Modélisation textométrique des textes

Bénédicte Pincemin

In Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles *(JADT 2008), Lyon, 12-14 mars 2008*
Serge Heiden & Bénédicte Pincemin (éds), Lyon : Presses Universitaires de Lyon, ISBN 978-2-7297-0810-8
vol. II, pp. 949-960.

Modélisation textométrique des textes

Bénédicte Pincemin[1]

[1]CNRS & ICAR, Université de Lyon – ENS-LSH – 15 parvis René Descartes – B.P.7000 – F69342 Lyon cedex 07 – France

Abstract

This contribution analyses the data model for textometry (for calculations like the ones in textual statistics (Lebart & Salem, 1994)). It also presents a review of the textual representations proposed by textometric software. Taking into account a linguistic point of view (especially Rastier's textual semantics), it then points out the limits of these representations, and explores new propositions for textual modelization in textometry.

Résumé

Les statistiques textuelles (ou textométrie) exploitent une représentation du texte sous forme d'une suite d'unités typées, éventuellement réparties entre des subdivisions du corpus. Pour rendre compte et tirer parti de la multiplicité des typages possibles, des logiciels ont mis au point des représentations tabulaires du texte, claires et très efficaces pour la recherche de motifs complexes. Mais les délimitations des unités et des partitions, et la catégorisation des unités, sont encore peu souples, car fortement prédéfinies. De nouvelles modélisations seraient à élaborer, qui élargiraient le rôle accordé aux propriétés, distingueraient des contextes locaux (passages) et globaux (entités), et s'associeraient avec des calculs textométriques adaptés à la caractérisation de corpus structurés, aux unités non uniformes. Nous proposons alors une schématisation des étapes d'un calcul textométrique, qui explicite les multiples sélections en jeu (niveaux de corpus, fond, forme, dimensions de lecture, d'analyse et d'affichage). Puis nous étudions la récente modélisation du texte comme trame et soulignons comment elle innove en relativisant le découpage en formes graphiques ; nous concluons en ébauchant une modélisation du texte comme trace, davantage en accord avec la sémantique différentielle et interprétative de Rastier.

Mots-clés : textométrie, lexicométrie, statistique textuelle, modèle de données, corpus étiqueté, annotation, trame textométrique, sémantique différentielle, sémantique interprétative.

1. Problématique

La définition et la caractérisation de la textualité restent des questions ouvertes en linguistique. En revanche, par les calculs qu'elle propose et par les outils qui l'instrumentent, la textométrie (dans la lignée de la statistique textuelle (Lebart & Salem, 1994)) est amenée à expliciter des représentations opératoires du texte. Notre objectif ici est de faire le point des modélisations actuelles et des nouvelles propositions, avec une préoccupation à la fois linguistique (trouver une modélisation en accord avec certains aspects fondamentaux, notamment de sémantique interprétative au sens de Rastier (1994, 2001)) et technique (expliciter, formaliser, envisager les conditions d'implémentabilité des propositions).

2. Représentations actuelles du texte en textométrie

2.1. Théorie textométrique : les éléments mobilisés par les calculs

Considérons quels sont les objets mobilisés par les traitements textométriques classiques (concordances, calcul de spécificités, etc.), autrement dit de quoi ont-ils besoin exactement

pour pouvoir lancer leurs calculs. Pour un traitement textométrique, le modèle de données est une succession d'unités (par exemple des mots). Chaque unité est une occurrence d'un type (ce qui permet de définir le vocabulaire –c'est-à-dire l'ensemble des différents types- utilisé dans le corpus, et la fréquence d'apparition de chaque type). Enfin, pour certains calculs, le modèle textuel est subdivisé en parties (corpus dit *partitionné*), de telle sorte que chaque unité relève d'une partie et d'une seule. Bref, « il y a des choses [unités] que l'on compte [grâce au typage] dans des trucs [parties]. »[1] L'exploitation textométrique d'un corpus suppose donc au moins deux procédures incontournables : *segmentation* en unités, et *identification* des types sous-jacents (Lebart & Salem, 1994).[2]

Ceci étant, un même corpus peut se prêter à différents découpages. Très souvent, on considère une multiplicité de partitions : le corpus est tantôt vu comme un ensemble de textes, comme une succession de périodes, ou comme les écrits de différents auteurs ou de différentes sources, par exemple. Grâce aux outils d'analyse linguistique, il est également très courant de disposer d'un corpus découpé en unités étiquetées par différentes informations, par exemple forme fléchie, lemme, catégorie morphosyntaxique. Etant donnée cette multiplicité des partitions et des types, le corpus est alors pris en charge comme une matrice à modèles textuels (ou vues textométriques), obtenus typiquement par projection : on retient un domaine (certaines zones du corpus peuvent être exclues du calcul) et éventuellement un découpage en parties, également un découpage en unités et un typage sur ces unités, les autres informations apportées par le corpus restant dans l'ombre pour le calcul.

2.2. Formalisations proposées par les logiciels

Les logiciels de textométrie doivent donc proposer une représentation du corpus qui articule cette multiplicité de partitionnements et de typages au modèle de données textométrique. Deux logiciels développent ainsi, dans leur manuel, une telle modélisation du corpus : Weblex, qui s'appuie lui-même sur le modèle très clair et efficace proposé par CQP[3] ; et SATO, qui, travaillant tout particulièrement au niveau de l'étiquetage (exploitation mais aussi transformation et ajout d'étiquettes en cours de traitement), a développé une représentation du texte originale et suggestive.

