In Situ
Revue des patrimoines

15 | 2011
Le patrimoine des guides : lectures de l'espace urbain européen

Vers une cartographie géo-lexicale

William Martinez

Édition électronique
URL : http://journals.openedition.org/insitu/590
DOI : 10.4000/insitu.590
ISSN : 1630-7305

Éditeur
Ministère de la culture

Référence électronique
William Martinez, « Vers une cartographie géo-lexicale », *In Situ* [En ligne], 15 | 2011, mis en ligne le 29 juin 2011, consulté le 30 avril 2019. URL : http://journals.openedition.org/insitu/590 ; DOI : 10.4000/insitu.590

Vers une cartographie géo-lexicale

William Martinez

Introduction

1 Quel vocabulaire emploie-t-on pour décrire une ville dans les guides de tourisme du XIX^e^ et XX^e^ siècles[1] ? À partir de cette question, nous avons entrepris de déterminer par le biais de la statistique textuelle les formes lexicales élémentaires qui constituent le vocabulaire courant pour la représentation de l'espace urbain dans les ouvrages touristiques. Nous présenterons ici les résultats obtenus à partir de deux guides de la France : le Guide Diamant de 1873 et le Guide Bleu de 1956, tous deux publiés chez Hachette.

2 Le traitement objectif et exhaustif qui caractérise la lexicométrie permet d'identifier les particularités lexicales saillantes de ces textes afin d'esquisser des cartes reliant les notions fondamentales qui y sont véhiculées et d'en étudier l'évolution chronologique (partie 1). Plus particulièrement, l'analyse des cooccurrences lexicales détermine des systèmes de mots associés qui construisent localement le sens dans chaque passage descriptif des guides et pour certains élaborent des réseaux structurant de manière globale les textes (partie 2). Au fil de ces expériences, la cartographique récurrente - plans factoriels et réseaux de cooccurrences - nous amènera à envisager une fusion des données afin d'associer les structures lexicales mises en évidence par la statistique avec le géo-référencement des Systèmes d'Information Géographique en vue de produire des cartes descriptives plus riches (partie 3).

Approche lexicométrique des guides de tourisme

3 Après la numérisation des textes d'origine qui les transforme de documents papier en fichiers informatiques, on obtient une base textuelle numérisée adaptée au dépouillement statistique des données lexicales. À ce corpus numérisé on peut appliquer une batterie de méthodes lexicométriques pour mettre en évidence l'évolution lexicale au fil des textes ainsi que la structuration de chacun des ouvrages[2].

Caractéristiques quantitatives du corpus des guides

4 À l'issue de la numérisation des guides un certain nombre de modifications du texte sont nécessaires. Parmi celles-ci on mentionnera la correction des scories de la reconnaissance optique, l'uniformisation de la casse des lettres (pour éviter la dilution fréquentielle des mots, toutes les majuscules sont transformées - ex. : *Le* devient *le*) et l'introduction de balises méta-textuelles qui délimitent les différentes descriptions de villes, routes, etc. On obtient alors un corpus double dont voici les caractéristiques quantitatives :

Tableau 1 : Caractéristiques quantitatives des corpus

	Hachette 1873	**Hachette 1956**
Nb. d'occurrences	287 823	603 850
Nb. de formes	32 725	35 600
Nb. de hapax (1 seule occ.)	21 564	19 722
Nb. de phrases	56 431	85 992

5 Sur le plan quantitatif on observe d'une part une différence majeure dans le volume des deux corpus puisque le guide de 1956 compte plus du double de mots employés dans celui de 1873, et d'autre part, des similitudes entre le nombre de formes (les mots distincts) et de *hapax legomena* (les mots employés une seule fois en contexte) durant les deux périodes [3].

6 La première phase de l'analyse lexicométrique consiste à identifier les unités lexicales employées en contexte. À l'issue de cette *segmentation*, on extrait un dictionnaire de fréquence pour chaque corpus dont les têtes de liste sont présentées ci-après (tableau 2) : 35 formes de fréquence supérieure ou égale à 10 pour le guide de 1873 et 51 formes pour le texte de 1956. À cette étape initiale on peut déjà tracer des parallèles entre les deux guides. On constate en effet des ressemblances pour ce qui concerne les mots les plus employés dans chaque texte et l'on observe notamment, malgré les divergences orthographiques qui compliquent la comparaison entre les textes (ex. : *kil* devient *k*), la persistance d'un axe lexical *on-route-pont-église-saint-château*. Ce dernier tend à montrer une continuité dans le mode de présentation de l'observateur, de l'itinéraire et des points d'intérêt touristique, autant de points d'ancrage lexicaux que nous allons étudier plus en détail.

Tableau 2 : Dictionnaire des fréquences pour chaque guide (extrait)

Hachette 1873		Hachette 1956	
Forme	Fréq.	Forme	Fréq.
de	18 855	de	36 405
le	9 746	la	27 985
la	7 570	du	14 465
et	5 935	le	12 478
du	5 727	à	12 256
à	5 398	et	11 940
s	4 565	s	11 521
d	3 968	l	10 014
l	3 940	les	9 949
les	3 843	d	7 908
a	3 293	on	6 875
h (heures)	3 202	des	6 808
r (route)	2 780	au	6 039
sur	2 715	en	5 844
en	2 639	a	5 476
des	2 629	par	5 174
par	2 618	sur	4 942
saint	2 479	ème	4 698
on	2 142	les	4 510
route	2 138	n	4 473
met	2 004	reste	4 270
au	1 836	un	4 247
église	1 730	une	3 808
les	1 630	saint	3 381
château	1 512	m	3 313
un	1 508	église	3 009
dans	1 438	au	2 956
c (classé)	1 413	est	2 696
mon (monument)	1 392	dans	2 626
hist (historique)	1 294	g (gauche)	2 424
fr (francs)	1 260	dr (droite)	2 402
une	1 167	ou	2 336
min (minutes)	1 091	hôt (hôtel)	2 194
pont	1 038	hab (habitants)	2 168
cl (classé)	1 037	ville	1 835
		vallée	1 677
		xv	1 667
		rue	1 647
		château	1 570
		xvi	1 548
		avec	1 546
		h (heure)	1 514
		e (Est)	1 473
		o (Ouest)	1 414
		pour	1 255
		place	1 190
		plus	1 188
		rive	1 162
		hôtel	1 154
		pont	1 069
		i (information)	1 004

Typologies lexicales

7 Afin de dépasser le stade de la statistique descriptive et déterminer s'il existe un vocabulaire récurrent dans les guides qui est employé pour la description des milieux urbains, il faut mobiliser une méthode statistique multidimensionnelle plus à même de prendre en compte tous les éléments lexicaux et de les étudier dans l'ensemble des deux corpus. Pour ce faire nous exploiterons l'Analyse Factorielle des Correspondances (AFC), une méthode d'analyse qui synthétise les liens d'attraction et de répulsion entre les formes et les parties du corpus et détermine quelles sont les régularités et les ruptures dans la structuration des guides.

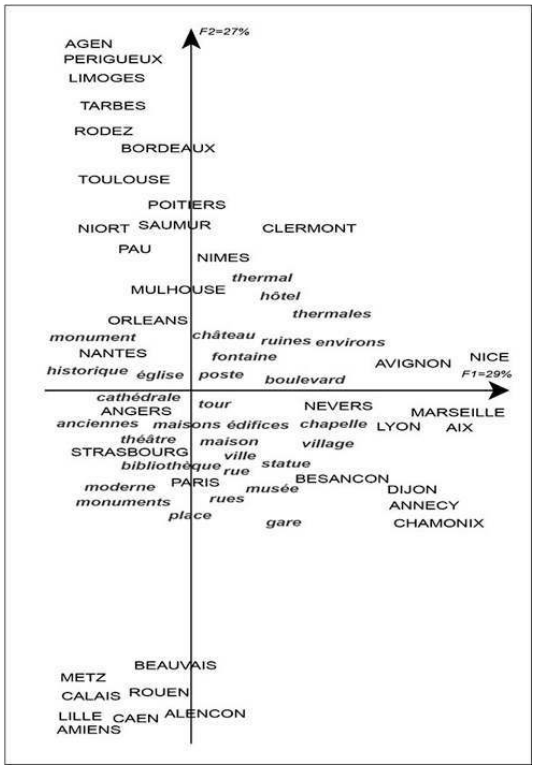
8 Le principe de l'AFC consiste à décomposer les données de répartition du lexique dans les différentes parties du texte en une somme de facteurs successifs et représentables deux à deux sur un plan graphique. L'AFC dégage ainsi une carte qui résume les faits les plus saillants du corpus qu'il s'agit d'interpréter visuellement en observant la disposition des points pour y repérer une continuité et d'éventuels points de rupture[4].

