Un modèle de données pour la textométrie : contribution à une interopérabilité entre outils

Serge Heiden[1]

[1]ICAR UMR5191 – ENS-LSH – 69342 Lyon BP7000 Cedex 07 – France

Abstract

The research community for textual data analysis is organizing itself to better develop textometric tools, and to be able to share the textual data they analyze. The goal is to make fonctionalities and datas to better interoperate. This is important to be able to clarify complex software architectures involving natural language processing tools and to be able to capitalize the work of preparation of input data. To be able to globaly compare the various fonctionalities of the different textometric tools, we propose a synthetic fonctional model composed of 4 axes: Statistical analysis, Text edition, Search engine and Text annotation. There exist several international initiatives to standardise textual data description (metadata) and the encoding of their content. Being diverse in their application and ever evolving, we propose a synthetic data model for textometric tools composed of 11 different parts. It has been build by the analysis of the data formats used by the textometric tools. We propose to make the tools interoperate with data at that level of description.

Résumé

La communauté des chercheurs en analyse de données textuelles s'organise afin d'unir les efforts de développement et de diffusion des outils de textométrie ainsi que l'échange des données textuelles qu'ils traitent. L'effort nécessaire pour rendre les calculs et les données interopérables est important pour clarifier les architectures de traitement textométrique complexes intégrant les outils de TAL et pour la capitalisation du travail de préparation des données. Afin de pouvoir comparer globalement les fonctionnalités des outils, nous proposons 4 axes de synthèse fonctionnels : Synthèses statistiques, Édition de texte, Moteur de recherche et Annotation de texte. Il existe différentes initiatives internationales de standardisation de la description (métadonnées) et du codage du contenu des données textuelles. Vue la grande diversité dans l'usage de ces standards et leur évolution permanente, nous proposons de situer l'effort d'interopérabilité avec les données à un niveau synthétique composé de 11 rubriques générales. Elles ont été construites à partir d'une synthèse des différents formats de données des outils de textométrie. Nous proposons de situer le travail d'interopérabilité des données à ce niveau de description.

Mots-clés : textométrie, interopérabilité, modèle de données, format de représentation, feuille de style d'entrée, XML, TAL, Weblex.

1. Introduction

La communauté des chercheurs en analyse de données textuelles s'organise afin d'unir les efforts de développement et de diffusion des outils de textométrie. L'enjeu est de pouvoir appliquer ces outils à des corpus de données de plus en plus riches et variés, tout en continuant à développer la méthodologie par l'enrichissement qu'offre la confrontation des différents points de vue correspondant aux disciplines impliquées : statistique, informatique, linguistique... Le réseau international ATONET[1], financé par le gouvernement canadien, est un premier exemple d'organisation des activités en ce sens à des fins d'enseignement et de

[1] http://www.atonet.net.

recherche. Cet article est l'occasion pour nous de présenter le résultat de notre réflexion sur les moyens de rendre les outils du domaine interopérables entre eux.

1.1. L'interopérabilité se situe aussi bien au niveau des calculs que des données

La théorie informatique des systèmes d'exploitation et des langages de programmation réduit l'ensemble des activités informatiques à une articulation entre les calculs effectués et les données manipulées, même si les rôles joués par ces deux notions ne sont pas étanches, comme en atteste le modèle du lambda calcul où il existe des règles de passage entre les deux (Pierce, 2002). Quoi qu'il en soit, pour les deux paradigmes dominants que sont la sémantique[2] dénotationnelle (Schmidt, 1986) et la sémantique opérationnelle (Plotkin, 1981), on retrouve toujours une notion de mémoire ou de données et une notion de calcul ou de processus. On peut donc analyser l'interopérabilité entre outils textométriques aussi bien au niveau des fonctionnalités offertes par les outils qu'au niveau des ressources et des corpus de données disponibles pour les analyses. Par exemple, un outil peut faire appel à un module de calcul d'un autre outil pour mettre en oeuvre une fonctionnalité qu'il n'implémente pas lui-même. Cet appel peut se faire au moyen d'un protocole d'invocation externe comme les RPC[3] (Sun, 1988), ou bien prendre la forme d'appels successifs des différents outils, chacun recevant en entrée le résultat de l'outil précédent. Mais l'interopérabilité peut aussi se concevoir au niveau des échanges de ressources et de corpus de données. Les échanges de données peuvent se faire dans le cadre des protocoles standardisés du web comme SOAP[4] (W3C, 2003). Mais dans tous les cas, l'échange correct des données entre les outils reposera sur une interprétation commune de la sémantique des codes qu'elles représentent. À commencer par la valeur des caractères (leur ordre lexicographique, leur participation éventuelle à la forme des unités lexicales, etc.), mais aussi les codes typographiques (comme les sauts de ligne ou de page), éventuellement les balises - ou les principes - codant les informations linguistiques explicitées (délimitation des unités lexicales, codage de leurs propriétés, délimitation des phrases, etc.), la macro-structure et les propriétés des documents ou des corpus dans leur ensemble, etc. L'enjeu d'une telle interprétation commune est de pouvoir comparer des calculs similaires en appliquant deux outils différents sur des entités clairement identifiées d'un corpus. Par exemple, cela peut permettre de comparer précisément la pertinence de classements issus de calculs de cooccurrence d'unités lexicales graphiques au sein de phrases orthographiques[5], selon deux modèles statistiques différents. Bien sûr, ce type de comparaison repose sur l'hypothèse que les calculs considéreront les données de manière identique : unité lexicale, contexte de phrase, etc. En cherchant à satisfaire ces conditions, l'effort d'interopérabilité participe, à sa manière, à une généralisation de la méthodologie textométrique car il force une certaine abstraction des données qu'elle se donne à analyser.