2.2.1. Le texte vu par Weblex

Le texte est considéré comme un tableau, qui croise, comme en un quadrillage, les dimensions syntagmatiques et paradigmatiques : chaque colonne correspond à une position (la suite des positions étant issue de la segmentation initiale du corpus), et les lignes décrivent des propriétés. Chaque position du texte reçoit donc une description sous forme d'une série de valeurs de propriétés, issues des étiquetages exploitables et pertinents dans le corpus.

[1] Ce bon mot, que nous devons à André Salem, abstrait donc, en une formule mnémonique simple, les structures fondamentales utilisées par les calculs textométriques.

[2] Quand nous considérons que le modèle de données est une suite d'unités typées éventuellement partitionnées, nous nous focalisons sur les structures mobilisées par les calculs de synthèses, sans exclure que d'autres informations, notamment d'*édition* (mise en page, typographie, etc.), soient également attendues et exploitées par les logiciels (pour l'affichage du texte dans l'indispensable retour au texte). De même, en pratique, la médiation entre le modèle de données et l'expression (numérique) du corpus est assurée par des *formats*, tant pour la reconnaissance des caractères que pour l'identification des découpages et des types (cf. les premières rubriques d'import dans (Heiden, 2006)). Nous n'abordons pas cette question ici, bien qu'en pratique elle corresponde à une étape incontournable pour l'intégration de données dans un logiciel textométrique.

[3] Weblex ne considère qu'une vue partielle sur la représentation WTC construite et manipulée par CQP.

Propriétés ↓	*Déroulement syntagmatique du texte (positions)* →		
forme	**ouvrirait**	**les**	**vannes**
catégorie	verbe	article	substantif
lemme	ouvrir	le	vanne
…	…	…	…

Figure 1 : Exemple de représentation textuelle dans Weblex (d'après (Heiden, 2002))

Ainsi, au moment de lancer un calcul, les différentes informations requises sont disponibles de la façon suivante : le découpage en unités est donné par la suite des positions, et un paramètre indique quelle propriété (ligne du tableau) sert pour définir le regroupement des occurrences en types.

En ce qui concerne les partitionnements, ceux-ci sont définis une fois pour toutes au moment de l'intégration du corpus dans l'outil. Le choix d'une partition (ou d'aucune) se traduit alors tout simplement par le choix d'une vue du corpus. Au plan de la modélisation, les partitions (comme les références de localisation) sont construites à partir des indications de contexte que l'on a choisi d'enregistrer, et qui se traduisent elles aussi de façon distribuée sous forme de propriétés attachées aux positions[4]. Il y a par ailleurs la possibilité de prédéfinir des segmentations locales, par exemple la délimitation de phrases, mobilisables pour la définition de voisinages (contrôle de la portée de recherches de motif, cooccurrences,…)[5].

2.2.2. Le texte vu par SATO

SATO modélise le texte comme un ensemble de points dans un plan, dit *plan lexique/occurrences*. Il se base sur un espace à deux dimensions, linguistiquement assimilables à des axes paradigmatique et syntagmatique. En effet, l'une de ces dimensions (conventionnellement représentée comme l'axe vertical) est l'axe lexical : cette dimension dresse la liste du vocabulaire (mots, formes lexicales) utilisé dans le texte. L'autre dimension (conventionnellement représentée comme l'axe horizontal, orienté de gauche à droite) représente l'ordre séquentiel de lecture, la linéarité du texte qui se donne comme une suite d'occurrences des formes lexicales.

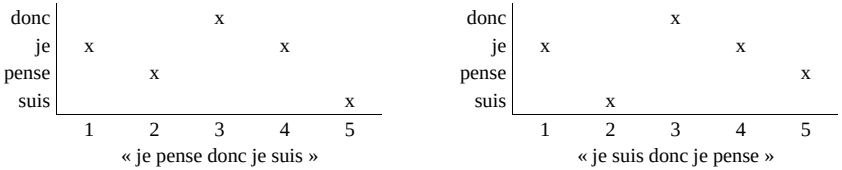

Figure 2 : Exemples de représentations textuelles dans SATO : un même lexique, deux textes (Daoust 2007)

Les propriétés permettant de varier les typages et les découpages ont alors deux natures possibles : propriété *lexicale* – qui annote un mot hors contexte, et est associée à l'axe

[4] Ceci permet la construction de partitionnements imbriqués, en croisant plusieurs critères successivement.

[5] Nous verrons plus loin (§3.2) que ces deux modes de définition de contextes pourraient s'apparenter à deux formes de contextualisation : globale (*entités*) et locale (*passages*).

lexical -, ou propriété *textuelle* – qui annote une occurrence, et est associée à l'axe séquentiel[6]. La représentation du texte et de ses annotations correspond alors non seulement à un ensemble de points dans le plan lexique/occurrences, mais aussi à deux tableaux superposant un certain nombre de propriétés aux positions sur chaque axe.

