Approche textométrique de la catégorisation des langues minoritaires

Giancarlo Luxardo, Françoise Rollan, Alain Viaut

EEE (Europe, Européanité, Européanisation) – CNRS
MSHA, 33607 PESSAC Cedex, France

Abstract

The notion of 'minority language' can be associated with a number of other notions such as 'regional language', 'immigration language' or even 'dialect'. The aim of the present study is to analyse the connections between these notions which are often seen to vary along the lines of national traditions or disciplinary approaches. The concepts needing to be considered belong to various domains, including language science (e.g. linguistic variation, language contacts) and institutional practices specific to different geopolitical realities (e.g. official status, territoriality). On the basis of a text corpus constructed from scientific publications containing the relevant terms and concepts, we carried out textual explorations. Several layers of statistical processing were applied (mainly, hierarchical classifications, correspondence analyses, co-occurrence analyses) and the results produced by two software tools (Alceste, TXM) were compared. These tools provide helpful hints in the design of a typology for the different notions examined in the study.

Keywords: textometry; sociolinguistics; legal, scientific and political discourse; lexical and semantic analysis

Résumé

La notion de langue minoritaire peut être mise en relation avec d'autres notions, comme celles de langue régionale, de langue d'immigration, voire de dialecte. On se propose ici d'analyser les connexions entre ces notions, présentant souvent des variations en fonction des traditions nationales ou des approches disciplinaires. Il s'agit de prendre en compte à la fois des concepts qui relèvent des sciences du langage (variété linguistique, contacts de langues, ...) que des pratiques institutionnelles propres à différentes réalités géopolitiques (statut officiel, territorialité, ...). A partir d'un corpus de textes constitué d'extraits de publications scientifiques centrés autour de notions retenues, nous avons procédé à des explorations textométriques. Plusieurs traitements statistiques ont été utilisés (principalement : classifications hiérarchiques, analyses factorielles de correspondances, analyses de cooccurrences) et les résultats obtenus avec deux logiciels (Alceste, TXM) ont été confrontés. Ces outils fournissent une aide dans l'établissement d'une typologie des notions étudiées.

Mots-clés: textométrie; sociolinguistique; discours politique, scientifique et juridique; analyse lexicale et sémantique

1. Introduction

Une politique linguistique concerne des choix sur l'usage réglé ou encadré par un appareil légal et institutionnel d'une langue ou de plusieurs langues et définit les relations entre ces

langues dans un territoire. Les débats autour de ces choix ont été renouvelés et réactualisés en ce qui concerne les langues moins répandues par la mise en application depuis une vingtaine d'années dans la majorité des pays européens de la Charte européenne des langues régionales ou minoritaires, adoptée dans le cadre du Conseil de l'Europe et en application depuis 1998. Celle-ci répond à des objectifs de sauvegarde du patrimoine culturel par la promotion de la diversité linguistique et la protection de langues menacées de disparition. Elle a permis de mettre en parallèle des traditions et des expériences différentes de mise en œuvre de ces politiques dans plusieurs pays. Dans le corpus qui nous servira de base de travail, un certain nombre d'extraits proviennent précisément d'études (cf. *infra*, Viaut, Woehrling) reposant sur cette convention.

Les commentaires portant sur les langues régionales ou minoritaires amènent souvent à choisir comment qualifier une langue et comment désigner ses propriétés vis-à-vis d'autres langues. Dans ce contexte, la langue est elle-même l'objet de l'étude à travers sa catégorisation. Il s'agit de classer et de définir des relations entre des notions en tenant compte de plusieurs variables qui dépendent de l'approche suivie (linguistique, géographique, juridique). On peut donc envisager des applications d'analyse multivariée de données, ce qui motive le travail présenté ici.

2. Le Corpus sur les langues minoritaires en Europe (CLME)

Notre étude des discours catégorisants sur les langues minoritaires procède d'une démarche comparative (sur plusieurs pays, plusieurs langues) et est basée sur la construction et l'exploitation en équipe d'un corpus de textes. La construction du corpus est réalisée par un travail de citation des notions utilisées par les experts du domaine ou dans des textes officiels. L'exploitation du Corpus «Catégories de langues minoritaires en Europe» (CLME), récemment mis en place à Bordeaux (Maison des sciences de l'homme d'Aquitaine, sous la responsabilité de A. Viaut)[1] devrait permettre de définir une typologie des notions, qui peut être utilisée suivant deux approches : soit une approche hypertextuelle, de mise en relation d'extraits à partir de termes communs, soit une approche intertextuelle, de recherche d'associations et de transformations des concepts de base dans leur contexte.