9 Les deux analyses qui suivent sont effectuées sur la base d'un découpage des corpus en routes, c'est-à-dire des itinéraires suggérés par les auteurs des textes qui associent plusieurs villes et régions dans le cadre d'un parcours proposé au lecteur. Ces routes partent des grandes métropoles (Paris, Lyon, etc.) et sont approximativement une trentaine par guide, soit un nombre de divisions du texte adéquat pour une analyse factorielle[5]. Si l'analyse porte sur l'intégralité des corpus et prend en compte l'ensemble du texte, pour les besoins de l'affichage nous ne retiendrons que les villes les plus souvent mentionnées dans les guides (37 villes évoquées au moins à 50 reprises dans les itinéraires) et les formes lexicales qui nous semblent relatives à l'espace urbain.

10 L'interprétation du plan factoriel suit une règle de base qui pose que la proximité graphique des points correspond à une proximité lexicale. Ainsi, deux villes sont proches sur le plan parce qu'on emploie un vocabulaire semblable pour les décrire, deux termes sont proches parce qu'on les mentionne souvent dans les mêmes contextes et, à l'inverse, villes et mots sont opposés sur le plan parce qu'on ne les combine jamais dans les descriptions touristiques. Cette règle d'interprétation énoncée, soulignons également que l'agencement des villes sur ces schémas repose exclusivement sur les similitudes lexicales dans leurs descriptifs respectifs sans aucune autre information de type géographique.

11 Pour ce qui concerne la figure extraite du guide de 1873[6] **(fig. n°1)**, on constate une concentration des mots décrivant l'espace urbain au centre du graphique, c'est-à-dire dans une zone signalant le vocabulaire commun (soient les mots qui sont équitablement répartis au fil du texte). Également près du centre du plan factoriel, se place la capitale française et plusieurs grandes villes dont la description dans les guides invoque ce vocabulaire commun : *ville*, *rue*(*s*), *place*, *maison*(*s*), *église*, *théâtre*, *musée*, *gare*, etc. Loin du centre et aux extrémités du graphique, on trouve des villes qui se distinguent de la masse lexicale commune et s'agglomèrent dans des groupes dont la cohérence géographique est assez nette : villes du Nord (*Lille*, *Amiens*, *Rouen*, *Calais*), villes du Sud (*Nice*, *Marseille*, *Aix, Avignon*) et villes du Sud/Sud-Ouest (*Tarbes*, *Rodez*, *Agen*, *Périgueux*). De fait, pour ces ensembles géographiques on vérifiera par une seconde analyse réalisée à des seuils de spécificité plus bas (afin de relever même les phénomènes lexicaux de moindre importance statistique) qu'aucun vocabulaire caractéristique, particulier, original ne se dégage dans les descriptifs de ces villes. Leur agencement sur le plan factoriel repose donc sur une logique statistique qui met en évidence leur répulsion par rapport au vocabulaire commun. Par exemple, on ne décrit jamais les villes du Nord en termes de *thermal*(*es*), *hôtel*, *monument*, *château*, *ruines*, etc. pas plus que l'on n'évoquera *monument*(*s*), *cathédrales*, *église* pour dépeindre Nice ou Marseille.

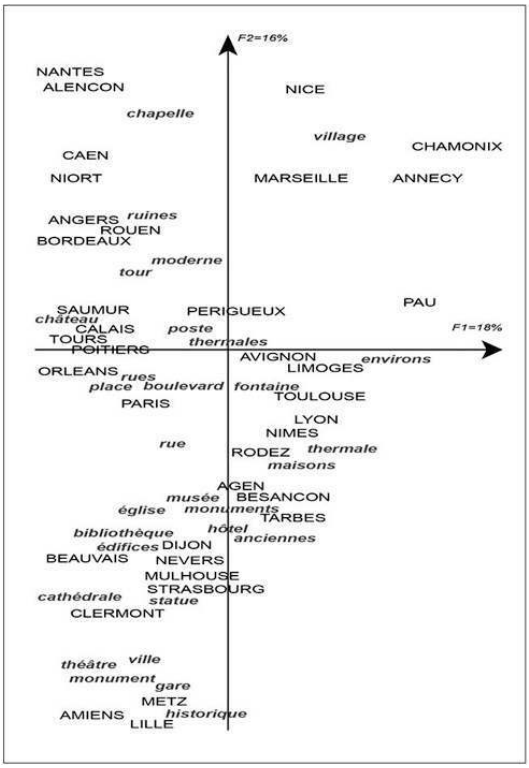
Figure 1



Plan factoriel – Guide 1873 – Lexique et villes les plus mentionnées.

12 Quand on compare ces résultats à ceux que l'on obtient pour le guide de 1956, la différence est d'emblée visible dans l'organisation générale du graphique. En effet, dans le second plan factoriel[7] **(fig. n°2)** l'ensemble des points suit un mouvement de déconcentration : le centre géométrique du graphique n'est plus son centre de gravité. Et, même si quelques formes qui sont presque incontournables dans les descriptifs des villes (*poste*, *rue*, *place*, *boulevard* et *fontaine*) occupent le centre du plan, on observe un effet de dilution du nuage initial. Ainsi, en 1956 tout un lexique culturel des guides de tourisme est dédié aux villes du Sud/Sud-Ouest et du Nord : *musée*, *monuments*, *bibliothèque*, *statue*, *théâtre*, *cathédrale*, *église*, etc. De même, les formes *hôtel* et *gare* sont plus souvent employées pour décrire ces villes excentrées dont on souhaite souligner les récentes infrastructures touristiques, facilités de transport et hébergement.

Figure 2



Plan factoriel – Guide 1956 – Lexique et villes les plus mentionnées.

13 L'Analyse Factorielle révèle donc que la différence structurelle principale entre les deux guides tient au fait que des ensembles géographiques existent qui ne sont pas décrits par le vocabulaire typique de l'espace urbain. La comparaison montre aussi qu'au fil des 83 ans séparant les guides de 1873 et 1956 un lexique caractéristique a été introduit pour qualifier ces villes différentes. On constate par exemple pour l'année 1956 que le groupe excentré formé autour d'*Alençon*, *Caen* et *Rouen* est qualifié par *chapelle* et *ruines* tandis que la zone *Nice*, *Marseille*, *Chamonix* et *Annecy* se démarque par de nombreuses évocations de *village*. Ces formes peu nombreuses ne suffisent pas à définir entièrement les univers lexicaux que les guides attachent à ces villes mais elles montrent clairement une évolution dans les choix de description touristique, culturelle et historique : on est passé d'une caractérisation lexicale négative (ces villes se distinguent en 1873 par l'absence de vocabulaire qualificatif) à une caractérisation positive (on dénombre en 1956 des formes nominales, adjectivales et autres employées spécifiquement pour parler de ces villes).

Les réseaux de cooccurrence lexicale

14 À la base de toute méthode de cooccurrence, il y a une hypothèse sur une distribution inégale des formes en contexte. À la manière du lecteur qui au fil de la lecture du texte perçoit des combinaisons lexicales récurrentes qu'il interprète comme des associations privilégiées, la statistique syntagmatique que nous appliquerons ici va comptabiliser minutieusement dans le voisinage d'un mot-pôle chaque occurrence lexicale et, au terme de cette lecture systématique, comparer les fréquences des mots en jeu pour déterminer quelles coïncidences lexicales sont statistiquement remarquables.