[2] Dans la suite de cet article, la mention « sémantique » sans précision fera référence à la sémantique informatique définie ici, et non aux sémantiques linguistique ou logique.
[3] Remote Procedure Calling protocol (protocole d'appel de procédure à distance).
[4] SOAP : Simple Object Access Protocol (protocole simple d'échange d'objets) est une généralisation des protocoles RMI, CORBA, etc. Il permet d'invoquer des services web à distance, ce qui inclu l'échange de données. Ces services peuvent être décrits formellement au moyen de WSDL (Web Services Definition Language), ce qui permet de les annoncer, en tant que logiciel fournisseur de services ou de données, et de les découvrir automatiquement en tant que logiciel client.
[5] Spécification TEILITE section 16.1 « Phrases orthographiques »
<http://www.gutenberg.eu.org/publications/autres/TEILITE/node33.html>.

1.2. Pour une capitalisation des corpus de données

La compilation de corpus est une étape coûteuse dans l'analyse de données textuelles. Pour obtenir ces données, il faut :

- transcrire s'il s'agit de manuscrits médiévaux ou de vidéos par exemple ;
- formater pour que les outils trouvent les éléments nécessaires à l'analyse ;
- enrichir de sorte à obtenir des niveaux d'observables satisfaisant en fonction des objectifs de l'étude et des ressources disponibles ;
- calibrer l'ensemble du corpus en termes de représentativité, de couverture, et d'homogénéité.

Selon les projets, chacune de ces opérations a un coût variable mais non négligeable. Il est donc intéressant de trouver des moyens de rendre la forme des corpus générique pour les traitements que l'on s'apprête à leur appliquer. Ceci permet, par exemple, de bénéficier dans une nouvelle étude, d'une partie des données compilées pour une étude précédente, réalisée avec un outil différent. Cela est d'autant plus important que, dans une perspective historique, on constate souvent une réelle difficulté à pouvoir analyser des données formatées pour des outils de générations anciennes. À cet égard, se donner les moyens d'interopérabilité de corpus de données entre outils, c'est peut-être aussi, dans une certaine mesure, formater les données aujourd'hui pour les calculs de demain[6].

1.3. Il faut intégrer les outils de TAL à l'analyse d'interopérabilité

On a longtemps considéré l'application d'outils statistiques au niveau de données textuelles riches de diverses informations linguistiques - le lemme d'un mot ou sa catégorie morphosyntaxique, etc. - comme sensible du fait du coût et du manque de maîtrise de la qualité des projections de connaissances linguistiques au sein des textes (Geffroy et al., 1974). Dans la mesure où la textométrie s'applique désormais à des données de plus en plus riches (Pincemin, 2004) et maîtrisées (Adda et al, 1999), notre étude a considéré l'ensemble du continuum allant des outils de textométrie à proprement parler (Lexico, Hyperbase…) aux outils de Traitement Automatique de la Langue (TAL) de manière générale (TreeTagger, Intex…). Ces derniers sont, en effet, de plus en plus sollicités en amont des analyses. Ils conditionnent donc beaucoup les résultats et intègrent par ailleurs de plus en plus de fonctionnalités d'analyse ou d'exploration. Les inclure dans l'effort d'interopérabilité, c'est d'une part clarifier leur rôle dans l'ensemble de la méthodologie, et, d'autre part, les considérer comme des modules fonctionnels autonomes. Tout outil de textométrie peut alors les solliciter de la même manière qu'il le ferait avec les modules d'autres outils de textométrie. Dans ces conditions, on pourra par exemple appeler tel ou tel analyseur morphosyntaxique, en fonction de performances reconnues sur tel ou tel type de données, tout en étant assuré de la bonne interprétation des données d'entrée par l'outil de TAL. Cette démarche permettra d'évaluer la participation des outils de TAL utilisés en amont et celle des outils de textométrie utilisés en aval, et donc l'ensemble d'une analyse textométrique.

