La topologie textuelle : légitimation d'une notion émergente

Sylvie Mellet, Jean-Pierre Barthélemy

La topologie textuelle : légitimation d'une notion émergente

Sylvie Mellet*, Jean-Pierre Barthélemy**

Introduction

Jusqu'à il y a peu, les principaux développements de la statistique linguistique – y compris lorsqu'elle s'appliquait à l'étude des textes et des discours – ont essentiellement recouru à des modèles qui tendent à négliger ce fait majeur qu'un texte est une structure ordonnée ; les dénombrements, les relevés de fréquences, les calculs de spécificités caractérisant l'Analyse des Données textuelles reposent pour l'essentiel sur le schéma d'urne qui transforme un texte en un sac de mots et renoncent à prendre en compte le positionnement dans le texte des unités dénombrées[1]. Certes, les résultats ainsi obtenus sont généralement intéressants et bien interprétables, et ils ont largement contribué au développement et aux succès de la discipline. Mais il se pourrait qu'ils fussent en train d'atteindre leurs limites. Ou, du moins, de ne plus suffire à donner entière satisfaction au chercheur. De plus en plus souvent en effet, celui-ci souhaite pouvoir étudier, à côté de la dimension paradigmatique appréhendée par ce type de calculs statistiques traditionnels, la dimension syntagmatique des données textuelles, saisie à courte ou à longue portée : distribution régulière ou non d'une entité linguistique susceptible d'arriver à intervalles à peu près égaux ou, au contraire, en paquets plus ou moins denses ; répartition d'un élément au fil du texte selon la structure globale de celui-ci et de ses parties constituantes ; phénomènes d'échos et d'alignements dans la mise en parallèle de deux textes ou deux portions de textes ; motifs répétés de cooccurrents multiples, etc. Le recours à diverses formes de partitionnement des corpus et à l'étude contrastive des différentes parties ainsi constituées représente une première tentative intéressante pour répondre à ces attentes et tenter de localiser les faits observés dans un espace précis du texte ou du corpus ; mais il convient sans doute d'aller plus loin dans le développement de méthodes nouvelles permettant de fournir une véritable représentation spatiale (ou cartographie) des faits et d'en induire pour chaque texte un profil caractéristique ouvrant la voie à la comparaison et à l'interprétation. Plusieurs travaux récents vont dans ce sens : voir ici même le bilan établi par Damon Mayaffre (2007b).

L'évolution des pratiques est désormais suffisamment sensible pour qu'on éprouve le besoin de nommer cette nouvelle approche, afin de lui donner un statut et de la faire reconnaître. Certains chercheurs ont d'abord parlé de « topographie textuelle »[2]. Nous-mêmes avons osé employer le terme de « topologie textuelle »[3]. Or, au fil de divers exposés et des articles publiés, ce qui n'était qu'une approximation terminologique a fini par s'imposer à nous à la fois comme une nécessité et comme une interrogation épistémologiques fondamentales : quelle est donc la pertinence et quelle est surtout la légitimité scientifique de ce terme « topologie » qui nous semble intuitivement juste et dont nous avons désormais du mal à nous passer ? Telle est la question à laquelle cet article va essayer de répondre.

1. Emergence d'une notion, première étape : l'emprunt étymologisant naïf.

*Laboratoire BCL, Université Nice Sophia-Antipolis, CNRS ; MSH de Nice, 98 bd E. Herriot, 06200 NICE
** ENST de Bretagne, BP 832, 29285 Brest Cédex

[1] Signalons toutefois l'article, très novateur de ce point de vue, de Lafon (1981).

[2] Sérant & Thoiron (1988) ; Lamalle & Salem (2002).

[3] Longrée D., Luong X. & Mellet S. (2004) et (2006) ; Luong X., Juillard M., Mellet S. & Longrée D. (2007).

Un texte n'est donc pas un sac dans lequel seraient rassemblées en vrac ses unités constitutives. Un texte est à tout le moins une chaîne linéaire, donc un espace ordonné à une dimension. Il arrive parfois que son inscription sur la page, signifiante, oblige à le placer (et à l'interpréter) dans un espace à deux dimensions : tel est le cas, par exemple, pour de nombreux poèmes dont la structure formelle se lit à la surface du texte (succession des différents types de strophes, alternance des rimes – sans même parler des acrostiches ou des calligrammes) ; tel est aussi le cas pour de nombreux textes d'affiches, publicitaires ou autres. Enfin, il faudrait ajouter la dimension temporelle qui, sous la forme de la mémoire à court et à long terme, joue un rôle fondamental dans la structuration et la réception de tout texte de quelque ampleur[4].