15 Les *concordances* suivantes – qui constituent une forme primaire de méthode de cooccurrence – permettent d'apprécier quelques environnements contextuels de la forme *ville* et organisent ces extraits de manière à ce que le lecteur puisse y détecter les affinités autour du mot-pôle.

16 ..que le lecteur qui veut aller d'une **ville** à une autre y trouve le meilleu
avant la description intérieure de la **ville**. hôtels. les hôtels, au choix e
de paris à versailles par suresnes et **ville**-d'avray, 539. de versailles et
collections de peinture moderne de la **ville** de paris, tandis que celle de l
s de saint-gervais s'élève l'hôtel de **ville**, vaste édifice de style renaiss
lignes de collines comprises dans la **ville**. celle de la rive dr., la plus
capitale du ponthieu, est une vieille **ville** de 19502 hab., située sur les d..

17 Sur la base de ces contextes d'apparition, la méthode de cooccurrence introduit un calcul probabiliste afin de hiérarchiser les affinités lexicales les plus inattendues sur le plan statistique. On identifie ainsi les *cooccurrences empiriques* du pôle, c'est-à-dire des « coïncidences de mots dans les mêmes contextes ». Cette définition présente l'avantage de faire abstraction de toute considération linguistique et n'impose ni orientation, ni contiguïté, ni distance minimale, ce qui rapporte un plus grand nombre de phénomènes de cooccurrence.

18 Nous appliquerons ce principe d'analyse dans trois méthodes variantes - les *cooccurrences spécifiques*, les *poly-cooccurrences* et les *trames de cooccurrence*.

Les cooccurrences spécifiques

19 La méthode des *cooccurrences spécifiques* détermine par la statistique les attractions lexicales remarquables autour d'une forme-pôle donnée[8] dans une fenêtre d'exploration contextuelle définie[9]. Le tableau ci-après présente les résultats de cette statistique des rencontres autour de la forme *ville* au cours de ses 519 apparitions dans le guide de 1873.

Tableau 3 : Cooccurrences spécifiques du pôle ville (Hachette 1873) (Ctx Phrase ; CFMin 5 ; SpMin 5)

Forme	Fréq.	Co-fréq.	Spéc.	Nb. ctx.
hôtel	364	230	+50	226
basse	35	18	+20	18
ancien	338	35	+12	35
haute	102	18	+11	18
vieille	33	11	+10	11
divisée	20	9	+10	9
moderne	202	22	+9	22
musée	215	22	+8	22
bibliothèque	160	19	+8	19
bâtie	64	12	+8	12
faubourg	43	10	+8	10
1750	14	6	+7	6
construit	55	9	+6	9
nouvelle	31	7	+6	6
beffroi	29	7	+6	7
cité	19	6	+6	6
divise	16	6	+6	6
ancienne	272	19	+5	19
rue	168	14	+5	10
vol	126	12	+5	12
parties	30	6	+5	6
gallo-romaine	17	5	+5	5

Guide de lecture du tableau : L'analyse des cooccurrences révèle les principales attractions autour d'un pôle en comparant entre autres données, la fréquence globale de chaque cooccurrent (*Fréq.*) avec sa co-fréquence *(Co-fréq.)* dans les contextes qu'il partage avec le pôle *(Nb. ctx)* et fournit un indice de spécificité (*Spéc.*) signalant son sur-emploi (*+x*).

20 La méthode d'analyse des cooccurrences peut aussi être appliquée à plusieurs formes associées en un système de mots tel que celui formé par l'axe lexical relevé initialement : *on-route-pont-église-saint-château*. On vérifiera de cette manière que l'emploi de ces termes associés en contexte évolue dans le temps en relevant pour le guide 1873 26 attractions dont 11 verbes de déplacement contre 53 cooccurrences dont 21 verbes de déplacement en 1956. Cette inflation montre que l'exploration piétonne est de plus en plus encouragée dans le texte des guides.

En 1873 :	En 1956 :
franchit, *passe*, *traverse*, *croise*, *suit*, *descend*, *remonte*, *entrer*, *atteint*, *quitte* et *sort*	*traverse*, *franchit*, *prend*, *sort*, *passe*, *laisse*, *arrive*, *reprend*, et *remonter*

21 Avec la méthode des cooccurrences spécifiques il devient possible de réaliser des comparaisons inter-guides et de mesurer l'évolution chronologique du vocabulaire afin d'apporter des éléments de réponse à des interrogations telles que :

22 • Quels mots emploie-t-on le plus souvent pour décrire la ville de *Marseille* ?
- en 1873, Marseille est pour l'essentiel un port de commerce proposant dans ses alentours quelques attractions géographiques naturelles : *méditerranée*, *boulevard*, *marine*, *docks*, *port*, *rue*, *colline*, *grotte*, *romains*, *plaine*, *navires*, *golfe*, *superficie*, *îles*, *chantiers*, *maritime* et *navigation*.
- en 1956, une infrastructure touristique a été mise en place alors le guide reprend l'information précédente et la complète par les caractéristiques culturelles du port, de la ville, de la région, désormais mises en valeur : *canebière*, *corniche*, *port*, *chaîne*, *calanque*, *quai*, *canal*, *services*, *hôtellerie*, *tramway*, *florales*, *transatlantiques*, *navigateurs*, *renseignements*, *bassin*, *mistral*, *porcelaine*, *paquebots*, *mer*, *côte*, *boulevard*, *avenue*, *alpes*, *autocars*, *grotte*, *butte*, *club*, *allées*, *cultures*, *étages*, *bateaux*, *méditerranée*, *faïences*, *habitants*, *faculté*, *arides*, *escale*, *magasin*, *trolleybus*, *publics* et *desservie*.

23 • Quels mots emploie-t-on quand on évoque la *rue* ?
- le guide de 1873 présente les lieux d'intérêt touristique et historico-culturel accessibles depuis la *rue* : *place*, *musée*, *port*, *hôtel*, *jardin*, *statue*, *dimanche*, *ville*, *navires*, *quai*, *statues*, *rues*, *histoire* et *industrie*.
- le guide de 1956 reprend ces indications en les complétant par de nombreux verbes incitant le touriste au déplacement en lui indiquant des itinéraires précis : *place*, *hôtel*, *maisons*, *maison*, *nº*, *conduit*, *aboutit*, *ouvre*, *anciennes*, *musée*, *sort*, *saint*, *prolonge*, *prend*, *grande*, *cathédrale*, *bordée*, *angle*, *artère*, *natale*, *courte*, *république*, *arrive*, *ville*, *trouve*, *principale*, *face*, *revient*, *ramène*, *porte*, *devant*, *vieille*, *horloge*, *relie*, *lycée* et *mène*.

Les poly-cooccurrences

24 Pour dépasser les limites interprétatives des cooccurrences spécifiques qui ne restituent que les attractions entre deux mots, et dans un effort de reconstruction du sens, une extension de la méthode a été imaginée qui explore les phénomènes d'attirance simultanée entre plusieurs formes. Ces systèmes d'attraction complexes sont identifiés par le calcul réitéré des cooccurrences qui révèle des systèmes lexicaux à l'œuvre dans le texte.

25 Prenons l'exemple de la forme *place* qui apparaît fréquemment dans les deux ouvrages. En examinant les systèmes d'attraction lexicale au-delà des paires de cooccurrence, on découvre des schémas descriptifs complexes et récurrents qui caractérisent chacun des guides :

En 1873 :	En 1956 :
place + statue + bronze	place + rue + ville + république + hôtel
place + statue + fontaine + marché	place + rue + ville + centre
place + statue + napoléon	place + rue + hôtel + aboutit
place + statue + général	place + rue + on + revient
place + rue + hôtel	place + rue + conduit + directement
place + belle + fontaine	place + rue + maisons + anciennes
(...)	(...)