Actuellement, ce corpus CLME comprend deux volets, linguistique et juridique, pour chacun des sous-corpus : en français, en russe et en anglais jusqu'à présent. Le corpus utilisé pour cet article est un sous-ensemble du corpus linguistique en français. Il contextualise une série de notions reliées à celle centrale et globale de langue minoritaire au sens large et comprend donc entre autres des commentaires sur la charte européenne. Il résulte d'un choix lié à la méthode qui a présidé à sa mise en place. Son caractère apparemment hétérogène résulte de à la première phase de mise en place pour laquelle il fut convenu de tester plusieurs types de sources : monographie d'auteur, ouvrage collectif, dossier de revue. Si, à ce stade, le but n'était pas encore de tendre vers l'objectivité de la représentativité de ces sources, il fut néanmoins tenu compte d'une certaine pondération laissant ainsi une place significative à des auteurs tels que Calvet, d'abord, et également Boyer. Un lien commun entre les références est certes assuré par l'objet langue minoritaire. Néanmoins, si des ouvrages figurent en entier à côté d'articles, certains d'entre eux n'ont été productifs que pour peu ou très peu d'extraits (ex. Ducrot). La récurrence d'auteurs comme Boyer et Calvet est en outre représentative de leur place dans le

1 Mis en place de ce coprpus dans le cadre du programme de recherche Région Aquitaine Langues minoritaires et marges linguistiques en Europe (2008-2012) coordonné par A. Viaut (CNRS).

contexte français lié à notre objet. L'empirisme de notre choix (une des visées premières, à ce stade, ayant été la mise au point d'un outil méthodologique) a au moins l'avantage de présenter une image d'ensemble à la fin relativement représentative du champ concerné.

Au résultat, les notions répertoriées au cours de ce travail sont au nombre de 104 (voir tableau 1, pour la liste des 40 premières dans l'ordre alphabétique). Elles sont tirées de 207 extraits comprenant de 200 à 1500 caractères chacun, selon le protocole préalablement défini pour mettre en place cette base de données. Une liste de références bibliographiques figure dans le tableau 2 avec les codes utilisés pour désigner les auteurs par la suite.

1	créole	21	langue des pays colonisé
2	dialecte	22	langue d'Etat[2]
3	dialecte régional	23	langue d'immigration
4	idiome	24	langue d'isolat
5	langue ancestrale	25	langue dominante minoritaire
6	langue autochtone	26	langue dominée
7	langue commune[1]	27	langue dominée écrite
8	langue de communication	28	langue dominée non-écrite
9	langue de communication interethnique	29	langue d'origine
10	langue de diaspora	30	langue du peuple
11	langue de la communauté	31	langue en diaspora
12	langue de la diaspora	32	langue en voie de disparition
13	langue de la dispersion	33	langue et culture locales
14	langue de l'émigration et de la diaspora	34	langue grégaire
15	langue de migrants	35	langue historique
16	langue de minorité nationale	36	langue historique de l'Europe
17	langue dépourvue de territoire	37	langue identitaire
18	langue des colonisés	38	langue identitaire ou grégaire
19	langue des immigrants	39	langue locale
20	langue des migrants	40	langue maternelle

Tableau 1 : notions répertoriées

2 «langue commune» apparaît ici dans le sens de langue quotidienne, plus ou moins formalisée, se distinguant d'une langue plus codifiée et employée dans des registres plus soutenus.

3 Concerne les cas où telle langue officielle d'État peut être en situation minoritaire comme par exemple le suédois en Finlande ou l'irlandais en Irlande.