1.4. Comment rendre les outils interopérables ?

L'interopérabilité entre outils de textométrie et de TAL repose sur des spécifications - de formats de données et de fonctionnalités - explicites et partagées par tous les concepteurs et

[6] Adaptation libre de la devise de la TEI : *Yesterday's information tomorrow*.

utilisateurs d'outils. On trouve un exemple de cadre de mise en place de telles spécifications au sein du JCP[7] pour la standardisation des développements logiciels en langage de programmation Java. Au sein du JCP, le processus d'appel à spécification, de proposition et de validation par des implémentations de référence est explicite et formalisé (SUN, 2004). Même si ce processus industriel est relativement lourd, il peut être une référence intéressante pour organiser les efforts d'interopérabilité entre outils de textométrie.

2. Structurer les fonctionnalités considérées

Sur la base d'une acception large de la textométrie, les logiciels d'analyse de données textuelles pris en compte dans notre étude d'interopérabilité sont : Hyperbase, Lexico, Alceste, Sato et Weblex. L'analyse a été réalisée à partir de la documentation disponible et des corpus fournis avec chaque logiciel comme jeu d'essai, et non à partir de la mise en œuvre effective de chaque outil sur des corpus similaires.

Avant de pouvoir nous focaliser sur les diverses données pouvant participer à l'interopérabilité, nous commencerons par regrouper les différents traitements disponibles dans les outils selon quatre axes[8] fonctionnels. Ces axes, qui sont autant de points de vue complémentaires « outillés » sur les mêmes données textuelles, nous serviront à justifier le rôle de chaque partie du modèle de données proposé. Nous réduirons, dans la suite, l'usage des outils textométriques, et donc leurs fonctionnalités, à un va-et-vient entre ces quatre modalités d'accès aux données textuelles :

- un axe S comme « synthèses statistiques » permettant de calculer, à partir des données textuelles, divers types de synthèses classant ou regroupant les apparitions de phénomènes en s'appuyant sur différents modèles probabilistes. Cet axe est en général le plus développé dans les outils de textométrie ;
- un axe E comme « édition » permettant de restituer à la lecture le contexte « natif » des phénomènes textuels, soit selon une forme la plus proche possible du fac-similé du texte à l'origine de sa production, soit sous une forme élaborée où l'information éditée et l'aspect qu'elle prend dans l'édition sont choisis à la demande selon l'étude réalisée ;
- un axe M comme « moteur de recherche » permettant de trouver efficacement les occurrences d'un phénomène textuel, mot, locution, etc. et d'afficher ou de compter ses différentes réalisations dans le corpus de données analysé ;
- un axe A comme « annotation » permettant d'ajouter ou de modifier une information du corpus de données au moyen d'un codage spécifique, explicite ou non. Tous les outils ne prennent pas en compte ces préoccupations de mutabilité du corpus. Dans le cas où l'outil considéré ne le fait pas, on considère que le travail d'annotation a été réalisé en amont de l'outil de textométrie.

Ces axes nous serviront à rendre compte des grandes catégories de fonctionnalités de manière synthétique. Ils n'ont pas vocation à être étanches entre eux et indépendants. Par exemple, les

[7] Java Community Process (http://jcp.org). En particulier, le JCP propose des spécifications JDM (Java Data Mining) pour l'interopérabilité des plates-formes de « data mining » (Hornick, 2004) qui sont les plus proches des données et des fonctionnalités de la textométrie, et notamment la description des tâches de « text mining » JDM2 (Hornick, 2005) et d'analyse multidimensionnelle OLAP (Poole, 2003).