26 Ces poly-cooccurrences confirment dans le guide de 1956 une invitation à la promenade et à la découverte de la ville avec de nombreuses suggestions d'itinéraires. Aussi, étant donné le mode de calcul des poly-cooccurrences, leur réalisation en contexte est garantie et l'on peut à partir des squelettes de phrase déterminés par la statistique extraire un nombre exact de phrases où se concrétisent ces associations complexes. Considérons par exemple le cas de l'ensemble *place + rue + conduit + directement* pour lequel on identifie cinq contextes de réalisation :

27 <ville=strasbourg>
*là s'ouvre la **rue** mercière qui **conduit directement** à la **place** de la cathédrale : maisons anciennes dont la maison kam-merzell, au n° 1.*
<ville=besançon>
*la **rue** de lorraine **conduit directement** à la **place** monge, centre de la ville.*
<ville=nice>
*de la **place**, la **rue** notre-dame **conduit directement** à l'ancienne cathédrale sainte-marie, bel édifice à trois nefs, de 1604-1625.*
<ville=orléans>
*de la nouvelle **place**, la **rue** de la hallebarde **conduit directement** à la **place** du martroi*
<ville=niort>
*en face de notre-dame, la **rue** mautrec **conduit directement** à la **place** de la comédie.*

Les trames de cooccurrence

28 La troisième méthode de cooccurrence à laquelle nous aurons recours vise une exhaustivité totale de l'analyse. En effet, les cooccurrences et les poly-cooccurrences mises au jour par notre approche dépendent entièrement des pôles choisis au départ de l'analyse (*ville*, *place* et autres formes qui nous semblent appartenir au vocabulaire de l'espace urbain). Or, ces choix subjectifs conditionnent toute l'exploration contextuelle et orientent les résultats. Une autre approche a donc été envisagée qui prend pour point de départ non pas un mot prélevé dans le dictionnaire des formes mais un ensemble de mots, une plage pouvant comprendre plusieurs milliers de formes. L'analyse cooccurrentielle en devient pour autant exhaustive et rapporte tous les phénomènes d'attraction en contexte jusqu'à révéler la colonne vertébrale du texte[10].

29 Le premier résultat de l'analyse des trames est une hiérarchisation des formes lexicales suivant leur capacité d'attraction. La *valence lexicale* est un indice fondamental du magnétisme des formes dans leurs apparitions successives en contexte et signale les formes dont l'emploi dans le discours déclenche le recours systématique à d'autres mots qui ensemble élaborent des champs sémantiques récurrents. On constate à la lecture de la liste des termes les plus magnétiques dans chaque corpus (tableau 4) que celle-ci ne se confond pas avec la liste des formes les plus fréquentes (tableau 2). En d'autres termes, la capacité attractive d'un mot n'est pas corrélée au nombre de ses apparitions[11].

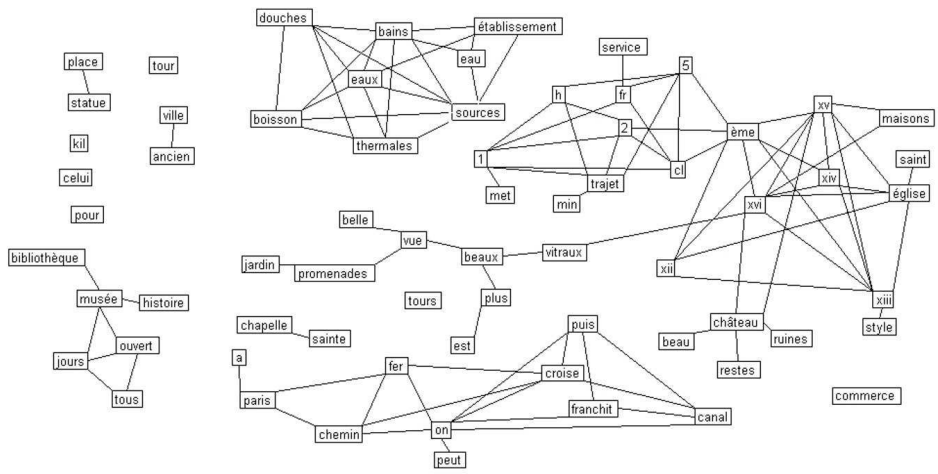
30 De cette analyse cooccurrentielle on dégagera un axe lexical *on-saint-r(rue)-église-place* qui est semblable à celui esquissé par l'analyse fréquentielle (1.1) mais qui introduit deux éléments centraux de l'espace urbain, la *rue* et la *place*, en lieu et place de *route*, *pont* et *château* signalant là peut-être un déplacement du tourisme champêtre vers la ville. Cet axe lexical constitue une colonne vertébrale qui traverse le texte de 1956 et provoque au fur et à mesure de ses occurrences l'apparition de nombreuses autres formes (adjectifs, etc.).

Tableau 4 : Valence lexicale des formes – Guides 1873 et 1956 (extrait)

Forme	Coocs	Forme	Coocs
saint	78	vallée	116
on	72	saint	113
r	70	on	105
kil	61	musée	103
musée	41	h	97
église	36	place	87
cl	36	rue	85
chapelle	32	sur	80
vitraux	31	r	80
h	29	église	74
ancien	29	plus	73
canal	29	station	72
tableaux	29	vue	69
commerce	29	hab (habitants)	67
statue	26	route	64
sources	25	renaissance	64
restes	24	place	61
bains	24	ville	60
chœur	24	plateau	59
paris	23	louis	58
eaux	23	port	57
place	22	pont	53
chemin	21	bd	52
tour	21	gothique	52
vue	21	franchit	51
château	20	plaine	50
ruines	20	balnéaire	50
maisons	20	château	48
bois	20	paris	47
établissement	20	ancienne	47

31 Le second résultat obtenu avec l'analyse des trames de cooccurrence est une visualisation de l'ensemble des attractions binaires en contexte sous la forme d'un graphe. À l'origine, le graphe complet compte plus de 1 800 mots reliés par plus de 4 000 liens, soit un objet mathématique parfaitement viable mais difficilement représentable (note nº8). On procède alors à un filtrage draconien qui ne conserve dans le réseau que les mots avec une valence lexicale très élevée (au moins 50 cooccurrences observées autour d'un mot-pôle). Le résultat[12] **(fig. n°3)** est un résumé représentatif de l'arbre initial qui réduit les cooccurrences à l'essentiel tout en reflétant la combinatoire des mots en contexte.

Figure 3

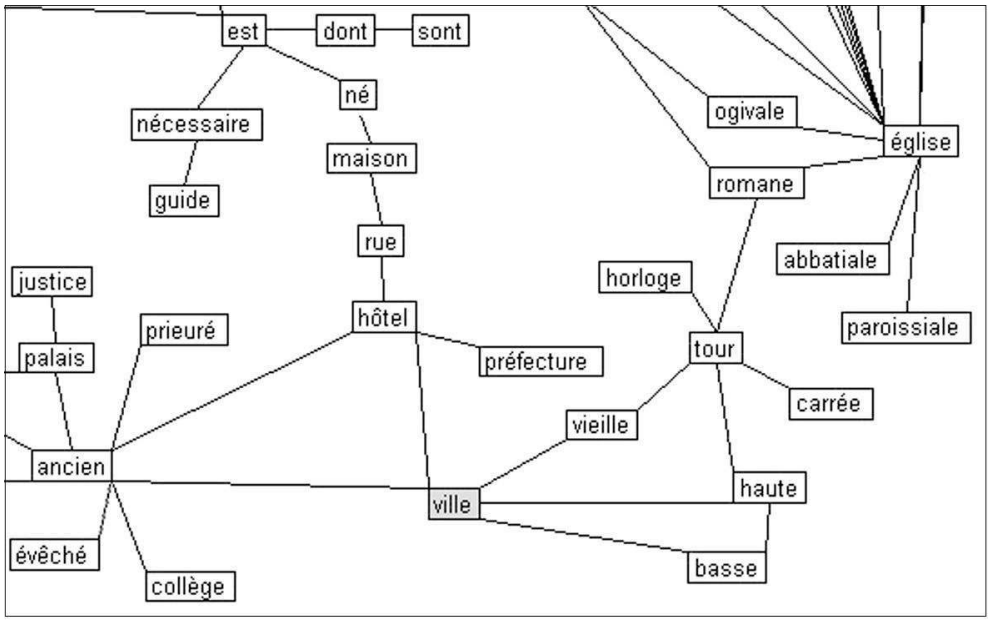


Réseau de cooccurrence (filtré par valence lexicale ≥50) – Guide 1873.

