sorbonne nouvelle - paris 3.

Laramée, François Dominic, 2019, « Introduction à la stylométrie en Python » traduction par François Dominic Laramée, *Programming Historian en français* 1, <https://doi.org/10.46430/phfr0003>.

Lavrentiev, Alexei et Heiden, Serge, 2014, « Exploration textométrique du corpus des dossiers de Bouvard et Pécuchet. », *Revue Flaubert*, Centre Flaubert, pp.1-12.

<https://halshs.archives-ouvertes.fr/halshs-00957029>
[file:///C:/A%20I-DOC/mes%20donn%C3%A9es/Gustave Flaubert revue 13 article Lavrentiev Heiden.pdf](file:///C:/A%20I-DOC/mes%20donn%C3%A9es/Gustave_Flaubert_revue_13_article_Lavrentiev_Heiden.pdf)

Lebart, Ludovic et Salem, André 1994, *Statistiques textuelles*, Paris, Dunod.
<http://www.cavi.univ-paris3.fr/lexicometrica/livre/st94-tdm.html>

- Leblanc, Jean-Marc, 2015, « Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie ». *Nouvelles perspectives en sciences sociales*, 11(1), 25–63. <https://doi.org/10.7202/1035932ar>
- Leblanc, Jean-Marc, 2017, *Analyse lexicométrique des vœux présidentiels*, (éd) ISTE, London.
- Longhi, Julien, 2020, « Explorer des corpus de tweets : du traitement informatique à l'analyse discursive complexe », In *Corpus, Bases, Corpus, Language-UMR 7320*. <https://doi.org/10.4000/corpus.4567>
- Loikkanen, Sinikka 1998, *Vocabulaire du roman kabyle (1981-1994)*, (dir) Salem Chaker, INALCO, Paris.
- Mahrazi, Mohand, Nahali, Djamel, 2012, « Essai de création d'un générateur automatique de l'état d'annexion », *Iles d'Imesli*, volume 4, numéro 1, pp. 207-228. <https://www.asjp.cerist.dz/en/article/37450>
- Mahrazi, Mohand, 2013, « Proposition d'un schéma d'annotation des passages entre guillemets », *Iles d'Imesli*, volume 5, numéro 1, pp. 215-229. <https://www.asjp.cerist.dz/en/article/45807>
- Maingueneau, Dominique 1994, *L'énonciation en linguistique française*, Éditions Hachette Livre, Paris.
- Mayaffre, Damon, 2010, *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique. Synthèse d'habilitation vol 1*, (dir) Sylvie Mellet, Université Nice Sophia Antipolis.
- Pincemin, Bénédicte, 2018, *Sept logiciels de textométrie*. (CNRS, Univ. Lyon, laboratoire IHRIM UMR 5317). <https://halshs.archives-ouvertes.fr/halshs-01843695/document>