Fréqtot	*Gramr*						
1	Con	**donc**			x		
2	Proper	**je**	x			x	
1	Vconj	**pense**		x			
2	Vconj	**suis**					x
			1	**2**	**3**	**4**	**5**
		Edition	maj	nil	cap	nil	nil
		Partie	prém	prém	conn	conc	conc

« [Je pense]$_{prémisse}$ [DONC]$_{connecteur}$ [je suis]$_{conclusion}$ »

Figure 3 : Exemple de représentation textuelle dans SATO : texte et propriétés (Daoust, 2007)

En ce qui concerne le découpage du corpus en parties, SATO propose une procédure segmentation basée soit sur le découpage préalable du corpus en documents et pages, soit sur une longueur fixée en nombre de mots, soit sur la reconnaissance d'un délimiteur (correspondant à une unité) repéré par un filtre (typiquement une ponctuation forte, mais le filtre permet aussi de désigner ou combiner des valeurs de propriété par exemple). Il existe également un autre mode de morcellement du corpus, par la sélection de l'ensemble des contextes d'un motif.

2.3. Interprétation des modélisations actuelles

2.3.1. Unités : un seul découpage

Que ce soit dans Weblex ou dans SATO[7], le découpage en unités est fixe et unique. Si l'analyse conduit à vouloir modifier le découpage de tout ou partie des unités, il faut soumettre un nouveau corpus au logiciel (ou/et modifier le paramétrage d'entrée), et remplacer l'ancien découpage par le nouveau. Si l'on veut pouvoir travailler sur plusieurs découpages d'unités en alternance, et varier les découpages, il faut créer plusieurs bases et les consulter parallèlement.

Cependant, la possibilité de considérer d'autres découpages que ceux initialement dans le corpus semble un besoin très naturel pour le travail sur corpus : en pratique, les découpages, souvent générés automatiquement, ne sont pas exempts d'erreurs ; et en théorie, il est admis (en sémantique interprétative par exemple) que les unités ne sont de toutes façons pas définissables préalablement et une fois pour toutes, elles sont construites par l'analyse et toujours remodelables, selon le contexte notamment.

[6] Notons qu'une propriété textuelle de SATO n'est pas en soi syntagmatique ; il suffit qu'elle ne soit pas compatible avec la base lexicale, qui sert d'ancrage à toutes les propriétés enregistrées de façon paradigmatique.

[7] Et cela se vérifie aussi pour d'autres logiciels se réclamant de la textométrie, tels que Lexico3 ou Hyperbase.

2.3.2. Typages : exploitation souple pour la recherche de motif (moteur de recherche), mais rigidité de la catégorisation pour les calculs statistiques

Les modélisations actuelles permettent une multiplicité de typages des unités. Weblex offre de travailler sur le corpus à partir d'une dizaine de propriétés différentes. Pour le repérage d'un motif dans le corpus (fonction de moteur de recherche), elles sont interrogeables de façon très souple : possibilité de combiner plusieurs propriétés, possibilité aussi de faire des sélections de type expression régulière sur les valeurs de propriétés, considérées et manipulées comme des chaînes de caractères. En revanche, pour le volet statistique, on ne travaille que sur une propriété à la fois, avec les regroupements prédéfinis par sa gamme de valeurs.

Les TGEN (types généralisés) implémentés dans Lexico 3 (Lamalle & al., 2006), et les topes proposés par Salem, visent à assouplir et généraliser le typage des unités[8]. Concrètement, ils servent à regrouper comme plusieurs occurrences d'un type unique et original des occurrences sinon dispersées dans des types différents. Ces modes de sélection seraient-ils une manière de définir dynamiquement de multiples typages, au lieu de recourir par exemple à des codages en propriétés ? Pas vraiment, car ils construisent un seul type, alors qu'une propriété définit un paradigme de types. Pour des calculs statistiques d'ensemble, le type généralisé ou le tope ne retravaille la distribution que d'un type (nouveau) ; alors que les propriétés permettent d'examiner des répartitions d'ensemble sur des systèmes de types différents.

SATO, dont la spécialité est le travail sur les annotations, a l'originalité de distinguer les types associés aux occurrences, et les types associés au lexique, ce qui effectue en quelque sorte une factorisation de l'information quand cela est pertinent. SATO prend en compte également diverses natures de propriétés, auxquelles sont associés des traitements et opérations différenciés et adaptés : numérique (entier), symbolique (énumérée), ensemble (les valeurs sont des ensembles de symboles), chaînes de caractères.

Toutefois, nous observons que nous n'avons pas pleinement la souplesse qu'introduirait la distinction entre *dimension élémentaire* et *dimension descriptive* (cf. §4.1). Actuellement, pour les calculs textométriques, les unités ne sont regroupables en types que selon les valeurs d'une propriété déjà associée aux occurrences (éventuellement via le lexique). Or il serait envisageable que l'analyste puisse définir plus finement les regroupements souhaités en combinant les valeurs de plusieurs propriétés (dans un corpus où, compte-tenu de cette possibilité, on a pris soin de faire un codage analytique qui décompose les informations en propriétés élémentaires). Dans SATO, on peut obtenir un comportement approchant : la dimension d'analyse ne serait pas générée à la volée au moment du calcul, mais elle pourrait être matérialisée par une nouvelle propriété calculée à partir des propriétés existantes.