32 La systématicité de l'analyse rapporte le treillis fondamental tel qu'il se construit au fur et à mesure de l'énoncé et la visualisation hiérarchisée du graphe reproduit fidèlement la dynamique du corpus en distinguant les lexies prépondérantes qui subsument la diversité initiale du texte. En effet, ce que la cooccurrence lexicale met au jour est une toile de co-dépendances binaires et réciproques que tissent les composantes lexicales en vue de former un système lexical sous-jacent au texte entier et qui le structure de part en part.

33 Contrairement à la disposition linéaire du texte d'origine, l'organisation fragmentaire de la trame de cooccurrence propose des îlots de mots dont la lecture relève davantage d'une interprétation topologique des données. On s'intéressera donc sur cette carte aux noeuds qui forment des zones de densité, qui dominent leur milieu ambiant, et qui opposés l'un contre l'autre polarisent la carte : établissements de villégiature, lieux de culture, bâtiments religieux, etc. Ces agrégats qui subsistent dans le graphe après élagage statistique du réseau constituent les principales thématiques évoquées dans les guides[13].

Figure 4

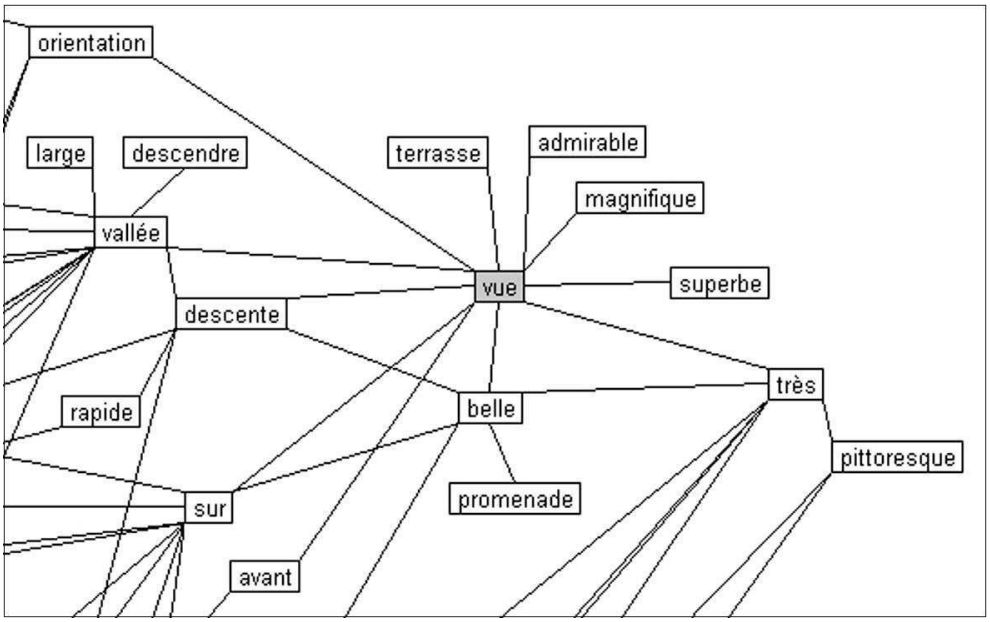


Réseau de cooccurrence (extrait autour de « ville ») – Guide 1873.

34 Si le graphe est une aide à la compréhension du territoire décrit dans les guides et s'il permet de formuler des hypothèses quant aux thématiques évoquées en contexte, il faut suivre certaines précautions pour son interprétation. Sur le plan linguistique, la lecture du réseau de cooccurrence peut s'effectuer, suivant les orbites autour d'un mot pôle, de deux manières illustrées par les schémas qui suivent :

35 - une lecture onomasiologique c'est-à-dire de l'idée vers les mots. Par exemple, par quels termes (noms, adjectifs, adverbes, verbes, etc.) exprime-t-on l'idée de *ville* ? **(fig. n°4)**
- une lecture sémasiologique c'est-à-dire des mots vers l'idée. Par exemple, quelle notion construisent ensemble les qualificatifs *admirable*, *magnifique*, *superbe*, (*très*) *pittoresque* et (*très*) *belle* ? Réponse : *vue* **(fig. n°5)**

Figure 5



Réseau de cooccurrence (extrait autour de « vue ») – Guide 1873.

36 On observera qu'en tant que visualisation du vocabulaire préféré par les auteurs des guides, le réseau de cooccurrence se rapproche de la *carte mentale* c'est-à-dire d'une méthode d'enquête sociologique qui tente de cartographier la représentation de l'espace qu'ont les individus par rapport à un territoire connu et pratiqué par eux.

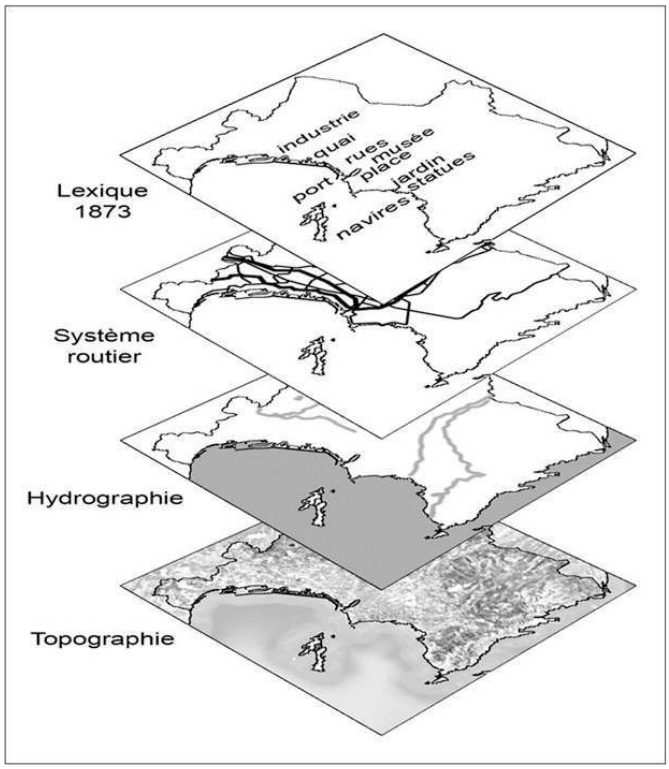
Fusion des données géographiques et lexicales

Les Systèmes d'Information Géographique

37 Définir l'existence d'une entité dans l'espace physique c'est bien sûr se référer à celle-ci par son toponyme. Mais, dans le cas d'un lieu sans dénomination exacte ou avec plusieurs appellations ou encore qui partage son nom avec d'autres endroits, l'identification géographique univoque se complique. Avec l'avènement de l'informatique on a donc évolué vers un vecteur de précision pour l'identification spatiale : un géo-référencement sous la forme d'un jeu de coordonnées qui distinguent sans ambiguïté un lieu précis – le SIG.

38 Le SIG (Système d'Information Géographique) est un système informatisé de gestion de données ordonnées dans l'espace qui fournit des informations géo-référencées sur des objets spatiaux. Typiquement un SIG est un composite formé de plusieurs bases de données où chaque base correspond à une variable : la topographie, l'habitat, l'élevage, etc. L'implémentation informatique de ces bases permet de les combiner afin de chercher les intersections, les corrélations et autres rapports entre les données et les entités géographiques[14] **(fig. n°6)**. En gérant à la fois des identificateurs uniques et l'ensemble des attributs physiques, sociaux, et autres pour des milliers d'objets géographiques, le SIG optimise l'information spatiale en vue de satisfaire une multitude de requêtes simples ou croisées.

Figure 6



Inclusion de données lexicométriques dans un SIG.