Références utilisées pour l'établissement du corpus	Code auteur
Airoldi, S. (2005). « Les choix linguistiques des entreprises multinationales : options diverses et contradictoires », in : Paulin, C. (coord.), *Multiculturalisme, multilinguisme et milieu urbain*, Presses universitaires de Franche-Comté, 7-22.	AIROLDI
Akin, S. (2006). « La Charte européenne des langues, les «langues des migrants» et les «langues dépourvues de territoire» », *Lengas, revue de sociolinguistique*, n° 59, 51-66	AKIN
Bidart, P. (1991). «La révolution française et la question linguistique», in : *1789 et les Basques*, Presses universitaires de Bordeaux, 145-170	BIDART
Blair, Ph. (2006). « Conception et expérience de la territorialité linguistique à travers la Charte européenne des langues régionales ou minoritaires », *Lengas, revue de sociolinguistique*, n° 59, 11-20.	BLAIR
Boyer, H. (1991). *Langues en conflit : études sociolinguistiques*, L'Harmattan, 274 p. Boyer, H. (2001). *Introduction à la sociolinguistique*, Dunod, 104 p. Boyer, H. (2005). « Continuité et prégnance d'une désignation stigmatisante sur la longue durée », *Lengas, revue de sociolinguistique*, n° 57, 73-92.	BOYER
Bruneau, M. (2004). *Diasporas et espaces transnationaux*, Anthropos, 249 p.	BRUNEAU
Calvet, L.-J. (1979). *Linguistique et colonialisme*, Payot, 228 p. Calvet, L.-J. (1993). *L'Europe et ses langues*, Plon, 234 p. Calvet, L.-J. (2002). *Le marché aux langues, les effets linguistiques de la mondialisation*, Plon,220 p.	CALVET-ouvrage
Calvet, L.-J. (1996). « La France a-t-elle une politique linguistique ? », in : Juillard, C. & Calvet, L.-J., *Les politiques linguistiques, mythes et réalités*, FMA, 89-101. Calvet, L.-J. (2000). « La guerre des langues et les chances d'un véritable plurilinguisme », *Panoramiques*, n° 48, 10-16. Calvet, L.-J. & Varela, L. (2000). « XXIè siècle : le crépuscule des langues ? Critique du discours Politico-Linguistiquement Correct ». *Estudios de Sociolingüística*, n° 1(2), 47-64.	CALVET-article
Commission Européenne (Euromosaic) (1996). *production et reproduction des groupes linguistiques minoritaires au sein de l'Union Européenne,* Office des publications officielles des Communautés Européennes, 66 p.	COMMISSION
Courouau, J.-Fr. ; Lieutard, H. (2006). « Petites langues d'Europe : le luxembourgeois, le sarde, et le croate du Burgenland », *Lengas, revue de sociolinguistique*, n° 60, 9-13.	COUROUAU
Drettas, G. (2007). « Formes de la langue grecque en diaspora », in : Bruneau, M., Hassiotis I., et alii (éds.), *Arméniens et Grecs en diaspora : approche comparative*, Ecole française d'Athènes, 549-562.	DRETTAS
Ducrot, O. et Todorov, T. (1972). *Dictionnaire encyclopédique des sciences du langage*, Le Seuil, 470 p.	DUCROT-TODOROV
Giblin, B. (2002). « Langues et territoires : une question géopolitique », *Hérodote*, n° 105, 3-14.	GIBLIN
Guillorel, H. (2006). « Démocratie, territoire et langue dans la Charte européenne des langues régionales ou minoritaires », *Lengas, revue de sociolinguistique*, n° 59, 37-50.	GUILLOREL
Herdam, A. (2005). « L'allemand kanak, de l'insulte au phénomène de mode », *Multiculturalisme, multilinguisme et milieu urbain,* Presses universitaires de Franche-Comté, 121-137.	HERDAM
Jetchev, G. (2006). « Elements de politique linguistique de l'Etat bulgare », *Lengas, revue de sociolinguistique*, n° 60, 190-203.	JETCHEV
Koulayan, N. (2003). « Le français, langue diasporique d'un genre spécifique ? », *Langues dépaysées, Diasporas, Histoire et sociétés*, n° 2, 120-132.	KOULAYAN
Léonard, J. L. (2006). « La variation interlangue et dialectale des langues finno-ougriennes de la Volga : planification linguistique et aspects structuraux internes », *Lengas, revue de sociolinguistique*, n° 60, 115-142.	LEONARD
Lespoux, Y. (2006). « L'instruction publique et les patois dans les Basses-Pyrénées des années 1880 aux années 1930, d'après le Bulletin de l'Instruction primaire des Basses-Pyrénées », *Lengas, revue de sociolinguistique*, n° 59, 165-181.	LESPOUX
Marcellesi, J.-B. (2003). *Sociolinguistique. Epistémologie, langues régionales, polynomie*, L'Harmattan, 308 p.	MARCELLESI
Rollan, Fr. (2006). « Les politiques linguistiques et les frontières en Asie centrale ex-soviétique », *Lengas, revue de sociolinguistique*, n° 60, 143-171.	ROLLAN
Schanen, Fr. & Lulling, J. (2006). « Lëtzebuergesch : la langue nationale du Grand Duché de Luxembourg », *Lengas, revue de sociolinguistique*, n° 60, 12-48.	SCHANEN
Sintas, S. (2005). « Le bilinguisme, cheval de bataille du parti populaire dans l'archipel baléare : du slogan politique à la réalité sociolinguistique », in : Paulin, C. (coord.), *Multiculturalisme, multilinguisme et milieu urbain*, Presses universitaires de Franche-Comté, 263-283.	SINTAS
Viaut, A. (2006). « Les langues historiques de l'Europe », *Lengas, revue de sociolinguistique*, n° 59, 67-81.	VIAUT
Woehrling, J.-M. (2005). *La charte européenne des langues régionales ou minoritaires : un commentaire analytique*, Editions du Conseil de l'Europe, 323 p.	WOEHRLING

Tableau 2 : sources

La principale variable utilisée dans la suite pour définir une partition du corpus est la variable *auteur*, qui, comme l'indique le tableau 2, comporte 25 modalités. On notera que les textes de L.-J. Calvet, dont le volume est le plus important dans le corpus (90 extraits sur 207) ont été associés à deux modalités (*CALVET-article* et *CALVET-ouvrage*) afin de réduire cette disproportion dans la distribution.

Ce corpus CLME a été soumis à des explorations textométriques au moyen de deux logiciels : Alceste et TXM. Il s'agit d'un choix méthodologique qui permet de mieux valider les résultats obtenus et qui est simplifié par la possibilité d'importer directement dans TXM un corpus au format Alceste.