[8] Un *type généralisé* se présente concrètement comme une sélection sur les types prédéfinis (par un ou plusieurs filtres sous forme d'expression régulière par exemple, et éventuellement retaillée en excluant certaines formes). Le regroupement opéré par type généralisé opère *en compréhension*, *paradigmatiquement*, via les types prédéfinis du corpus. Un *tope* est une sélection directe sur les occurrences, à même la *syntagmatique* du texte, non médiée (factorisée) par les types. Les topes procèdent donc quant à eux *en extension*, par désignation directe des occurrences à regrouper. Notons qu'un tope pourrait être rendu formellement équivalent à un type en définissant une propriété sélectionnant les mêmes occurrences. On pourrait donc interpréter la distinction entre types et topes non pas comme une différence fondamentale de nature, mais comme une différence d'intégration dans un système paradigmatique. Salem s'oriente d'ailleurs vers une unification de ces deux opérations, qu'il propose d'appeler *sélection*.

2.3.3. Partitions : faiblement dynamiques et peu liées à l'étiquetage en propriétés

Alors que les modèles textométriques du texte intègrent la multiplicité des propriétés (comme autant de vues sur le texte), ils utilisent peu ou pas ces informations pour la répartition des unités entre plusieurs parties[9]. Les partitions apparaissent fixées préalablement et généralement en exploitant des propriétés supratextuelles (texte, auteur, genre, date de publication, etc.), alors que les propriétés au niveau des occurrences semblent usuellement utilisées pour des informations infra-textuelles.

Si les logiciels tirent parti du codage du corpus pour définir des partitions, ils le font de façon quelquefois très fine (Xaira, Weblex, tirent parti de l'infoset XML), mais pour une exploitation ensuite toujours relativement statique et prédéfinie. Lexico 3 ou Xaira permettent d'ajuster une partition par élimination et par regroupement de parties : cela s'avère déjà extrêmement utile en pratique, même si l'on voit bien que l'on reste cependant sur la base de quelques découpages prévus *a priori*.

Il y a là quelque chose de décevant : alors qu'on dispose d'une très grande richesse potentielle d'information *via* les propriétés associées aux unités, cette information semble encore faiblement mise à profit pour composer des partitions au fil de l'analyse. Le découpage à la volée basé sur l'annotation proposé par SATO se fait sur la reconnaissance d'unités délimitatrices (qui peuvent être identifiées par une condition (filtre) sur leur forme graphique ou les valeurs de propriétés associées). Cependant, cette procédure définit moins des parties que des frontières : il s'agit d'un découpage du corpus en tranches (selon son déroulement linéaire), qui ne donne pas accès à une subdivision du corpus en parties non connexes[10].

Cependant, dans quasiment tous ces logiciels, la définition des partitions au moment de l'entrée du corpus dans le logiciel est liée à la nécessité de la précompilation d'index, requis pour l'efficacité des calculs statistiques élaborés (par ex. calcul des spécificités par exemple) sur de gros corpus (centaines de millions d'occurrences). Il s'agirait de voir si cette contrainte est contingente à la puissance des machines, auquel cas on pourrait envisager d'avoir des partitions dynamiques (découpage de base non fixé *a priori*) à plus ou moins long terme, ou s'il y a un obstacle algorithmique fort[11]. Entrent aussi en considération des contraintes d'implémentation, notamment le choix éventuel d'une interface web, et la volumétrie des corpus.

[9] Nous considérons ici la représentation du corpus pour des calculs de distribution des unités et de caractérisation de différentes parties. La possibilité de constituer un sous-corpus par sélection d'une partie du corpus n'est pas en jeu ici, lorsqu'il s'agit simplement de focaliser l'espace de recherche, et non de contraster une partie sur un tout et surtout par rapport à d'autres parties.

[10] C'est la sélection (par filtrage) d'un *sous-texte* qui permet de travailler sur la portion du corpus réalisant certaines valeurs d'une propriété. Cependant, pour le moment, cette procédure de SATO relève davantage d'une démarche de focalisation que de contraste, elle construit un extrait mais non un système de parties se prêtant directement à l'observation de répartitions. Il reste possible néanmoins de faire certaines mesures tour à tour sur différents sous-textes formant système, de les enregistrer dans des propriétés, pour pouvoir ensuite les recueillir pour une étude contrastive d'ensemble.

[11] SATO, qui privilégie la souplesse des découpages en parties et le remaniement constant du corpus via l'utilisation et l'ajout de propriétés, fait le choix de renoncer presque totalement aux index. Il démontre donc un niveau de faisabilité actuel pour des partitionnements dynamiques. Ceci étant, SATO n'intègre pas certains calculs statistiques réputés lourds.

3. Pistes de recherche

3.1. Elargir encore le rôle donné aux propriétés et unifier ainsi la modélisation du fonctionnement textuel

Une plus grande souplesse pourrait être donnée en unifiant et généralisant le rôle des propriétés. Actuellement on privilégie encore l'usage des propriétés pour rendre compte d'informations infra-textuelles, et pour typer les unités. Or il apparaît que les propriétés pourraient intégrer une vision davantage unifiée du fonctionnement à la fois local et global des textes : ne pas cantonner les informations infra-textuelles (respectivement supra-textuelles) à la caractérisation des unités (respectivement des parties), utiliser les propriétés non seulement pour typer mais aussi pour segmenter (en unités et en parties). En effet, si nous revenons au modèle de données textométrique, il a déjà été remarqué que parties et unités se comportent comme des rôles, de contenant ou de contenu, dévolus à des segments textuels. Autrement dit, il n'y a pas d'unités ou de parties par nature, mais par fonction, et ce qui est partie (contenant) dans une analyse peut devenir unité (contenu) dans une autre, par un simple changement de granularité de l'analyse. Segmentation en unités typées, et délimitation d'une partition, sont alors simplement deux manières de voir une même réalité.