Or qui dit espace dit aussi lieux : lieux privilégiés de concentration de l'information, lieux plus atones de redondance ou de pause informative, lieux de transition et zones de rupture, entre deux développements thématiques différents par exemple. Autant de « topoï » intrinsèques, inhérents à tout déploiement textuel. Tout cela est bien connu de toutes les disciplines qui ont pour objet d'étude le texte, mais on avait cru pouvoir l'occulter dans les traitements statistiques (ou ne le récupérer qu'*a posteriori*).

La reconnaissance, la prise en compte de l'existence dans les textes de ces lieux différenciés nous autorisent-elles à parer nos études du label de la « topologie » ? Non, bien sûr. Tout au plus suggèrent-elles l'ébauche d'un rapprochement possible, lequel devient un peu plus insistant cependant lorsqu'on déclare simultanément qu'un texte c'est aussi un ensemble (E) d'unités linguistiques qui ne sont pas indépendantes les unes des autres, muni d'une structure ou, plus exactement, de plusieurs structures imbriquées dont l'union constitue cet ensemble. Ces structures dessinent des sous-ensembles de (E), délimités par des zones frontières dont les propriétés sont souvent très intéressantes à étudier. Il devient alors tentant de comparer le texte à un espace topologique[5] dont les formes pourraient être analysées quantitativement ***et*** qualitativement. L'approche topologique, conformément aux fondements qui sont à l'origine de la discipline, permettrait en effet au linguiste statisticien d'une part de sortir du tout quantitatif, d'autre part d'articuler le local et le global[6], les micro-structures aux macro-structures.

2. Emergence d'une notion, deuxième étape : le voisinage.

[4] On devine ici l'importance de l'approche « réticulaire » défendue notamment par Jean-Marie Viprey (voir Viprey 2006-b).

[5] Rappelons la hiérarchie traditionnelle des structures mathématiques auxquelles on associe des classes particulières de fonctions (appelées aujourd'hui « applications »). Ainsi, en allant du moins contraint au plus contraint, on distingue :

- les structures ensemblistes dans lesquelles les points sont indépendants les uns des autres et auxquelles on peut appliquer des fonctions quelconques ;
- les structures topologiques dont les éléments sont liés par des relations de voisinage et auxquelles on appliquera des fonctions continues (sur l'ensemble des nombres réels par exemple) ;
- les structures différentiables, telles que les courbes et les surfaces, pour lesquelles on peut, entre autres, calculer des dérivées ;
- les structures métriques telles que les espaces euclidiens, hyperboliques, etc.
- les espaces vectoriels.

[6] L'articulation du local et du global était l'un des objectifs principaux de Poincaré lorsqu'il introduisit la topologie en mathématique sous le nom d'« analyse *in situ* ».

Nous considérons donc qu'un texte est une structure linéaire constituée d'un ensemble d'événements linguistiques (occurrence d'un mot, d'un syntagme, d'une catégorie grammaticale, etc.) qui, chacun à leur tour, peuvent être considérés comme des points remarquables de la chaîne textuelle. Mais ce faisant, l'analyste ne saurait détacher l'unité observée de son contexte immédiat, c'est-à-dire de la portion textuelle jugée pertinente pour l'analyse et qui comprend un certain nombre d'autres mots (ou, plus généralement, d'autres événements linguistiques) situés avant et après lui. La pertinence de la taille du contexte varie selon les faits étudiés, la taille globale du texte, l'objectif de la recherche, etc. L'empan contextuel doit donc être défini par le linguiste en fonction de ces différents paramètres ; si l'ensemble des paramètres n'est pas maîtrisable, le choix d'une taille arbitraire ajustée par essais/erreurs peut apparaître légitime et donner d'excellents résultats.

Or qu'est-ce donc qu'un nombre [arbitrairement] fixé d'éléments x_i de (E) entourant le point x (*i.e.* l'occurrence étudiée), sinon un ***voisinage*** de x[7] ? Et puisque l'on peut, selon les besoins de l'étude, faire varier ce nombre et donc la taille des contextes ainsi définis, ne peut-on considérer que chaque point x se trouve muni d'une ***famille de voisinages*** ? Le texte étant ainsi composé d'un ensemble d'occurrences auxquelles sont associées des familles de voisinages, il peut légitimement être étudié comme un espace à topologie discrète.