Comme la plupart des logiciels de statistique textuelle de l'école française, Alceste et TXM sont des logiciels qui s'appliquent aussi bien à des corpus littéraires (textes monolithiques d'un seul auteur) qu'à des entretiens ou des questionnaires (agrégation des réponses d'un nombre élevé d'individus). Notre corpus CLME présente des caractéristiques intermédiaires puisqu'il est constitué de courts extraits et que l'on cherche à trouver des relations entre des positions exprimées par différents auteurs.

Dans un premier temps, Alceste a été utilisé pour classer des énoncés et reconnaître les thèmes émergents du corpus. Les résultats fournis par TXM sont également présentés : essentiellement, analyse du lexique et exploration des relations entre les différents auteurs.

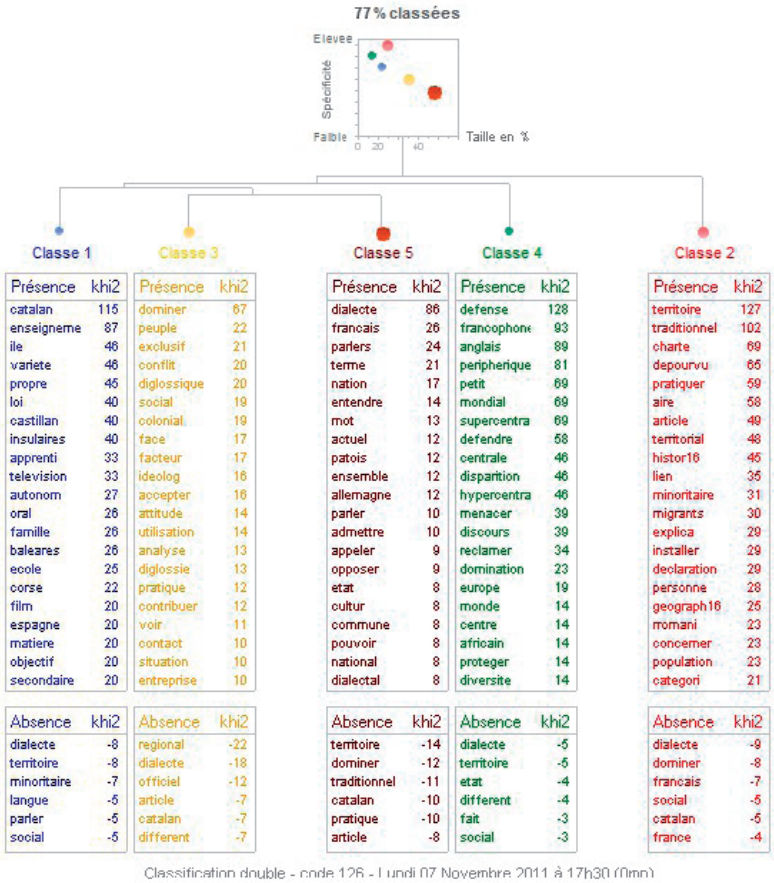
3. Classification en mondes lexicaux

L'originalité d'Alceste dérive du découpage du corpus en unités de contexte élémentaires (UCE), identifiées de façon semi-automatique. La principale fonctionnalité met en œuvre une classification descendante hiérarchique faisant émerger des "mondes lexicaux" qui peuvent aider à déceler la structure latente du discours.

Dans un travail précédent, portant sur un corpus similaire établi à partir d'écrits de L.-J. Calvet (Bassac *et al.*, 2009), Alceste a permis de reconnaître certaines composantes énonciatives qui se manifestent dans le discours de l'auteur. Un clivage a été identifié entre deux dimensions sémantiques : d'une part du point de vue de la gestion administrative des langues, d'autre part du point de vue de l'activité des locuteurs.

Etant donnée la nature du corpus CLME, qui relève de plusieurs auteurs, mais qui est aussi plus homogène en raison d'un sujet plus ciblé, celui de la notion de langue minoritaire, on se propose ici de fournir un point de vue différent, à partir d'une analyse contrastive des positions exprimées par les différents auteurs.

Plusieurs essais de classification ont été effectués avec Alceste. La classification retenue (représentée figure 1) fournit cinq classes, décrites ci-dessous dans l'ordre suivant lequel elles sont structurées par le dendrogramme.



Dans un premier temps, on ne retiendra que quelques mots significatifs (avec un Khi-2 plus élevé) qui peuvent caractériser ces classes dont voici une brève description en l'attente d'une analyse ultérieure :

Classe 2 : *territoire, traditionnel, charte, dépourvu, pratiquer, aire, article, historique, lien, migrant*

Classe 4 : *défense, francophone, anglais, périphérique, petit, mondial, supercentrale, hypercentrale*

Classe 1 : *catalan, enseignement, variété, propre, loi, castillan, insulaire*

Classe 3 : *dominer, peuple, conflit, diglossie, social, colonial, idéologie*

Classe 5 : *dialecte, français, parlers, nation*

Dans la classe 2, le thème émergent est celui du *territoire* avec un vocabulaire appartenant au domaine de la géographie ou de l'histoire et avec, aussi, la présence du thème de la *migration*. Il y a ici des références à la charte européenne des langues (*déclaration*, *article*, *rapport explicatif*). Les définitions de *« langues dépourvues de territoire »* (*rromani*) ou de *« langues de migrants »* apparaissent.