Concrètement, l'annotation en propriétés, plutôt que d'être utilisée soit pour la segmentation en parties, soit pour le typage d'unités, pourrait de même être généralisée à la segmentation comme à l'identification de contenants comme de contenus. La définition d'un contenant comme d'un contenu mobiliserait alors deux propriétés (ou deux combinaisons de propriétés) : l'une pour donner l'empan (en groupant des unités du découpage élémentaire de base), et l'autre pour grouper ces unités (sous forme de répétition pour le contenu –les occurrences sont distinguées-, et sous forme de fusion pour le contenant –les segments de même valeur sont fondus en une partie)[12].

3.2. Affinement des rôles. Deux formes de contenant : passages et entités

Le modèle textométrique, tel que nous l'avons résumé initialement (§2.1), ne rend pas pleinement compte des distinctions opérées au moment des analyses textométriques. Nous avons une structuration en deux rôles (contenants *vs* contenus), mais il y a peut-être lieu de dédoubler le rôle de contenant et d'introduire une contextualisation intermédiaire. Pour situer les *unités*, on distinguerait ainsi des contextes locaux (ou *passages*) et des contextes globaux (ou *entités*). Les passages fonctionnent comme des fenêtres, comme l'entour d'un motif, comme un voisinage ; ils sont relatifs aux unités, et ne sont pas nécessairement considérés pour eux-mêmes ni en tant qu'ensemble, ils ne forment pas nécessairement une partition. Les entités ont à l'inverse, dans le rôle qu'on leur fait jouer, une véritable consistance : ce sont des contenants nommés, répertoriés, représentant une structuration globale du corpus d'étude. On peut également avoir une seule entité (corpus distingué) que l'on oppose au « reste » (corpus de référence) au sein du corpus d'étude (cf. §4.1).

[12] Plus généralement, les propriétés pourraient ne pas directement calquer les groupements et découpages potentiels, mais être interprétées par une fonction (au sens mathématique), qui n'est pas nécessairement la fonction identité. Ainsi, pour la propriété (ou la combinaison de propriétés) délimitante, on se donne une fonction qui, à la valeur de la propriété, associe une unité syntagmatique –voire plusieurs, pour d'éventuels recouvrements ; et pour la propriété (ou la combinaison de propriété) identifiante, on se donne une fonction qui, à un ensemble de valeurs, associe un type paradigmatique –voire plusieurs, si polyvalence / plurivocité.

Notons qu'à ces trois « niveaux » (unité, passage, entité) sont associées des références de localisation (explicitement ou implicitement). Au niveau de l'unité, la localisation sert à qualifier les positionnements relatifs immédiats, par exemple l'opposition avant / après. Au niveau du passage, on matérialise l'opposition du proche et du lointain. Au niveau de l'entité, on a une traduction de l'opposition intérieur vs extérieur (dans *vs* hors).

Notons également que la distinction entre passage et entité est une distinction de fonction (rôle) plus que de nature. En particulier, elle ne correspond pas nécessairement à une opposition entre infra-textuel et supra-textuel. Par exemple, sur un corpus de numéros de journeaux, la contextualisation d'un numéro qui engloberait les trois numéros précédents serait de type passage. Et inversement, sur un corpus de pièces de théâtre d'un auteur, une contextualisation par personnage correspondrait à des entités tout en étant infra-textuelle.

Les procédures textométriques ne mobilisent pas toutes les deux niveaux de contexte, et c'est ce qui fait que cette distinction peut rester inaperçue. Un calcul de cooccurences fait intervenir une contextualisation locale, dans la mesure où ces contenants ne sont pas considérés en eux-mêmes, mais ne servent qu'à définir des voisinages. Inversement, un calcul de spécificités considère des contextes entités fortement individualisés. Mais un (bon) calcul de concordance articule les deux niveaux de contexte, local et global : l'unité est le motif recherché, servant de pivot ; la ligne de concordance est un passage ; la référence associée à la ligne et permettant de situer l'extrait dans le corpus, de trier et regrouper les lignes selon l'organisation globale du corpus, nomme et individualise des entités.

3.3. Théorie textométrique et corpus structurés

Les principaux calculs textométriques ont été mis au point à une époque où les éditions numériques étaient peu structurées. L'état de l'art actuel de formatage des corpus, notamment avec XML (et les recommandations de la TEI), apporte des richesses nouvelles, et l'enjeu est d'adapter ou d'étendre les techniques textométriques pour une exploitation des corpus plus fine et plus complète.