Notons au passage que le concept d'espace topologique donne un cadre formel et un modèle mathématique à la notion relativement intuitive d'objet construit. En effet, comme tout objet scientifique, l'objet de la linguistique textuelle est un objet construit : aucun chercheur ne saurait prétendre travailler sur des données brutes, ne serait-ce que parce qu'il sélectionne son objet parmi une infinité de possibles en fonction de ses intérêts et de son projet de recherche. La notion de voisinage, parce qu'elle affiche explicitement la sélection d'une unité remarquable parmi d'autres et parce qu'elle prévoit la pluralité des « focales » d'observation (tailles des voisinages), donne un statut épistémologique clair à l'objet construit. Elle précise aussi la notion de « contextualisation », souvent mise en avant par F. Rastier par exemple. En voici une illustration sommaire dans laquelle on prêtera attention aux processus de choix et de réduction de l'objet étudié.

Soit un texte (fictif) qu'on a transformé en objet d'étude par réduction à ses seules formes verbales et, plus précisément, à la succession des temps verbaux de l'indicatif qui le structure. Nous avons affaire là à une suite, généralement signifiante et caractéristique. Ainsi, une suite telle que :

[IMP. IMP. PQP. IMP. IMP. PS. IMP. PS. PS. PS. IMP. PQP. PS. PS. PS. PS. PS.][8]

relève, selon la classique distinction benvenistienne, de l'histoire (par opposition au discours) et trahit d'emblée une première partie introductive posant probablement le cadre des événements narrés (succession de cinq imparfaits et plus-que-parfaits), suivie d'une séquence narrative au passé simple très nette (les cinq derniers verbes), avec, entre les deux, une zone de transition (mélange de passés simples et d'imparfaits ou plus-que-parfaits) qui rend la frontière entre les deux parties difficile à poser. Dans ce cadre ainsi schématisé de la narration historique, chaque élément de la forme textuelle étudiée peut être caractérisé par un voisinage assez étroit (par exemple de taille (5) et par une mesure de ce voisinage évaluant la ***densité*** (d, $0 \leq d \leq 5$) des paramètres pertinents présents dans ce voisinage, par exemple le nombre des formes temporelles descriptives ; une telle mesure peut rester sommaire, ne prenant en compte que la fréquence des imparfaits ou devenir plus complexe en tenant compte aussi de la présence des plus-que-parfaits jugés équivalents des imparfaits au regard de la propriété

[7] Le concept de *voisinage* d'un point dans un ensemble, au sens intuitif *d'entourer*, *d'être proche*, est à la base de la définition axiomatique des *espaces topologiques* ; ainsi, dans **R**, ensemble des nombres réels, tout intervalle $]x - h, x + h[$, avec $h > 0$ est un *voisinage* de x.

[8] IMP = imparfait, PQP = plus-que-parfait, PS = passé simple.

descriptive, mais aussi des participes apposés ou de certaines subordonnées qu'il faudrait alors réintégrer dans la chaîne syntagmatique et dont on pourrait éventuellement pondérer le poids dans la mesure. C'est là que le qualitatif peut être réintroduit dans l'analyse en fonction des hypothèses de travail du chercheur et par le biais de la multiplication des paramètres descripteurs du voisinage qui permettent d'atteindre des niveaux de complexité propres à rendre compte de la richesse textuelle.

On peut aussi ne pas prendre en compte chaque forme verbale successive, mais ne s'intéresser qu'à l'une d'entre elles en particulier. Par exemple centrer l'étude sur le temps narratif par excellence, à savoir le passé simple, et caractériser les variations de ses voisinages au fil du texte. D'autre part la taille (5) du voisinage étant arbitraire, elle peut être modifiée au gré du chercheur ; celui-ci peut aussi vouloir travailler à partir d'une base de voisinage linguistiquement pertinente (par exemple, la phrase – auquel cas il faudra réintégrer dans les données les marques de ponctuation forte), et passer ensuite à une famille de voisinage plus large si le besoin s'en fait sentir (rareté des occurrences d'un phénomène, significativité de séquences plus longues que la phrase telles le paragraphe, etc.). On trouvera une application de cette méthode dans Longrée, Luong & Mellet (2004).