La classe 4 semble mettre en relief la question de la défense de la francophonie par rapport à la domination de l'anglais. Les verbes: *défendre*, *menacer*, *réclamer*, *protéger* caractérisent ce

champ lexical. On rencontre ici le discours développé par L.-J. Calvet à partir d'une hiérarchie : *langues centrales*, *supercentrales*, *hypercentrales*. Le terme *petit*, qui apparaît avec un Khi-2 élevé, peut se référer à *« petite langue »*.

La classe 1 contient des termes caractéristiques de la situation de l'Espagne et en particulier des Îles Baléares. On est ici dans une approche principalement descriptive (*enseignement*, *loi*, *télévision*, *famille*) avec des références aux notions de *« langue propre »* et de *« variété »*. L'apparition de cette thématique peut être représentative des échos suscités par le développement, au cours des dernières décennies en Espagne, d'aménagements linguistiques en faveur de langues propres à ce pays autres que le castillan.

Les deux dernières classes, 3 et 5, sont les plus volumineuses et sont assez proches (d'après l'indication de spécificité donnée par le graphique et la présence de valeurs de Khi-2 moins élevées). La classe 3 est construite autour de termes politiques (*dominer*, *peuple*, *conflit*) mais également illustrée par la notion de *diglossie*, contextuelle par rapport à celles qui nous occupent en première analyse ici.

Enfin, la classe 5 est définie autour de la notion de *dialecte* (mais avec également la présence des termes : *parler* et *patois*). Elle s'applique surtout ici aux situations du français et de l'allemand (Luxembourg, France) avec des références politiques génériques (*État*, *nation*, *culture*).

On notera que les deux lemmes les plus fréquents du corpus (voir ci-dessous), *langue* et *linguistique*, n'apparaissent pas dans la classification (ce ne sont pas des termes qui créent du contraste).

Un retour au texte et des observations qui seront détaillées dans la suite permettent d'associer la classe 1 aux textes de Sintas (cf. «langue propre» en Espagne), la classe 4 à certains textes de Calvet, et la classe 3 principalement à Boyer mais également à Calvet.

4. Analyse lexicale, reconnaissance des lexies

Alceste, à la différence d'autres logiciels, permet de reconnaître de façon automatique les mots-outils et inclut des fonctions de lemmatisation qui, toutefois, résultent d'une analyse morphologique et non syntaxique du texte : le logiciel produit des *lexèmes* à partir de différents dictionnaires et catégories grammaticales, modifiables par l'utilisateur. On peut donc rencontrer des situations d'ambiguïté, comme par exemple : dans le syntagme « la langue corse », c'est le verbe « corser » qui peut être reconnu.

Le logiciel TXM intègre au contraire un étiqueteur morphosyntaxique, *TreeTagger*, dont l'utilisation a été privilégiée pour l'approche lexicale du corpus.

Selon TXM, le corpus CLME est caractérisé par un nombre total d'occurrences de 22671, par 4189 formes graphiques, dont 3126 après lemmatisation. Il s'agit d'un volume relativement faible mais, étant donné le caractère homogène et la précision apportée dans le formatage du texte, un travail d'interprétation peut être entrepris.

Un examen des lemmes les plus fréquents du corpus, après élimination de quelques mots-outils et en tenant compte de l'étiquette morphosyntaxique (voir tableau 3), met en évidence :

- deux lemmes principaux : *langue* et *linguistique*, auxquels on peut ajouter deux autres lemmes dérivant du même radical : *sociolinguistique* et *plurilinguisme*,

- les autres noms par ordre de fréquence (*territoire, dialecte, groupe, cas, locuteur, pays*) peuvent servir à résumer une vue très schématique du corpus ; ceci est dû au fait que les extraits sont ciblés autour d'un sujet commun,
- parmi les lemmes les plus fréquents, on peut remarquer une prédominance des adjectifs bien qu'une comparaison n'ait pas été faite par rapport à un corpus de référence : ceci est dû au caractère à la fois descriptif et comparatif de ces textes.

langue_NOM	804	groupe_NOM	57	dominant_ADJ	34
linguistique_ADJ	114	cas_NOM	47	situation_NOM	33
français_ADJ	79	national_ADJ	46	dominer_VER:pper	32
régional_ADJ	77	officiel_ADJ	44	politique_ADJ	30
minoritaire_ADJ	75	locuteur_NOM	39	maternel_ADJ	30
territoire_NOM	68	France_NAM	38	parler_NOM	28
Etat_NOM	66	social_ADJ	37	variété_NOM	27
dialecte_NOM	64	pays_NOM	35	exemple_NOM	26
pouvoir_VER:pres	62	rapport_NOM	34	discours_NOM	26