Les éditions numériques ont notamment fait évoluer la représentation des textes sous deux aspects liés au modèle de données textométrique : d'une part, l'étiquetage des unités (de tous ordres, mais notamment lexicales), que la textométrie a commencé à intégrer dans ses modélisations et procédures, notamment via les propriétés attribuées aux unités (Pincemin, 2004). D'autre part, les corpus XML font la part belle à une segmentation non uniforme des unités, et à leurs imbrications : le texte XML manifeste plus directement des segmentations en unités de tailles inégales, il code également des motifs non seulement syntagmatiques mais aussi hiérarchiques, par exemple des régularités d'emboîtement d'un élément dans un autre. Ce chantier de développement textométrique est encore complètement ouvert. Il n'est pas du tout sûr que les techniques traditionnelles, même astucieusement appliquées (par exemple, faire des segments répétés sur des motifs d'emboîtements au lieu de motifs de succession), permettent de bien caractériser les nouvelles formes de régularités et de contrastes liées à ces nouvelles structures. Autrement dit, le modèle textométrique du texte, pour le moment linéaire et uniforme, pourrait être amené à évoluer par la conception de nouveaux calculs adaptés à ces corpus structurés.

Par exemple, considérer pleinement des unités avec une certaine étendue suppose que le modèle de données sache représenter deux unités successives (voire chevauchantes) de même type en les distinguant : cela implique une représentation globale de l'unité, individualisée avec un début et une fin, voire des composantes non connexes, qui n'est pas réductible à une

vision locale, où la nature de l'unité étendue serait distribuée comme une propriété attachée aux unités élémentaires sous-jacentes. Ou bien, il faut mobiliser deux propriétés locales, l'une pour la délimitation et composition des unités, l'autre pour leur typage.

4. Propositions

4.1. Eléments de terminologie

Le texte (le corpus) se définit selon diverses vues, que nous avons vu jusqu'à présent codées par des *propriétés* attachées à des unités (syntagmatiques –positions, ou/et paradigmatiques). Pour la souplesse des analyses, nous avons proposé (Pincemin 2004) que ces propriétés enregistrent des informations élémentaires, combinables ensuite pour construire les différentes catégories souhaitées (au lieu de ne pouvoir travailler que sur des propriétés prédéfinies). Le corpus entré dans un logiciel de textométrie serait alors enregistré avec une caractérisation selon un certain nombre de *dimensions élémentaires*. Les *dimensions descriptives*[13] qui organisent les données pour les calculs (notamment le regroupement d'occurrences en types) sont ensuite définissables dynamiquement, à partir des dimensions élémentaires.

Une propriété est donc une dimension (dimension élémentaire si elle peut être combinée avec d'autres, dimension descriptive si elle n'est utilisable que directement). Une dimension (dimension élémentaire ou dimension descriptive) est donc un *jeu structuré* (ou *système*) de *valeurs*. Une telle valeur correspond à un *type*, au sens lexicométrique qui l'oppose à *occurrence*. Les valeurs sont des catégories, associables à des unités, et sur la base desquelles ces unités sont assimilables à des manifestations d'un même objet (autrement dit, la valeur – ou type textométrique – permet de grouper des occurrences en les déclarant comme répétition d'un même type).

Cet usage du mot *type* en textométrie, dans l'opposition *type* / *occurrence*, interfère avec l'usage informatique, qui nous est pourtant aussi utile ici. Le *type* informatique définit la structure et le comportement d'un objet, il fixe les opérations qui lui sont applicables. Ici, nous souhaiterions de même considérer les dimensions comme typées[14] : chaque dimension a un domaine de valeurs qu'elle structure d'une certaine manière, et sur lesquelles sont possibles certaines opérations appropriées. En particulier, il est très utile de disposer d'un ordre canonique selon lequel trier les valeurs de façon significative.

Ainsi, une dimension qui décrit la linéarité du texte, par une succession de positions, a une structure syntagmatique, qui définit par exemple une relation d'ordre, un avant et un après. Une dimension qui répertorie des étiquettes de parties du discours peut avoir une structure ensembliste, voire même arborescente, avec une organisation en catégories et sous-catégories. Les types numériques peuvent être très différents selon qu'ils expriment par exemple des entiers, des réels (induisant une structure continue), ou des rangs. D'autres structures sont possibles, par exemple des structures de traits, pour une dimension qui s'intéresserait au repérage d'un certain motif linguistique par exemple. Il est vrai cependant que le choix de

[13] Dans (Pincemin 2004) ces *dimensions descriptives* étaient appelées *dimensions d'analyse*, mais cela a l'inconvénient de confondre sous une même appellation deux opérations, d'une part l'élaboration de dimensions (dites maintenant *descriptives*) à partir de dimensions élémentaires, et d'autre part le choix de la dimension (toujours dite *d'analyse*) fixant le typage des unités.

[14] Ce que fait déjà SATO, pour lequel les propriétés sont typées.

coder des dimensions élémentaires tend à structurer les dimensions initiales de façon simple, les structures plus élaborées revenant alors le cas échéant aux dimensions descriptives.