La notion de voisinage est donc susceptible de fonder en partie une approche topologique de l'analyse textuelle. Elle reste néanmoins insuffisante dans la mesure où elle ne prend que partiellement en compte l'ordre linéaire des éléments et la récurrence des figures plus ou moins complexes, caractéristiques de certains textes ou parties de textes : elle ne suffit donc pas à donner accès à cette propriété intrinsèque de l'objet texte qui pourrait être décrite comme une structure topologique – ou qui, du moins, pourrait être évaluée à cette aune. Il faut pour cela s'intéresser à des marqueurs linguistiques qui constituent des schèmes ***ordonnés*** récurrents, et qui sont susceptibles de caractériser telle ou telle partie du texte à différents niveaux de granularité et d'aider ainsi au repérage des diverses zones textuelles. Les motifs pourraient être un bon candidat à ce rôle.

3. Les motifs : confrontation aux axiomes de la topologie sur un espace fini.

On appelle ici « motif » l'association récurrente de n éléments de l'ensemble (E) muni de sa structure linéaire qui donne une pertinence aux relations de successivité et de contiguïté. Ainsi, si l'ensemble (E) est composé de x occurrences des éléments A, B, C, D, E, F, un premier motif pourra être la récurrence du groupe linéairement ordonné ABD, un autre motif pourra être la récurrence du groupe AA[9].

Apparemment simple, cette définition soulève pourtant un certain nombre de difficultés dès que le linguiste cherche à la concrétiser :

- quelles unités linguistiques peuvent prétendre intégrer un motif ? A-t-on le droit de constituer des motifs hétérogènes, c'est-à-dire associant des éléments de nature différente (lexèmes, catégories grammaticales, …) ? Comment distinguer les motifs entièrement libres des motifs contraints par les structures de la langue (non seulement expressions lexicalisées, mais aussi enchaînements grammaticaux régis par des règles de dépendance hiérarchique) ?
- quelle place accorder à la ponctuation ? En d'autres termes, un enchaînement AB,D relève-t-il du motif ABD ? La phrase est-elle une unité pertinente de localisation des motifs ? En est-elle la seule ?

[9] On le voit, les motifs sont susceptibles d'englober les « segments répétés » (A. Salem) ; mais ils sont plus larges puisqu'ils intègrent aussi des codes grammaticaux et pourraient peut-être admettre l'insertion d'une variable parmi les éléments stables qui les définissent.

- comment gérer les espacements ? Si x est une occurrence quelconque au sein du texte, peut-on construire un motif ABxD ? Voire assimiler ABxD à ABxxD ?

Ces questions ne semblent pas pouvoir recevoir de réponses *a priori*. C'est à l'usage et, dans un premier temps, au coup par coup, que le linguiste aura à trancher. Ultérieurement, un ensemble de décisions concrètes pourra sans doute être généralisé et théorisé.

Pour l'instant, et pour revenir à notre propos, nous travaillerons sur un exemple qui met en jeu les divers types de propositions, principales et subordonnées, d'un texte narratif latin (la variation des unités sera donc relativement contrainte au sein d'un cadre grammatical défini) ; nous admettrons de ce fait que la phrase est une unité pertinente pour la détection des motifs liés à ce type d'unités linguistiques (cadre syntaxique phrastique classique) ; enfin nous n'exclurons pas *a priori* la possibilité de motifs complexes intégrant des occurrences quelconques en leur sein.

Les propositions qui nous intéressent sont d'une part la proposition principale, d'autre part certaines subordonnées qui contribuent à poser un cadre circonstanciel à l'action, tels les ablatifs absolus (*i.e.* une forme de proposition participiale) et les subordonnées en *cum* + subjonctif (*i.e.* l'équivalent approximatif des propositions en *alors que*, *comme*). L'organisation de toutes ces propositions structure en effet la narration et les stylisticiens ont depuis longtemps montré qu'elles contribuaient à caractériser le style d'un auteur : certains écrivains en effet les utilisent avec parcimonie, d'autres ne reculent pas devant leur accumulation. Certains préfèrent placer les subordonnées en début de phrase, pour poser d'abord le cadre de l'action principale, d'autres au contraire pratiquent volontiers la « relance syntaxique » en fin de phrase, ajoutant après coup des éléments informatifs secondaires. Les différents motifs et leur fréquence respective caractérisent donc le style d'un auteur ; leur distribution pourrait aussi caractériser les différentes parties d'une œuvre (les parties introductives ou de commentaires entre les passages narratifs offrant un cadre plus accueillant aux accumulations de subordonnées, par exemple). Le niveau local ouvre ici sur le niveau global.