Convergence des données

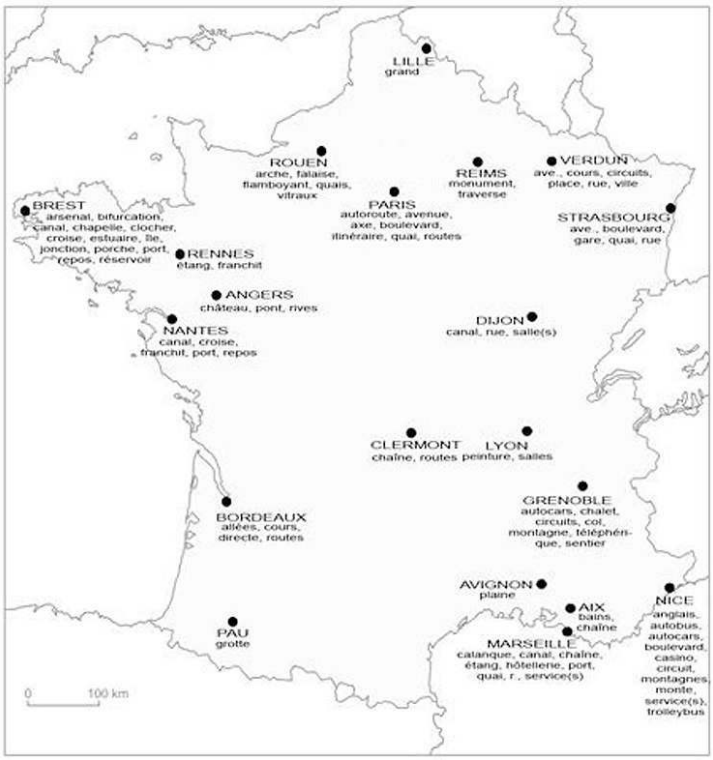
39 Même si elle n'est pas inscrite formellement dans les textes sous forme de coordonnées, la réalité géographique des guides de tourisme émerge systématiquement dans les résultats lexicométriques. Ceci est particulièrement visible dans les plans factoriels où les villes s'assemblent en agrégats sur la base du vocabulaire choisi pour les décrire et forment des ensembles qui font sens au niveau géographique. Dans les réseaux de cooccurrence, ce sont des champs sémantiques qui s'esquissent pour définir les éléments récurrents d'une « micro-géographie » : la *place* et sa *statue*, le *chemin* qui *franchit* le *canal*, l'*établissement* d' *eaux thermales*, etc.

40 Durant la manipulation de ces différents schémas et diagrammes, une interaction fertile entre ces données a successivement permis d'identifier les termes-clefs de la description touristique, d'étudier leur comportement en contexte et de projeter les systèmes lexicaux qu'ils forment sur des cartes. Ci-après nous présenterons deux expériences qui illustrent une exploitation efficace des données lexicométriques et géographiques et montrent comment les emplois lexicaux et les agglomérats qui sont suggérés dans les plans graphiques de l'analyse factorielle se confirment dans les projections du SIG en tant que zones territoriales.

41 Une première expérience de fusion des données lexicométriques et géographiques consiste à projeter sur une carte de la France les principales formes lexicales issues du réseau de cooccurrence élaboré dans le guide de 1956. Sur la figure 7 on constate que les affinités lexicales et les réseaux de cooccurrence qu'elles forment contribuent largement à la différenciation spatiale : le *milieu urbain* (Est), la *nature* (Nord-Ouest) et la *villégiature*

(Sud-Est)[15] **(fig. n°7)**. La projection de quelques villes sur une carte vierge suffit donc pour que des champs de force se définissent, des lignes de partage se dessinent et des clivages apparaissent.

Figure 7



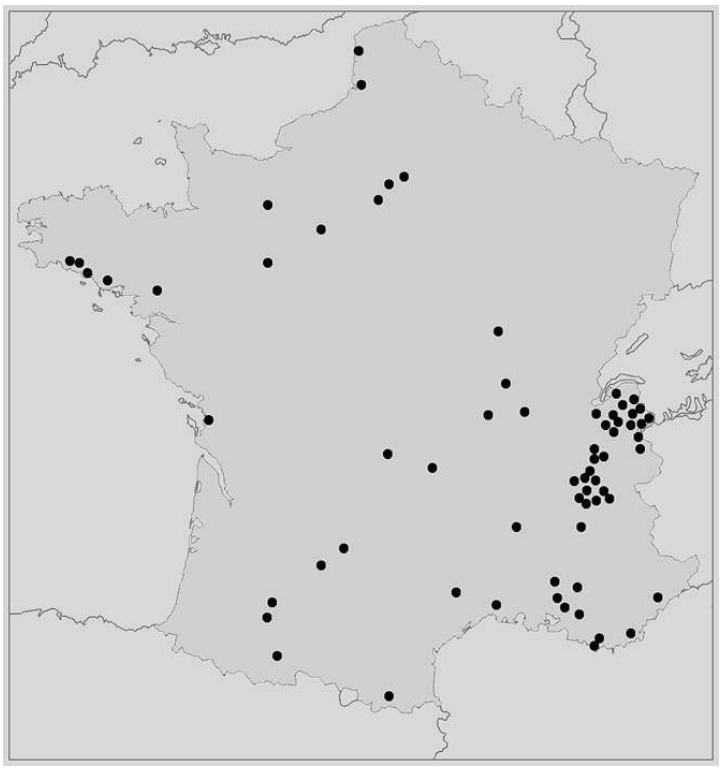
Projection de lexique spécifique sur une carte géographique (Guide 1956).

42 Si une telle combinaison de données cooccurrentielles et géographiques est possible et si elle produit des résultats pertinents qui respectent une logique géographique c'est parce que le réseau, qu'il soit concret ou virtuel, est la structure qui rend possible l'identité distincte des entités[16]. En effet, en science géographique on considère un territoire comme un ensemble de lieux et de connexions qui mettent ceux-ci en relation, soit un espace où des réseaux naissent de l'émergence de noeuds centraux qui génèrent une activité multiple et intense (zones de peuplement, centres de production, etc.). De toute évidence, cette dynamique est éminemment proche de celle qui régit les réseaux de cooccurrence...

43 Considérons maintenant un autre exemple, plus proche de la pratique du SIG, et qui exploite à bon escient l'analyse cooccurrentielle - notamment sa capacité exploratrice des textes et sa comparaison pondérée des usages des mots en contexte. Imaginons une requête à propos de l'adjectif *pittoresque*. La statistique descriptive nous informe que de 72 occurrences de l'adjectif dans le guide de 1873 on passe à 472 emplois dans le texte de 1956. Ce constat d'augmentation fréquentielle est en soi une information intéressante que l'on pourrait représenter dans un SIG afin d'apprécier l'inflation du terme, mais ici, nous exploiterons le calcul des cooccurrences pour relever les occurrences les plus surprenantes de l'adjectif au sens lexicométrique, c'est-à-dire les emplois en contexte où la forme n'est pas prévisible par la probabilité soient environ 50 emplois du mot pour 1873 et 90 pour 1956 **(fig. n°8, n°9)**. Sur le plan linguistique cela signifie que si le texte

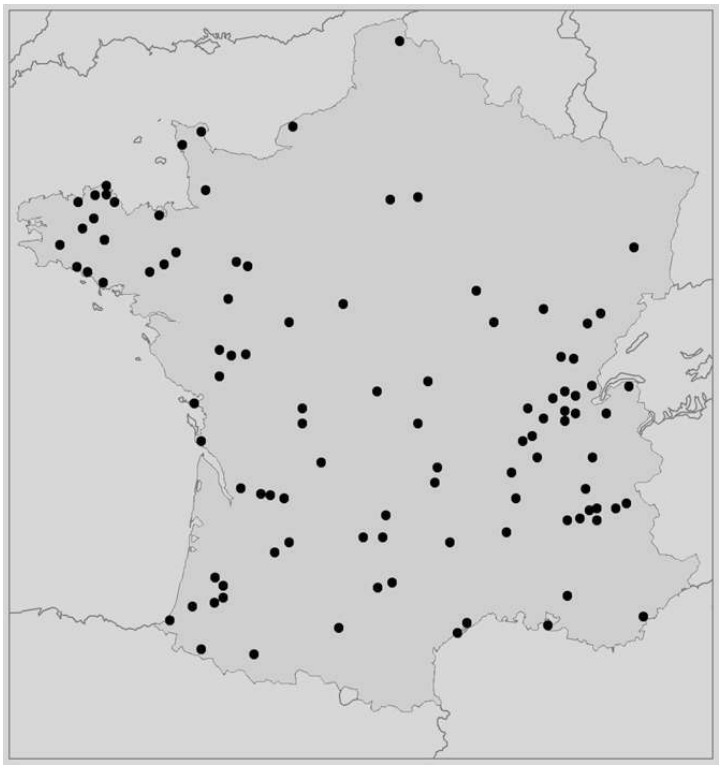
descriptif d'une ville ou d'un village est réduit à quelques lignes l'auteur choisira souvent l'adjectif *pittoresque* pour dépeindre concisément le lieu visité.