Il faut aussi expliciter les articulations entre les données disponibles et celles qui sont mobilisées. Nous proposons de distinguer les rôles suivants : *(i)* le *corpus existant*, ou *base*, qui correspond à l'ensemble des données disponibles, une fois entré dans le logiciel le corpus constitué selon les objectifs que l'on s'est donnés ; *(ii)* le *corpus d'étude* (visualisé, englobant) : partie du *corpus existant* constituant le terrain sur lequel on veut mener une analyse, une série de calculs ; *(iii)* le *corpus de référence* : corpus utilisé comme modèle de répartition des unités pour les statistiques contrastives -il correspond souvent au corpus d'étude ; *(iv)* le *corpus distingué* (focalisation, filtre) : partie du corpus d'étude sur laquelle on centre l'analyse, en cherchant à la caractériser par rapport à l'ensemble formé par le corpus d'étude (sinon, si l'on choisit de perdre de vue la mise en perspective par rapport à un corpus d'étude engobant, c'est que l'on définit un nouveau corpus d'étude).

L'analyse textométrique s'appuie sur les dimensions et doit faire des choix pour plusieurs moments du traitement (Pincemin 2004), outre la détermination des différents corpus : *(i)* sélection d'un *fond*, à savoir des unités prises en considération dans les calculs ; *(ii)* sélection éventuelle d'une *forme*, pour un calcul focalisé (cette sélection peut se faire en compréhension, par filtrage sur des valeurs, ou par désignation directe) ; *(iii)* détermination d'une *dimension de lecture*, qui soit l'espace d'observation des *occurrences* ; *(iv)* détermination d'une *dimension d'analyse*, dont les valeurs sont les *types* catégorisant les occurrences ; *(v)* détermination éventuelle de *dimensions d'affichage* complémentaires pour la visualisation des résultats. Les sélections du fond et de la forme peuvent s'appuyer sur plusieurs dimensions, indépendamment des dimensions de lecture, d'analyse et d'affichage.

4.2. La trame : un référentiel arbitraire servant d'ancrage et de lien entre de multiples vues

La notion de trame textométrique a été proposée récemment par André Salem et Serge Fleury, et est expérimentée dans l'outil *Trameur* (Fleury 2007). Au lieu de faire de la segmentation initiale du texte en formes graphiques (mots) la base de toutes les annotations (de typage et de segmentation : propriétés, partitions), il s'agit de se donner une trame suffisamment fine, et sans contenu propre (donc éventuellement modifiable) à laquelle rapporter toutes les informations (les différentes dimensions), y compris donc la segmentation en formes graphiques. La trame se présente typiquement comme une suite de positions. N'étant pas nécessairement liée à une analyse particulière (par exemple à une segmentation en forme graphiques), la trame peut être ajustée pour devenir un « dénominateur commun » aux analyses de plusieurs logiciels, et donc permettre des échanges de données (import et export de segmentations, d'annotations, etc.). C'est clairement un des objectifs majeurs de cette nouvelle proposition de modélisation.

Il est intéressant de noter que d'une part cette trame, réduite en quelque sorte à être un système de coordonnées, est un artéfact support et non un objet linguistique : elle n'a aucunement vocation à condenser un savoir sur la textualité, ce qui la rend particulièrement adaptable. D'autre part, la trame confère un rôle symétrique à toutes les dimensions, et exprime leur relativité. Autrement dit, un texte n'est plus une suite de mots (vue comme le socle incontournable) enrichie d'annotations : c'est un faisceau de représentations synchronisées par une trame textométrique qui ne fait que traduire une cadence commune, elle-même non fondamentale.

4.3. La trace : le texte comme déploiement

Par construction, du fait de son rôle central, la trame textométrique que nous venons de présenter rapporte toutes les dimensions à un référentiel commun, qui les rend donc comparables. Or nous pouvons pressentir que la réalité textuelle ne se plie pas naturellement à cette contrainte : rien ne dit que tous les points de vue sur le texte, toutes les manières de le décrire, s'articuleraient tous les uns par rapport aux autres, et que l'on pourrait toujours trouver une perspective commune unique.

Une représentation davantage en accord avec une conception différentielle et interprétative du texte (Rastier 2001) semble suggérée par la proposition de SATO, qui fait du texte une trace dans un plan lexique/occurrences. La différence fondamentale de conception avec la trame, c'est que le texte se dessine à la croisée de deux dimensions, plutôt que d'être en tous ses aspects ancré à un socle. Autrement dit, au texte-à-trame s'oppose le texte comme trace.

Cependant la proposition de SATO restait entièrement quadrillée par deux dimensions ainsi privilégiées : une segmentation (textuelle) en formes graphiques et un recensement (lexical) de ces formes graphiques[15]. Elle n'a pas ce recul par rapport à la relativité des descriptions, et illustré par le modèle de la trame textométrique. Nous voulons donc ébaucher ici une généralisation du texte comme trace qui intègre la multiplicité et la relativité des dimensions.

Un texte, plongé dans un espace de description, est une trace, une forme, une trajectoire, un déploiement. Sa significativité explique d'ailleurs qu'il soit reconnu comme une forme, par opposition à une masse informe. Nous ne le saisissons jamais totalement –l'espace dans lequel nous l'observons n'épuise pas toutes ses facettes-, et nous ne travaillons guère que sur son ombre (sa projection) dans un espace de lecture. Le corpus textuel se déploie donc à la croisée de ses caractérisations selon différents points de vue, concrètement des systèmes de description, traduits dans des dimensions. Une dimension est un domaine de valeurs structuré et typé. La surface d'une unité est mesurable à l'aune d'une dimension par rapport à laquelle elle est définissable.