Pour repérer plus aisément ces configurations potentielles, nous donnerons donc à la proposition principale le code P, aux subordonnées circonstancielles, quelle que soit leur nature exacte (participiales ou en *cum*), le code S et au point final de phrase le code $; x restera le symbole d'une occurrence quelconque de toute autre proposition.

Les motifs pertinents pour cette recherche sont donc susceptibles de prendre, entre autres, les formes suivantes :

-subordonnées circonstancielles en tête de phrase : $SP, $SSP, $SSSP, $SxP, $SSxP, $SSSxP

-subordonnées circonstancielles en fin de phrase : PS$, PSS$, PSSS$, PSxSS$

-configuration mixte avec pondération forte sur la fin de phrase : $SxPSS$, $SxxPSS$, $SxxPSSS$, $xSxPSS$

On voit, à partir de cette énumération non exhaustive, que les motifs sont des fermés qui ont la propriété de ne pas être nécessairement séparés. L'ensemble du texte est un motif à lui seul (à condition de transiger un peu sur la propriété de récurrence accordée initialement aux motifs) ; l'ensemble vide peut aussi être considéré comme un motif.

Par ailleurs, il est clair que l'intersection de deux motifs fournit un motif (plus ou moins pertinent pour la recherche – mais c'est là un autre problème qui ne relève plus de la topologie).

Un ensemble important de propriétés topologiques sont donc satisfaites.

En revanche, l'union de deux motifs ne fournit un motif que lorsque les deux motifs sont consécutifs. L'un des axiomes des structures topologiques *stricto sensu* est ici pris en défaut. Cependant, on peut trouver des séquences qui, au moins pour quelques motifs,

fournissent une borne supérieure[10] : ainsi, dans la liste ci-dessus, $SSSxP est la borne supérieure de l'ensemble des motifs rassemblant les circonstancielles en tête de phrase ; $SxxPSSS$, est la borne supérieure des deux motifs mixtes $SxPSS$, $SxxPSS$, mais pas du dernier motif de l'énumération $xSxPSS$. De la même façon il existe des motifs qui peuvent servir de borne inférieure[11] à un petit nombre d'autres motifs : ainsi $S constitue la borne inférieure de tous les motifs plaçant les circonstancielles en tête de phrase et des trois premiers motifs des configurations mixtes. On est donc peut-être légitimé à considérer que l'on a bien affaire à un ensemble muni d'un ordre partiel et dans lequel certains éléments au moins répondent aux deux conditions de borne inférieure et de borne supérieure, ce qui définit une structure topologique de « treillis ». Au sens strictement mathématique, tous les éléments de l'ensemble devraient répondre à ces conditions ; on peut néanmoins estimer que, dans la perspective qui est la nôtre, l'existence de plusieurs points signifiants du texte qui satisfont à ces deux conditions suffit à valider la transposition des notions topologiques à l'analyse des structures textuelles.

4. De la topologie comme propriété intrinsèque du texte à la topologie des espaces de représentation

Les motifs sont donc de bons candidats à fonder une approche proprement topologique des textes : ils permettent d'articuler le local au global et répondent à la plupart des axiomes de la topologie sur un espace fini. Comment les exploiter ?

Tout d'abord il est clair que pour appréhender un texte dans toute sa complexité on ne saurait se contenter de travailler sur une seule de ses dimensions. Le linguiste et/ou le stylisticien doit donc essayer de détecter plusieurs des dimensions pertinentes constitutives de la structure du texte ou des textes sous étude, de caractériser chacune de ces dimensions par la fréquence et la distribution des motifs afférents et de voir ensuite comment ces différentes classes de motifs s'articulent entre elles pour donner une image multidimensionnelle de la structure textuelle. Cette méthode qui progresse pas à pas et cumule les résultats repose bien évidemment sur une hypothèse de robustesse statistique ; on peut estimer que l'ensemble des études de linguistique statistique accumulées depuis les années 1950 valide suffisamment cette hypothèse. Elle nous permet en tous cas de quitter la simple linéarité de la structure de surface pour faire émerger les structures emboîtées auxquelles on prétend atteindre.

D'autre part, en linguistique quantitative et analyse des données textuelles, la caractérisation d'un texte est très souvent menée dans un but comparatif : il s'agit généralement d'évaluer les ressemblances ou les différences entre plusieurs textes au moyen de divers calculs de distance, et parfois même d'aboutir à une classification, – le tout étant sous-tendu par un projet interprétatif (caractérisation du style d'un auteur, d'un genre ou d'un sous-genre littéraire, d'une époque, détection d'une évolution diachronique, etc.). Sur ce point on prendra bien garde que les méthodes ici décrites ne relèvent pas des statistiques inférentielles : elles ne peuvent donc conduire qu'à la formulation d'hypothèses dans une démarche heuristique, mais ne sauraient en aucun cas valider une hypothèse préalable ; en particulier, elles n'ont aucune valeur probatoire en matière d'attribution d'auteur.