[8] Les trois premiers de ces « prismes » de lecture augmentée d'un corpus de données sont ceux proposés par l'outil Weblex.

concordances peuvent être interprétées comme une réédition du texte (axe E) résultant de l'affichage systématique des contextes d'apparition d'événements textuels trouvés à l'aide d'un moteur de recherche (axe M).

3. Données primaires (sources ou annotées) et secondaires

Du point de vue des données, nous proposons de distinguer les données primaires – sources ou annotées – des données secondaires.

Les données sources forment le support de l'expression des notions de sciences humaines et sociales manipulables par la textométrie. Dans le domaine de corpus de données issues de réalisations de la langue écrite, on trouvera par exemple les monographies numérisées en texte ou encore les éditions critiques de manuscrits, éventuellement liées aux images des folios de manuscrits. Dans le domaine de corpus de réalisations de la langue parlée, on trouvera les transcriptions de ce qui est dit et fait lors des interactions, éventuellement synchronisées avec le son ou la vidéo numérisés de l'interaction. Dans tous les cas, l'aspect numérisé de la donnée source implique des choix de représentation de l'information sous une forme codée pour la machine. Le code peut être directement orienté vers certains outils, et donc influencé par eux, ou plus largement propre aux systèmes d'exploitation des machines qui le stockent. En général, une étude reçoit des données sous la seconde forme et commence par les transformer pour les mettre sous la première forme de sorte à ce que l'outil de textométrie choisi puisse y interpréter correctement les informations dont il a besoin pour appliquer ses traitements.

Les données annotées relèvent d'un niveau d'interprétation plus riche et plus synthétique. Elles représentent le résultat d'un enrichissement de données sources, ou déjà annotées. Si le travail d'annotation se définit, dans toute sa généralité, par l'explicitation d'interprétations au moyen de codes, il peut s'avérer difficile de pouvoir dégager un état « originel » des données textuelles sources. En effet, si chaque caractère constitutif d'un texte ou d'une transcription porte toujours potentiellement une interprétation implicite - pour lui-même, pour le mot qu'il constitue, etc. - on peut considérer que les traitements textométriques portent toujours sur des données annotées. L'enrichissement peut avoir été réalisé de manière automatique, ou semi-automatique, par un outil de TAL, de sorte à trouver des informations nouvelles propres à l'outil : comme le lemme d'un mot, sa catégorie grammaticale, la limite de syntagmes ou de phrases, la délimitation et le typage d'entités comme des noms de personnes ou de lieux, etc. L'enrichissement peut aussi avoir été réalisé par des codeurs en vue d'ajouter toutes sortes d'informations nouvelles utiles à l'étude à réaliser ou à l'outil à appliquer, et ne ressortissant pas des possibilités des outils de TAL. C'est typiquement pour ce type d'informations, au delà des mots d'un texte, que le consortium de la « Text Encoding Initiative » (TEI, 2006) propose d'harmoniser le codage : structure logique d'un texte, attribution d'un émetteur à un discours rapporté, références au sein d'un texte, notes critiques, marques éditoriales, délimitation des éléments graphiques, délimitation du hors texte, des index, etc. Le travail consiste essentiellement à discuter collectivement de la structuration et de la terminologie des éléments codés. La qualité de ces éléments au sein d'un texte dépend des personnes ou des logiciels les ayant codés et des personnes ou des logiciels les interprétant. Leur première qualité est d'être homogènes au sein d'un même texte ou corpus. L'enjeu de cette standardisation des codes est celui de la capitalisation des interprétations par la transmissibilité des données entre collègues ou pour les générations futures. Mais à tout instant, chacun est libre de ne pas tenir compte de certains codes et de considérer la donnée textuelle sans informations supplémentaires particulières. Dans ce dernier cas, il faudra cependant s'assurer de la bonne interprétation du

texte correspondant[9] et éventuellement prendre certaines précautions liées à l'interopérabilité. Par exemple, dans le cas de certains outils de TAL appelés en amont, il faudra s'assurer qu'ils puissent restituer les informations correspondant au résultat de leur travail en accord avec les informations qu'ils ont reçues initialement même s'ils les ont ignorées en partie (Issac, 2005).

Les données secondaires correspondent aux informations finales, ou intermédiaires, de calculs de textométrie. Il peut s'agir, par exemple, de la liste hiérarchique du vocabulaire d'un corpus, de tableaux de contingence de différents phénomènes linguistiques ou de tableaux d'indices de spécificité d'apparition dans chacune des parties du corpus.