Figure 8



Projection du mot « pittoresque » dans le SIG (1873).

Figure 9



Projection du mot « pittoresque » dans le SIG (1956).

44 La cartographie de type SIG que nous avons réalisée pour les deux guides permet de prendre du recul sur le texte en proposant une focale entièrement différente. De la récurrence d'un adjectif en contexte tel qu'il peut être relevé par la lecture et comptabilisé par la statistique, nous passons à une carte qui montre clairement la dilution dans l'emploi du terme qui à l'origine est un adjectif employé de manière originale pour qualifier quelques rares zones territoriales et qui devient quelques années après un qualificatif plus commun. En effet, en 1873, les emplois originaux de *pittoresque* sont concentrés dans les régions Alpes et Côte d'Azur c'est-à-dire le Sud-Est de la France, zone où, comme l'a relevé la lexicométrie **(voir fig. n°7)**, une infrastructure touristique est déjà implantée au XIX^{e} siècle. En revanche, en 1956 plusieurs occurrences remarquables - au sens probabiliste - de l'adjectif sont relevées dans les régions Centre, Limousin, Poitou-Charentes et Aquitaine jusqu'à là vierges de tout emploi du mot. Cette dissémination de *pittoresque* sur le territoire s'explique par une mise en valeur de villes et de villages jusqu'à là ignorés par le développement touristique.

Conclusion

45 Pondérant les volumes textuels et comparant les fréquences lexicales, l'analyse lexicométrique de deux guides de tourisme de la France du XIX^{e} et XX^{e} siècles produit des cartes d'emploi du vocabulaire descriptif qui dévoilent le jeu des interactions entre lieux et lexique. D'une part, l'approche typologique de l'Analyse factorielle met en évidence un fonds de vocabulaire commun qui est à la base de la structuration globale des textes ainsi que des éléments lexicaux qui évoluent dans le corpus et au fil du temps, ce qui permet de

discriminer linguistiquement des villes et des régions de manière spatio-temporelle. D'autre part, l'analyse des cooccurrences qui opère au plus près du texte examine les tendances associatives des lexies qui tissent progressivement une toile et structurent le corps du texte. Au final, ces images qui résument les corrélations entre toponymes et éléments descriptifs prééminents esquissent une représentation des structures fortes du territoire.

46 Le rapprochement naturel entre les cartographies factorielle, cooccurrentielle et géographique nous a amené à imaginer une intégration directe des données lexicostatistiques dans un Système d'Information Géographique. Les expériences menées et les résultats obtenus ont permis de définir une méthodologie pour une exploitation complète et automatisée des ressources documentaires nouvelles et variées (Internet et toute base de données textuelles) qui, accouplées aux données géo-référencées du Système d'Information Géographique, viennent enrichir l'information cartographique et améliorer la compréhension de l'espace.

47 Au-delà de la combinaison efficace d'outils et de données informatiques nos expériences mettent en évidence des parallèles naturels entre les réseaux géographiques matériels et les réseaux lexicométriques immatériels. En effet, de prime abord la démarche peut surprendre : à l'opposé du géo-réferencement exact et précis du SIG, le vocabulaire descriptif des guides qui est à la fois subjectif et variable au cours du temps semble être une approximation insuffisante pour intégrer un système censé gérer des données objectives et générer des cartes conformes à la réalité. Pourtant, le dépouillement lexicométrique de milliers de pages de descriptions touristiques prouve que l'inventaire linguistique constitue un vecteur descriptif de grande valeur. Indéniablement les mots organisent l'espace à leur manière et une analyse systématique de leur organisation produit une géographie des représentations de l'espace vécu parfaitement cohérente au regard des impératifs de la science géographique. Par conséquent, l'étude de la complexité spatiale doit, au-delà des descriptions formelles, inclure la représentation de l'espace telle qu'elle s'élabore dans les esprits.

BIBLIOGRAPHIE

BRUNET, R. **La carte, mode d'emploi**. Paris : Fayard, Montpellier : Reclus, 1987.

STEINBERG, J. **Cartographie, télédétection, systèmes d'informations géographiques**. Paris : Armand Colin, coll. Campus, 2002.

LEBART, L. et SALEM, A. **Statistique textuelle**. Paris : Dunod, 1994.

MARTINEZ, W. **Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels**, Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3, soutenue le 17 décembre 2003.

SALEM, A. **Approches du temps lexical. Statistique textuelle et séries chronologiques**. Mots nº 17, 1998.

SANDERS, L. **L'analyse des données appliquée à la géographie**. Montpellier : Reclus, 1989.

NOTES

1. - Les résultats présentés ici proviennent d'un projet de recherche mené au Pôle Sciences de la Ville à l'Université Denis Diderot Paris 7 entre 2007 et 2009. Il s'agit de déterminer par le biais de méthodes de statistique textuelle le vocabulaire de l'espace urbain tel qu'il se développe dans un ensemble de guides de voyage depuis le milieu du XIX^{e} siècle jusqu'à nos jours. Le projet originel porte sur un ensemble de 20 guides rédigés en français, anglais et allemand et répartis sur la période 1867-1881 et les années 1950. Ces guides couvrent plusieurs pays dont l'Italie, l'Angleterre, l'Allemagne, l'Autriche, le Luxembourg, la Hongrie, la Tchécoslovaquie, la Grèce, la Russie et les pays d'Orient.

2. - Les analyses cooccurrentielles présentées ici ont été réalisées avec *CooCS* (www.williammartinez.fr/coocs) outil informatique pour formaliser et visualiser les réseaux de mots associés. Ce logiciel opère sur tout corpus constitué en base textuelle avec le programme *Lexico3* (www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW).

3. - On peut s'étonner des similitudes entre le nombre de formes et de hapax malgré la différence des volumes textuels mais le rapport « volume/vocabulaire » est régi par une loi statistique d'accroissement du vocabulaire qui limite progressivement l'émergence de vocables originaux dans tout discours en expansion : la répétition lexicale est inévitable. Signalons aussi que tous les calculs statistiques réalisés compensent les déséquilibres observés dans les volumes textuels afin de calculer des indices de spécificité des termes qui reflètent l'originalité du vocabulaire de manière pondérée.

4. - L'analyse factorielle opère à partir d'un tableau qui croise les mots et les parties discernées dans le texte. À l'intersection d'une ligne et d'une colonne de ce tableau on trouve la fréquence du mot dans une partie donnée. Le calcul factoriel vise à résumer toute cette information sur la ventilation des mots dans le texte sous la forme de plusieurs facteurs représentables sur des graphiques. Pour le corpus 1873, le premier facteur résume à lui seul 29 % de l'information et si on l'associe au second facteur ils apportent ensemble 56 % d'information, les 44 % restants étant des correctifs représentés par les facteurs suivants. Pour le corpus de 1956, les deux facteurs présentés ici résument 19 % et 17 % de l'information soit une bonne approximation de la structuration du texte.

5. - On opte pour un découpage des guides par itinéraires car le nombre de villes décrites dépasse les 1 500 ce qui est irreprésentable par analyse factorielle.

6. - Guide de lecture de la figure 1 : À travers sa représentation graphique l'analyse factorielle révèle les clivages entre les individus et les variables statistiques (respectivement les mots et les villes) en disposant l'information sur un plan graphique où elle place les formes et les parties par rapport aux axes : au centre les individus peu spécifiques, aux extrémités les individus très spécifiques.