Comment donc s'opère la comparaison ? Il s'agit d'affecter à chaque texte un profil (vectoriel ou autre) qui le caractérise en sommant l'ensemble des informations apportées par l'étude des différents motifs. Chaque profil, constitué d'un même nombre de descripteurs, est ensuite intégré à une matrice rectangulaire dont les textes fournissent l'intitulé des lignes et

[10] Il s'agit du plus petit élément de l'ensemble – donc du plus petit motif – englobant le motif A et le motif B.
[11] Fait pendant à la borne supérieure et peut être assimilée à un plus grand dénominateur commun (plus grand des motifs inclus à la fois dans le motif A et dans le motif B).

dont les descripteurs de profil fournissent les colonnes. A partir de là peuvent être appliquées les méthodes classiques de calcul de distance.

Or on sait que l'interprétation de ces calculs passe généralement par une représentation graphique sur un espace à deux dimensions[12] ; les plus fréquentes d'entre elles sont « les méthodes en axes principaux, toutes dérivées de la *décomposition aux valeurs singulières* : analyse des correspondances (AC), Latent semantic Indexing (LSI), analyse en composantes principales (ACP). Ces méthodes peuvent être qualifiées de linéaires dans ce contexte, parce qu'elles projettent les points sur des droites ou des plans. » (Lebart 2006 : 593). Les classifications, quant à elles, recourent le plus souvent à divers types de dendogrammes, sans que soient totalement exclues les cartes ou les grilles comme, par exemple, dans le cas des cartes auto-organisées de Kohonen.

Quoi qu'il en soit, toutes ces représentations graphiques à des fins de visualisation pour aider l'interprétation possèdent leur propre topologie qui est une *topologie externe* au corpus étudié. Il s'agit en outre d'une topologie séparée alors que la topologie intrinsèque des textes est une topologie non séparée. La question qui surgit alors est celle du lien à établir (ou pas) entre ces deux topologies : est-il possible de fonder épistémologiquement le passage de l'une à l'autre[13] ?

Notons d'abord que cette topologie externe n'a jamais prétendu être l'image de la topologie interne des textes. Elle n'est que l'image de la structuration – ou, plutôt, d'une structuration possible – du ***corpus***, elle-même établie à partir de paramètres d'analyse qui prennent en compte la structure topologique interne de chaque texte. Le lien est très indirect et se reconnaît comme tel. La question initiale doit donc être déclinée en deux autres questions, plus précises et complémentaires : est-il légitime d'appuyer la structuration d'un corpus aux différentes topologies structurant chacun des textes qui le constitue ? Existe-t-il des représentations graphiques plus appropriées que d'autres à la visualisation de cette structure de corpus ?

La réponse à la première question sera, à notre avis, apportée par les résultats des différentes études en cours ou en projet. Seule l'expérimentation nous dira s'il est plus ou moins pertinent de comparer entre eux des textes et d'organiser leurs relations réciproques en prenant appui sur des faits de structure interne ou sur des faits autonomes, isolés de leur contexte comme on l'a généralement fait jusqu'ici. A dire vrai, un pan très large de la linguistique textuelle du XXème siècle oriente vers le primat du structurel (ou du compositionnel[14]). Et la première expérience que nous avons menée dans ce domaine a donné des résultats prometteurs puisque la prise en compte des faits de structure a permis d'aboutir à une classification automatique de 11 textes latins mettant en lumière des oppositions sous-génériques fines (biographies *vs* commentaires et annales) se superposant aux oppositions d'auteurs et d'époques, et qui échappaient aux autres méthodes d'analyse[15]. Il convient bien entendu de multiplier ces expérimentations pour confirmer ces résultats, en gardant en outre présent à l'esprit le fait que tout corpus est susceptible de plusieurs structurations. Mais, précisément, c'est là que réside l'un des atouts majeurs de la démarche topologique : c'est qu'elle ne peut être que multidimensionnelle. Le profil affecté à chaque texte et intégré à la matrice de calcul des distances est un profil complexe qui répercute la diversité des

[12] Avec le développement des outils informatiques, l'ergonomie de telles visualisations peut être améliorée par l'ajout d'une troisième dimension, par l'emploi de la couleur, voire par des procédés d'animation ; cf. Viprey (2006-a).