Tableau 3 : mots-pleins les plus fréquents

Afin de localiser les notions identifiées initialement, les collocations autour du lemme "langue" ont été recherchées au moyen du processeur de requêtes intégré dans TXM. Les résultats fournis ont été fusionnés et épurés des segments non significatifs pour notre problématique (*« autre langue »*, *« langue berbère »*, *« langue différente »*, etc.) et sont fournis par le tableau 4.

langue régional	35	langue commun	6	langue régional et minoritaire	3
langue minoritaire	34	langue de origine	6	langue mixte	3
langue officiel	32	petit langue	5	langue d'immigration	3
langue dominer	26	langue de diaspora	5	langue véhiculaire	3
langue maternel	26	langue de communication	5	langue grégaire	3
langue dominant	26	langue périphérique	5	langue commun ou véhiculaire	2
langue régional ou minoritaire	25	langue identitaire	5	langue spécifique	2
langue national	23	langue en diaspora	5	langue de population	2
langue local	15	langue autochtone	4	langue traditionnel	2
langue de Etat	13	langue minoré	4	langue historique	2
langue propre	12	langue de immigration	3	langue immigrer	2
langue de migrant	7	langue de le Etat	3	langue menacer	2
langue dépourvu	11			langue hériter	2

Tableau 4 : collocations autour de « langue »

Ces résultats ont été fusionnés avec d'autres lemmes *simples* (par opposition aux *polyformes*) afin de constituer un tableau de *lexies-notions* (ici non lemmatisées) qui sera utilisé dans la suite (voir tableau 5). Un seuil de fréquence minimum de 5 a été établi afin d'appliquer ce tableau à des analyses factorielles.

dialecte	64	langue dominante	26	langue dépourvue de territoire	7
langue régionale	35	langue régionale ou minoritaire	25	langue commune	6
patois	35	langue nationale	23	langue d'origine	6
langue minoritaire	34	langue locale	15	langue de communication	5
langue officielle	32	langue d'Etat	13	langue de diaspora	5
parler	28	langue propre	12	langue en diaspora	5
langue dominée	26	langue de migrants	7	langue identitaire	5
langue maternelle	26			petite langue	5

Tableau 5 : lexies-notions

5. Partition du corpus et spécificités des auteurs

TXM permet de procéder principalement suivant une approche de partition sur le corpus. En partitionnant le corpus par la variable *auteur*, on peut effectuer un calcul de *spécificités* (basé sur une application de la loi hypergéométrique), et donc en particulier déterminer les lemmes qui sont en surnombre ou en sous-effectif selon chaque auteur.

Les résultats de cette analyse (en utilisant un seuil minimum de 4) sont résumés dans le tableau 6. A noter que l'on n'a pas conservé les termes plus fréquents (jusqu'à un seuil de 100, y compris: *langue*, *linguistique*), même s'ils ont été indiqués par le logiciel, étant donné qu'ils ne semblent pas apporter une information significative.

AIROLDI	entreprise, prestige
AKIN	territoire, dépourvu, immigration
BLAIR	aire, territoire
BOYER	sociolinguistique, diglossique, conflit, représentation, résistance, dominer, occitan
CALVET-article	droit, défense, France, francophone, périphérique, défendre
CALVET-ouvrage	droit, exclusif, local, dialecte, colonial, pays
COMMISSION	groupe, Etat, social, société, minoritaire
COUROUAU	dénomination, patois
DUCROT-TODOROV	parler, apparenter, patois
HERDAM	allemand, primaire
KOULAYAN	diaspora, maternel, origine
MARCELLESI	régional, référer, classe, corse
ROLLAN	communication, commun
SINTAS	propre, variété, castillan, île, catalan, standard
VIAUT	historique, territoire, Charte, article, historicité, déclaration, migrant
WOEHRLING	régional, officiel, territoire

Tableau 6 : spécificités lexicales

Le tableau des spécificités a été épuré des termes exclusifs (présents chez un seul auteur), qui sont présentés par le tableau 7. Les lemmes dont la fréquence minimale est de 3 ont été conservés ici. Ce tableau fournit quelques indications supplémentaires sur les aires géographiques de spécialisation des auteurs.

AKIN	kurde
BLAIR	sâme
BOYER	résistance, faveur, partager, inscrire
CALVET-article	PLC, central, réclamer, galicien, ratification, peur
CALVET-ouvrage	coloniser, colonisation, choisir, superstructure, mépris, libération, baptiser, swahili, endogène, véhicule, souligner, oppression, nécessité, graphie, exogène, débarrasser
COMMISSION	reproduction, irlandais, irlandais, individu, concentrer
DRETTAS	Grecs, Russie, déplacement, diasporéique
HERDAM	typique, ethnolecte, kanak
KOULAYAN	diasporique, natif
MARCELLESI	hégémonique, étendue, ressort, oïl
SINTAS	insulaire, Baléares, film, CAIB, îlien, éducation, sous-titrer
VIAUT	lien, relier, LRM[3]
WOEHRLING	coût

Tableau 7 : présences exclusives

6. Nouvelle partition et analyses des correspondances

Afin de procéder à des analyses des correspondances, la partition sur la variable *auteur* a été utilisée mais en laissant de côté un certain nombre d'auteurs dont le volume de texte représenté n'était pas considéré comme étant significatif (il s'agit de : Airoldi, Bidart, Bruneau, Giblin, Jetchev, Rollan).