En effet, certaines dimensions peuvent être orthogonales, au sens où il n'y a pas d'interrelation permettant de passer de l'une à l'autre. Il n'est peut-être pas pertinent d'entre-définir toutes les dimensions. Le déploiement du texte peut se faire en composantes non connexes. Il s'agit ici de rendre compte de phénomènes d'indétermination réciproque. C'est en ce sens que le texte n'est plus *à la croisée* de descriptions finalement liées en faisceau, mais se déploie *entre* les dimensions descriptives. La sémantique différentielle explique en effet que le sens n'est pas *dans* les unités ou les dimensions, mais il se crée *entre* elles, par rapprochements et contrastes.[16] Certaines dimensions sont de bons points d'entrée : par exemple, la linéarisation syntagmatique du texte est souvent une référence claire, même si elle n'est ni supérieure en soi, ni complète : il y a certains aspects du texte qui sont mal voire pas décrits selon cette linéarisation, par exemple des découpages plus fins, ou des phénomènes sémantiques diffus.

[15] Le plan lexique/occurrences est le croisement d'*un* découpage syntagmatique avec *un* système de catégories paradigmatiques ; les autres propriétés, pour pouvoir se greffer à la représentation, doivent être isomorphes.

[16] Rastier (2005) observe dans les écrits autographes de Saussure que le signe est noté non pas comme le cercle plein θ popularisé par le *Cours de Linguistique Générale*, mais par des figures concaves et ouvertes : le signe est en lui-même vide (*kénôme* ∩), ou du moins ouvert sur le contexte (*sème associatif* ⊃ ⊂) ; au lieu de contenir un signifiant et un signifié, il ne prend valeur qu'en relation avec d'autres.

Une telle modélisation textométrique du texte comme trace en est au stade de l'intuition, et laisse encore songeur quant à sa faisabilité. En revanche, le modèle du texte comme trame est d'ores et déjà formalisé et éprouvé dans un logiciel (*Le Trameur*). Le texte comme trace n'a peut-être pas à devenir prochainement une modélisation textométrique pour de nouveaux logiciels ; plus fondamentalement, en textométrie, le caractère réducteur des représentations est assumé par le fait que les techniques d'analyse donnent accès à des régularités globales, des lignes de force, et rendent ainsi déjà puissamment comte de réalités textuelles de valeur. Souhaitons néanmoins que les observations et principes qui inspirent les présentes propositions puissent contribuer à orienter les développements actuels de la textométrie.

Cette communication a été préparée dans le cadre du projet Textométrie ANR-06-CORP-029.

Je remercie vivement les collègues du projet Textométrie et ceux du séminaire CoLiGram, et particulièrement François Daoust, Serge Heiden, Marie-Hélène Lay et André Salem, ainsi que les relecteurs, qui m'ont aidée avec bienveillance et expertise à consolider et à mûrir cette réflexion.

Références

Christ O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proc. of COMPLEX'94 (3rd Conf. on Computational Lexicography and Text Research)*, pp. 23-32.

Daoust F. (2007). *SATO 4.3, Manuel de référence*, mars 2007. En ligne : http://www.ling.uqam.ca/sato/index.html.

Daoust F. and Marcoux Y. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés. In Viprey J.-M., ed., *Actes des 8es JADT*. Besançon, Presses Universitaires de Franche-Comté, pages 327-340.

Fleury S. (2007). *Le Métier Lexicométrique aka Le Trameur. Manuel d'utilisation*. Septembre 2007. En ligne : http://tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf.

Heiden S. (2002). *Weblex. Manuel Utilisateur*. Version 4.1 (janvier 2002), Lyon : Laboratoire ICAR, UMR 5191, CNRS & Université de Lyon. En ligne : http://weblex.ens-lsh.fr/doc/weblex.pdf.

Heiden S. (2006). Un modèle de données pour la textométrie: contribution à une interopérabilité entre outils. In Viprey J.-M., ed., *Actes des 8es JADT*. Besançon, Presses Universitaires de Franche-Comté, pages 487-498.

Lamalle C., Fleury S. and Salem A. (2006). Vers une description formelle des traitements textométriques. In Viprey J.-M., ed., *Actes des 8es JADT*. Besançon, Presses Universitaires de Franche-Comté, pages 581-591.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Paris, Dunod.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In Purnelle G., Fairon C. and Dister A., eds, *Actes des 7es JADT*. Louvain-la-Neuve, Presses Universitaires de Louvain, pages 865-873.

Rastier F. (2001). *Arts et sciences du texte*. Paris : Presses universitaires de France.

Rastier F. (2005). Saussure au futur : écrits retrouvés et nouvelles réceptions. *Texto !*, mars 2005. En ligne : http://www.revue-texto.net/Saussure/Sur_Saussure/Rastier_Saussure.html.

Rastier F., Cavazza M., Abeillé A. (1994). *Sémantique pour l'analyse*. Paris, Masson.

Logiciels cités

Xaira : http://www.xaira.org/ Lexico 3 : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW
Weblex : http://weblex.ens-lsh.fr/wlx/ Hyperbase : http://ancilla.unice.fr/~brunet/pub/hyperbase.html
SATO : http://www.ling.uqam.ca/ato/sato/ Le Trameur : http://tal.univ-paris3.fr/trameur/