[13] On ne parle pas ici de l'évaluation des algorithmes et donc de l'adéquation entre les groupements de points obtenus et les données traitées : il existe de nombreuses méthodes permettant de calculer cet intervalle de confiance et de se prémunir ainsi contre des lectures interprétatives trop rapides ou trop simplistes. Voir Lebart (2004) et (2006).

[14] Cf. notamment les travaux de J.-M. Adam.

[15] Voir Longrée & Mellet (sous presse).

paramètres nécessaires pour rendre compte d'un texte dans sa richesse et sa singularité. Les colonnes de la matrice n'accueillent point ici de simples mots ni même des catégories grammaticales ; elles accueillent des éléments de structures complexes tels que les motifs évoqués ci-dessus, affectés de leur fréquence globale et de leur fréquence partielle dans chacune des zones du texte qu'ils ont contribué à délimiter. On peut estimer que ce profil fournit une image abstraite du texte, une sorte de schème sous-jacent et que l'ensemble du corpus acquiert, en effet, une topologie propre liée à la réunion de ces différents schèmes au sein d'un même ensemble.

Quant au choix de la représentation graphique la plus appropriée, aucun argument ne nous paraît décisif en la matière. On soulignera simplement que le lien entre la topologie textuelle et la topologie des espaces de représentations (qui, redisons-le encore, n'est pas un lien direct, mais un lien à double détente) varie très certainement en fonction de la représentation choisie. Par exemple, l'Analyse Factorielle des Correspondances (AFC), en dépit de la projection planaire à laquelle on est contraint de la réduire, conserve très exactement le nombre de dimensions initiales ; on peut donc toujours convoquer l'ensemble des plans factoriels successifs qui décrivent la structure globale sous la forme d'un nuage de points. En revanche l'AFC ne permet pas de constituer des classes autour des éléments les plus saillants de la structure ni de décrire les positions relatives des classes dans l'espace.

L'analyse arborée permet au contraire de constituer des classes, en imposant en outre une contrainte de proximité. L'algorithme présenté par Luong & Barthélemy (1987) a l'avantage d'associer à cette structuration classificatoire une représentation des distances (ce qui n'est généralement pas le cas dans les représentations arborées). Il donne en outre accès à la dynamique de construction de l'arbre, laquelle fournit des éléments d'information importants sur les proximités topologiques ; par ce biais, on retrouve – au niveau du corpus dans son entier – l'articulation entre le local (la constitution de chaque nœud de l'arbre) et le global (sa structure générale). En revanche l'analyse arborée, contrairement à l'AFC, ne permet pas de représenter simultanément dans un seul et même graphe les variables textuelles (lignes) et les paramètres descripteurs (colonnes)[16].

C'est pourquoi, le linguiste est tenté de recourir tantôt à l'un, tantôt à l'autre de ces deux modes de représentation spatiale. Mais, en toute rigueur, une étude ne devrait pas les mélanger, du moins sans précaution et sans argumentaire.

Conclusion

- La notion de topologie textuelle pour appréhender la structure interne d'un texte nous paraît légitimée à travers la construction (certes artéfactuelle) de voisinages et de motifs, qui les uns et les autres répondent parfaitement à la fois à la réalité de l'objet d'étude (encore une fois en tant qu'objet construit), aux besoins et aux enjeux de l'analyse et aux axiomes d'une véritable topologie.

- Le passage d'une topologie interne textuelle à une topologie externe des représentations se fait par le sas de la constitution du corpus et de sa structuration qui d'une part repose sur l'analyse textuelle préalable, d'autre part conduit à une représentation spécifique des éléments séparés qui le constituent.

Références bibliographiques

[16] Pour une approche complète des représentations arborées, voir Barthélemy & Guénoche (1988).

Adam J.-M. (2006). « Autour du concept de *texte*. Pour un dialogue des disciplines de l'analyse de données textuelles », in *JADT 2006* [texte en ligne sur *Lexicométrica (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)*].

Barthélemy J.-P. & Guénoche A. (1988). *Les arbres et les représentations des proximités*. Paris : Masson.

Barthélemy J.P. & Luong X. (1987). « Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmiques et applications à l'analyse de données textuelles », *Mathématiques et Sciences Humaines* 100 : 57-80.