7. - Guide de lecture de la figure 2 : À travers sa représentation graphique l'analyse factorielle révèle les clivages entre les individus et les variables statistiques (respectivement les mots et les villes) en disposant l'information sur un plan graphique où elle place les formes et les parties par rapport aux axes : au centre les individus peu spécifiques, aux extrémités les individus très spécifiques.

8. - Tous nos modules d'analyse exploitent le Modèle Hypergéométrique. Celui-ci est fondé sur la distribution en probabilité du nombre de rencontres de toutes les permutations possibles des formes étudiées dans l'hypothèse d'équiprobabilité, et il détermine la valeur la plus probable d'après les paramètres suivants : T : le nombre d'occurrences dans le corpus ; t : le nombre d'occurrences dans les contextes du pôle ; F : la fréquence du cooccurrent dans le corpus ; f : la

fréquence du cooccurrent dans les contextes du pôle. À partir de cette valeur probable, on calcule un diagnostic de spécificité signalant l'écart par rapport à la valeur attendue - un écart qui peut être positif, négatif ou nul. Si la fréquence réelle est supérieure à la fréquence attendue, alors la forme est spécifique positive et nous l'indiquons par le code *+x*. Si la fréquence réelle est inférieure à la fréquence attendue, la forme est spécifique négative et nous l'indiquons par le code *-x*. Enfin, si la fréquence réelle est égale à la fréquence attendue, alors la forme est banale. La valeur numérique de l'indice mesure quant à elle le degré de probabilité de l'événement : un indice de 3 signalera une probabilité de 1 sur 1 000, 4 une probabilité de 1 sur 10 000, etc.

9. - La taille de la fenêtre d'exploration contextuelle est déterminante puisque celle-ci circonscrit une zone textuelle où le module de comptage repère les attractions autour du pôle. La phrase constitue à ce titre l'unité contextuelle privilégiée car on la considère en linguistique comme le lieu de l'expression fondamentale qui relie le sujet et le prédicat c'est-à-dire ce dont on parle et ce que l'on en dit. Toutefois, les phénomènes d'anaphore (le remplacement du sujet par un pronom ou un synonyme pour éviter les répétitions en contexte comme par exemple *Paris*, *la capitale*, *elle*) compliquent le repérage des cooccurrences autour des villes mentionnées dans les guides. Pour obtenir les résultats présentés ici, nous avons parfois contourné ce problème en créant des super-contextes qui fusionnent plusieurs phrases et permettent la détection d'attractions à distance.

10. - De fait, les résultats initiaux du calcul exhaustif sont si volumineux qu'il faut instaurer certains filtrages pour ne recueillir que les cooccurrences les plus significatives au plan statistique. Après le calcul de l'ensemble des attractions binaires en contexte, les cooccurrences sont filtrées par spécificité (on retiendra les attractions de spécificité +5 c'est-à-dire celles qui n'avaient qu'une chance sur 100 000 de se produire – Voir la note 6) et par réciprocité (dans une cooccurrence entre deux mots, chaque mot doit attirer l'autre avec le même indice de probabilité).

11. - On peut vérifier cette observation en calculant le rapport *valence lexicale / fréquence* pour certaines formes : en 1956, *place* a un ratio *VL/F* (87/1190 = 0.073) équivalent à celui de *vallée* (116/1677 = 0.069) qui est pourtant un mot bien plus fréquent. Autre exemple : de 1873 à 1956, *musée* conserve un ratio de valence lexicale comparable avec respectivement une valence lexicale pondérée de 41/215 = 0.16 et 103/717 = 0.14.

12. - Guide de lecture de la figure 3 : Le réseau de cooccurrence est construit à partir d'une analyse exhaustive des attractions observées autour de chaque forme lexicale en contexte avec une fréquence d'apparition supérieure ou égale à 5 soit un ensemble de près de 1 900 mots. Malgré le filtrage des cooccurrences non réciproques, la combinatoire est si grande que le graphe obtenu est illisible tant sa complexité est grande : 1 849 nœuds et 4 014 arcs. On opte alors pour un filtrage des formes les plus magnétiques : seuls les mots avec au moins 50 cooccurrents en contexte seront affichés. Le résumé ne présente alors que les 68 formes les plus magnétiques du corpus mais il faut garder à l'esprit que cette sélection repose sur une masse cooccurrentielle non visible sur le graphe mais qui contribue à projeter ces formes vers l'avant. Ce filtrage a un effet de clarification des données puisqu'il dégage les principaux agrégats et représente par des formes-concept les plus importants champs sémantiques du corpus : établissements de bains, transports, châteaux et églises, culture, etc.

13. - Une analyse globale du corpus montre que la classe lexicale qui est employée pour la description de la *ville* n'est pas la seule qui s'élabore dans les textes. De fait, plusieurs autres catégories sémantiques toutes aussi homogènes ont été dévoilées par l'analyse des attractions lexicales parmi lesquelles : - le déplacement (*route, trajet, aller*, etc.) - l'architecture (*maison, palais, ruines*, etc.) - les activités commerciales ou industrielles (*commerce, grains, houille,* etc.) - paysage et observation (*panorama*, *vue*, etc.) - villes d'eaux (*bains*, *thermal*, *source*, etc.). Ces ensembles lexicaux semblent *a priori* étrangers au vocabulaire de la ville mais en réalité ils convergent vers

la description de l'espace urbain en lui ajoutant des nappes conceptuelles complémentaires (comment regarde-t-on la ville ? comment s'y déplace-t-on ? etc.).

14. - Guide de lecture de la figure 6 : Un SIG (Système d'Information Géographique) associe plusieurs bases de données afin de combiner la richesse des différentes informations qui y sont collectées (hydrographie, habitat, élevage, etc.) avec la précision de coordonnées géographiques univoques en vue de produire des cartes ad hoc pour une requête donnée. Ces cartes multidimensionnelles gagnent à être complétées par une couche lexicométrique qui livre pour chaque zone territoriale (ici Marseille) le vocabulaire employé pour sa description dans les guides de tourisme.

15. - Guide de lecture de la figure 7 : En exploitant la précision du géo-référencement propre au Système d'Information Géographique, il est possible de projeter les données issues de l'analyse lexicométrique autour de différents toponymes sur un carte géographique. Le placement exact des points-villes permet de visualiser les corrélations entre termes descriptifs et villes décrites, et au-delà, à une échelle territoriale plus vaste de délimiter des zones de cohérence sémantique.

16. - Un parallèle s'établit avec la structure différentielle que F. de Saussure décrit comme la nécessaire base de construction du sens dans toute langue naturelle : chaque mot prend sens de par ses différences avec les autres mots du système.

RÉSUMÉS

L'analyse statistique de la distribution du vocabulaire dans des guides de tourisme du XIX^{e} et XX^{e} siècle permet d'identifier la terminologie descriptive essentielle telle qu'elle est employée dans ces textes. Une interprétation plus structurée de ces données est possible grâce aux méthodes de cooccurrence qui produisent des cartes de mots associés décrivant l'usage préféré de noms, adjectifs, adverbes, etc. à propos d'une ville, une région ou un itinéraire. À partir de cette visualisation originale de structures de mots, nous envisagerons la convergence des données lexicales et cartographiques dans une base de données de type SIG (Système d'Information Géographique).

The statistical analysis of vocabulary distribution in French tourist guides of the 19th and 20th century reveals the essential descriptive terminology used in these texts. A more structured interpretation of this data is made possible by way of co-occurrence methods that produce associated word maps describing the preferred usage of nouns, adjectives, adverbs etc. regarding a given town, region or route. Based on this original visualizing of word structures we will consider the convergence of lexical and cartographic data in a GIS-type database (Geographic Information System).

INDEX

Mots-clés : cooccurrences lexicales, réseaux de mots associés, SIG, statistique textuelle
Keywords : associated word networks, GIS, lexical co-occurrences, textual statistics

AUTEUR

WILLIAM MARTINEZ

Instituto de Linguística Teórica e Computacional (ILTEC), Lisbonne, will_martinez@hotmail.com