Une première analyse factorielle des correspondances a été effectuée sur le vocabulaire suivant :

> *{ aire, autochtone, communauté, communication, conflit, défense, dialecte, diaspora, diglossie, dominer, droit, géographique, historique, identité, immigrer, local, maternel, migrant, minoritaire, minorité, national, origine, patois, plurilinguisme, politique, régional, social, territorial, traditionnel, variété }*

Préalablement, une fusion de lemmes dérivés du même radical a été effectuée (exemple : *diaspora*, *diasporique* et *diasporéique*). Ce vocabulaire a été déterminé en fonction des résultats obtenus par les opérations décrites précédemment, en particulier les analyses de spécificités, et en privilégiant des termes pertinents pour notre problématique.

Les cinq premiers facteurs, qui expliquent 70 % de l'inertie totale, ont été étudiés. La projection des points-colonnes (c'est-à-dire des auteurs) sur le plan formé par les deux premiers axes est représentée par la Figure 2.

Les plus fortes contributions de points-lignes relevées sur l'axe 1 opposent, d'une part, en positif : *diaspora*, *origine*, *maternel*, *communauté*, et, d'autre part, en négatif : *conflit, diglossie, patois, dominer, communication, local*. Cet axe définit deux types de motivations dans les écrits : l'une à partir des questions d'origine et d'identité (ce sont les « langues de migrants », envisagées au sens large en incluant les « langues dépourvues de territoire »), l'autre à partir des problèmes de conflit (entre « langues officielles » et « langues locales »). Cet axe oppose principalement KOULAYAN (dans une moindre mesure : DRETTAS et HERDAM) à BOYER (avec CALVET-ouvrage et aussi COUROUAU).

4 Sigle pour «langues régionales ou minoritaires».

Sur l'axe 2, on trouve les termes : *minoritaire*, *régional*, *aire*, *géographique, traditionnel, territorial, historique* (en négatif) qui s'opposent à la plupart des termes cités pour l'axe 1. Il est possible d'associer cet axe au concept de « lien au territoire ». Cet axe est construit principalement par BLAIR, WOEHRLING, VIAUT et GUILLOREL. Il y a également une contribution de SINTAS mais dans une direction opposée.

L'axe 3 est associé très nettement au terme *variété,* qui s'oppose à la plupart des termes précédents, en particulier à *diaspora*. On considèrera qu'il représente une approche descriptive, utilisée surtout par SINTAS, mais également par HERDAM, par opposition à une approche conceptuelle.

L'axe 4 décrit une composante politique avec, également, les termes *défense, droit, plurilinguisme* : il est associé à CALVET-article.

L'axe 5 est construit par les termes : *traditionnel, historique.* Il est associé à VIAUT (en opposition à COMMISSION et BLAIR).