L'interopérabilité visée se situera au niveau des données annotées ou secondaires.

4. Format, modèle de données et métadonnées

Un format est une représentation des données au sein d'un fichier. Si l'interprétation des divers formats permet aux logiciels de textométrie de manipuler les données nécessaires sous-jacentes, c'est bien au niveau du modèle de ces dernières que nous devons poser les premières bases d'interopérabilité des données. Parmi celles-ci, on peut en distinguer certaines qui ne servent explicitement qu'à en qualifier d'autres : les métadonnées. Par exemple, dans le domaine du format XML et du modèle associé, une balise de début de paragraphe servira à délimiter et à qualifier l'ensemble des données suivantes, jusqu'à la balise de fin de paragraphe correspondante, d'une sémantique du type « ces informations se trouvent au sein d'un paragraphe ». Une fois le format linéaire XML interprété, on peut disposer dans les traitements d'un « élément paragraphe » composé de sous-éléments, contenus et qualifiés par lui, et situé lui-même au sein de la hiérarchie des éléments composant le document XML[10] représentant le texte. Au delà du rôle et de la place du paragraphe dans le texte, et donc des informations qu'il contient, ce dernier peut aussi être enrichi d'autres informations au moyen d'attributs, comme le numéro du paragraphe par exemple, qui influenceront également l'interprétation des données contenues. Bien sûr il existe un ensemble riche de structures englobant un paragraphe jusqu'à contenir l'unité textuelle entière ou le corpus de données lui-même. Et chaque niveau de structure peut se comporter comme une métadonnée pour le niveau inférieur. Pour un traitement textométrique, nous proposons d'appeler données les informations permettant de définir les éléments traités par les outils, et métadonnées ce qui permettra de les organiser en fonction des besoins des différents traitements. Dans le cas du modèle XML, par exemple, on trouvera les premières et les dernières parmi n'importe quel élément ou attribut XML en fonction de la structure logique choisie pour le document. À l'usage, on peut malgré tout distinguer un ensemble de métadonnées privilégié qui correspond à l'entête du texte, ou à celle du corpus dans son ensemble, regroupant toutes les informations qualifiant la totalité du contenu situé dans le corps du texte ou du corpus[11]. Nous proposons, dans ce cadre, de distinguer les métadonnées d'entête, globales, des métadonnées de corps, localisées au sein des observables. Les

[9] En effet, dans un format comme XML par exemple, il peut s'avérer délicat d'identifier la séparation entre les unités lexicales si ces dernières sont codées explicitement, au moyen d'une balise <W> par exemple, car dans ce cas les espaces de séparation initiaux peuvent avoir complètement disparus car redondants avec les balises. De même, les divers sauts de ligne forcés d'origine porteurs d'information peuvent avoir été remplacés par des interprétations explicites (marques de paragraphes <P>, de titres de section <HEAD>, etc.). Les techniques d'annotation débarquée offrent des solutions pour ne pas porter atteinte à l'intégrité des données sources. Elles étendent le modèle simple d'arborescence XML à une forêt d'arborescences pouvant se référencer mutuellement à l'aide de mécanismes de pointage, au caractère près (Habert, 2005 : 31).

[10] Appelé « infoset XML » (W3C, 2004).

[11] La TEI a consacré ces informations en développant un descripteur très riche appelé teiHeader.

métadonnées d'entête serviront à produire des comparaisons globales entre les différentes qualités des unités textuelles : leur genre, leur datation, etc. Les métadonnées de corps serviront à produire des comparaisons internes aux textes : évolution thématique entre les chapitres, spécificité du vocabulaire de chacun des acteurs d'une pièce de théâtre où le texte de chacun est codé au fil des prises de parole de la pièce.

5. Interopérabilité et standardisation des données

Nous avons déjà évoqué les efforts de standardisation du codage de la sémantique des données textuelles avec les recommandations de la TEI. Il existe des initiatives internationales orientées vers les descripteurs de données (les métadonnées) et vers le codage du contenu lui-même.

5.1. Métadonnées d'entête

Du point de vue le plus abstrait, le standard RDF[12] du W3C cherche à fédérer la manière de décrire l'ensemble des ressources disponibles sur l'Internet. Dans le domaine plus particulier des documents, le standard Dublin Core[13] propose un consensus minimal de descripteur pour renseigner l'auteur, la date, le titre, les conditions de diffusion, etc. Pour les ressources linguistiques à proprement parler, l'OLAC[14] étend le descripteur Dublin Core avec des champs de métadonnées supplémentaires. Enfin, l'IMDI[15] standardise la description des données multimédias et multimodales.