Barthélemy J.-P. & Luong X. (1998). « Représenter les données textuelles par des arbres », in *JADT 1998*, *Actes des 4èmes Journées Internationales d'analyse de données textuelles*, Univ. de Nice : UMR 6039, pp. 49-70.

Brunet É. (2006). « Navigation dans les rafales », *in* J.M. Viprey (éd.), *JADT 06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. Besançon : Presses universitaires de Franche-Comté, vol. 1, pp. 15-29.

Cahiers de praxématique (2005). « Hétérogénéités énonciatives et types de séquences textuelles », n°45.

Lafon P. (1981). « Statistique des localisations des formes d'un texte », *Mots* 2 : 157-187.

Lamalle C. & Salem A. (2202). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », *in* A. Morin & P. Sébillot (éds) *JADT 2002, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*. Saint-Malo : IRISA / INRIA, vol. 1, pp. 403-411.

Lebart L. (2004). « Validité des visualisations de données textuelles », *in* G. Purnelle, C. Fairon & A. Dister (éds), *JADT 2004, 7èmes Journées internationales d'Analyse statistique des Données Textuelles*. UCL : Presses universitaires de Louvain, vol. 2, pp. 708-715.

Lebart L. (2006). « Explorer l'espace des mots : du linéaire au non-linéaire », *in* J.M. Viprey (éd.), *JADT 06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. Besançon : Presses universitaires de Franche-Comté, vol. 2, pp. 593-600.

Longrée D. & Luong X. (2003). « Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés », *Corpus* 2 : 119-140.

Longrée D., Luong X. & Mellet S. (2004). « Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé », *in* G. Purnelle, C. Fairon & A. Dister (éds), *JADT 2004, 7èmes Journées internationales d'Analyse statistique des Données Textuelles*. UCL : Presses universitaires de Louvain, vol. 2, pp. 743-752.

Longrée D.& Mellet S. (sous presse). « Temps verbaux et prose historique latine : à la recherche de nouvelles méthodes d'analyse statistique », *in Actes du 13ème Colloque international de Linguistique latine* (ICLL 13, Bruxelles 2005).

Longrée D., Mellet S. & Luong X. (2006). « Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées », *in* J.M. Viprey (éd.), *JADT 06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. Besançon : Presses universitaires de Franche-Comté, vol. 2, pp. 643-654.

Luong X., Juillard M., Mellet S. & Longrée D. (2007). « The Concept of Text Topology. Some Applications to Verb-Form Distributions in Language Corpora », *Literary and Linguistic Computing* 22,2 : 167-186.

Juillard M. & Luong X. (1989). « Unrooted Tree Revisited : Topology and Poetic Data », *Computers and the Humanities* 23 : 215-225.

Juillard M. (1998). « Les lexèmes dans l'espace du texte : analyses arborées et bases de voisinage », *Cycnos* 15 : 57-75.

Mayaffre D. (2007a). « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », *in* F. Rastier et M. Ballabriga (éds) *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*. Toulouse : Put, pp. 15-26 (à lire aussi sur Texto ! Textes et cultures, http://www.revue-texto.net/Archives/ Archives.html).

Mayaffre D. (2007b). « L'analyse des données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle », *Lexicometrica* 7 « Topographie et topologie textuelles ».

Piérard S., Degand L. & Bestgen Y. (2004). « Vers une recherche automatique des marqueurs de la segmentation des discours », », *in* G. Purnelle, C. Fairon & A. Dister (éds), *JADT 2004, 7èmes Journées internationales d'Analyse statistique des Données Textuelles*. UCL : Presses universitaires de Louvain, vol. 2, pp. 859-864.

Piérard S. & Bestgen Y. (2006). « A la pêche aux marqueurs linguistiques de la structure des discours », *in* J.M. Viprey (éd.), *JADT 06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. Besançon : Presses universitaires de Franche-Comté, vol. 2, pp. 749-758.

Sérant D. & Thoiron Ph. (1988). « Topographie des formes répétées », *Revue Informatique et Statistique dans les Sciences humaines* 24 : 333-343 (*Le nombre et le texte : Hommage à Étienne Évrard*).

Viprey J.-M. (2006-a). « Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection 'géodésique' », *in* J.M. Viprey (éd.), *JADT 06, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*. Besançon : Presses universitaires de Franche-Comté, vol. 2, pp. 989-1000.

Viprey J.-M. (2006-b). « Structure non séquentielle des textes », *Langages* 163 : 71-85.