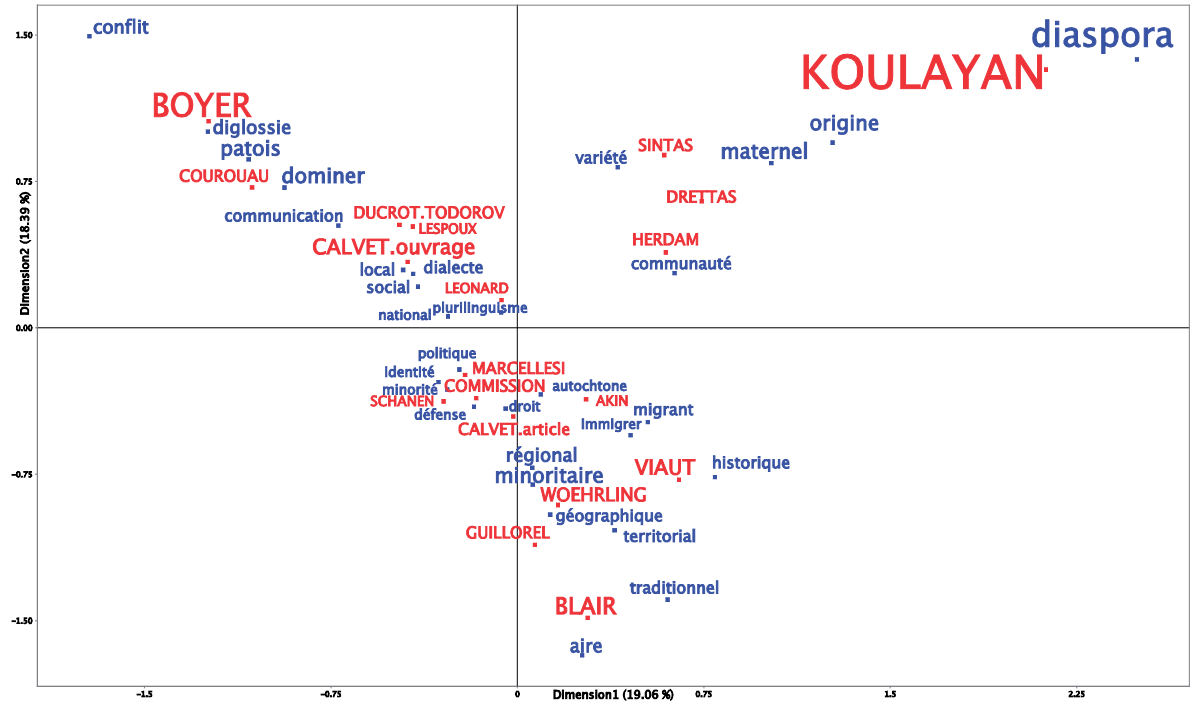


Figure 2 : AFC sur vocabulaire lemmatisé

7. Tentative de catégorisation

A partir de la même partition en seize parties, on effectue une autre analyse des correspondances, mais cette fois-ci sur les lexies-notions définies plus haut. La projection sur le plan formé par les deux premiers axes est représentée par la Figure 3 qui explique 37 % de l'inertie[5]. L'étude

5 Les deux notions : « langue de diaspora » et « langue en diaspora » ont été analysées en tant qu'éléments (lignes) supplémentaires.

des contributions permet de reconnaître quatre zones dans lesquelles sont agrégées les notions suivantes:

- *langue propre, langue d'Etat, langue minoritaire*
- *langue régionale ou minoritaire, langue dépourvue de territoire, langue de migrants*
- *langue de diaspora, langue d'origine, langue maternelle*
- *patois, dialecte, parler, langue dominante, langue dominée*

Figure 3 : AFC sur lexies-notions

Compte tenu des plans définis par les cinq premiers axes de l'analyse des correspondances et à l'aide d'un retour au texte par une recherche de concordances, on peut regrouper les auteurs en quatre classes associées à un certain nombre de lexies :

- les « *territorialistes* » : SINTAS, COMMISSION, VIAUT (*langue propre, langue* d'Etat)

  On reconnaît dans cette classe des notions traitées par la charte et une application à la situation espagnole, le lien est représenté par le territoire.

- les « *identitaires* » : DRETTAS, KOULAYAN, HERDAM, AKIN (*langue d'origine, langue maternelle, langue dépourvue de territoire*)

  C'est l'ensemble des situations « expatriées » (grecs, arméniens, kurdes) autour des questions d'identité.

- les « *politistes* » : CALVET-ouvrage, BOYER, DUCROT (*langue nationale, langue dominante, langue dominée, langue locale*)

  Il s'agit ici de la question du statut et du rapport entre langues locales et langues dominantes, en particulier dans des situations de plurilinguisme.

- les « *historiques* » : BLAIR, MARCELLESI, WOEHRLING, GUILLOREL (*langue traditionnelle, langue commune, langue régionale, langue de migrants*)

  On retrouve ici des notions traitées par la charte, le lien est représenté par l'histoire et la tradition.

8. Conclusion

L'utilisation combinée de deux outils logiciels permet d'obtenir des visions complémentaires sur un corpus. Une procédure de classification automatique suggère des dimensions sémantiques latentes. Celles-ci sont progressivement confirmées par une présentation paradigmatique du lexique défini par le corpus. Dans notre cas, il apparaît que des rapprochements s'opèrent en fonction des dates de publication (hypothèse à vérifier pour les textes précédents ou postérieurs à la Charte européenne des langues) et du genre (monographies, articles).

Références

Bassac C., Busquets J., Versel M. (2009). Analyse statistique des données textuelles à partir de publications de Calvet concernant les langues minoritaires. *Lengas, revue de sociolinguistique*, n° 66, pp. 57-78.

Bolasco S. (2005). Statistica testuale e text mining : alcuni paradigmi applicativi, *Quaderni di Statistica*, Vol. 7, pp. 17-53.

Daoust F., Dobrowolski G., Dufresne M., Gélinas-Chebat C. (2006). Analyse exploratoire d'entrevues de groupe : quand ALCESTE, DTM, LEXICO et SATO se donnent la main, in : Viprey, J.M., Condé C.C., Lelu, A., Silberztein, M. (éds.), *JADT 2006, Actes des 8es Journées internationales d'analyse statistique des données textuelles*, Besançon, Presses Universitaires de Franche-Comté, pp. 313-326.

Lebart L., Salem A. (1988). *Analyse statistique des données textuelles*, Paris, Dunod, Paris.

Reinert M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours, *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, 12-14 mars 2008, Lyon : Presses Universitaires de Lyon, pp. 981-993.