5.2. Codage du corps

Dans le domaine de la standardisation du codage du contenu des corpus de données, on peut citer les initiatives fédératrices suivantes[16] :

- la TEI édite ses recommandations pour le codage des textes littéraires et assimilés, les dictionnaires, etc. Il existe une extension, appelée XCES[17], pour coder les corpus linguistiques annotés et alignés ;
- l'OASIS[18] édite sa spécification OpenDocument[19] pour le codage des documents de bureautique (texte, tableau, dessin, diaporama) et DocBook[20] pour la documentation informatique (manuel de référence, guide utilisateur, etc.) ;

La communauté des concepteurs d'outils de textométrie et de TAL suit l'évolution de ces recommandations. L'intérêt fondamental de ces dernières est qu'elles sont explicites et consensuelles. On peut malgré tout constater une très grande diversité dans l'application de ces standards et une certaine variabilité dans leur évolution. C'est pourquoi nous proposons de porter l'effort d'interprétation des données pour l'interopérabilité des outils de textométrie à un niveau d'abstraction supérieur. Au lieu de se situer directement au niveau des données produites et manipulées par une communauté d'usagers donnée, nous proposons un modèle

[12] Resource Description Framework <http://www.w3.org/RDF>.
[13] De la Dublin Core Metadata Initiative <http://dublincore.org>.
[14] Open Language Archives Community <http://www.language-archives.org>.
[15] ISLE Meta Data Initiative <http://www.mpi.nl/IMDI>.
[16] Bien que ces standards recommandent le codage de certaines métadonnées d'entête, leur effort porte surtout sur le codage du corps ou du contenu des textes ou documents.
[17] <http://www.xml-ces.org>
[18] Organization for the Advancement of Structured Information Standards <http://www.oasis-open.org>.
[19] <http://www.oasis-open.org/committees/download.php/12572/OpenDocument-v1.0-os.pdf>
[20] <http://www.oasis-open.org/docbook/sgml/4.1/docbk41.zip>

générique des informations manipulées par tous les outils de textométrie. Ceci devrait permettre une première discussion à ce niveau d'échange, et une mutualisation des outils d'interfaçage de chaque standard vers ce modèle commun.

6. Structurer les données de la textométrie : un modèle de données type

Notre modèle de données type est constitué d'une liste de rubriques obligatoires servant à renseigner les différentes informations nécessaires aux traitements correspondant aux axes fonctionnels identifiés. Ils nous servent à ré-interpréter les formats actuels de représentation des données utilisés par les logiciels.

6.1. Codage utilisé

Si le format de représentation des données est ouvert[21], on peut spécifier la méthode qui sera utilisée pour y interpréter les codes. Par exemple, une spécification de codage XML permettra de donner un statut particulier aux caractères « < » et « > » pour pouvoir délimiter les éléments de codage, ou bien une spécification de « texte brut » forcera tous les caractères à avoir le même statut de codage. Le standard XML semble être un bon candidat pour spécifier le système de codage.

6.2. Encodage des caractères

Afin de pouvoir manipuler les caractères d'un texte, nous devons spécifier leur codage (Haralambous, 2004). Ce dernier sera à l'origine de la définition d'un ordre lexicographique et permettra de représenter les caractères dans les éditions au moyen de polices appropriées. Le standard Unicode semble être un bon candidat pour spécifier l'encodage des caractères.

6.3. Segmentation en unités

Cette rubrique concerne les moyens de segmentation en unités lexicales ou en unités plus complexes (locution, syntagmes, etc.). La segmentation est exprimable selon deux méthodes complémentaires :

- par intention (ou implicite) : par le paramétrage d'outils de calcul automatique de la segmentation (tokenizers intégrés à l'outil) - définition des caractères constituants de mot ou non, liste de cas particuliers pour les « mots » à morphologie spéciale, etc. On obtient des unités graphiques ;

- par extension (ou explicite) : par l'encodage explicite de la segmentation au sein des données elles-mêmes[22], et donc l'énumération exhaustive des unités, typiquement en TEI à l'aide de la balise XML `<w>`[23] enserrant chaque mot. On obtient des unités de statut quelconques.

[21] En opposition aux formats propriétaires dont la documentation n'est pas publiée.
[22] Pour toutes les propositions de codage au sein des données, nous considérerons que ce dernier peut s'exprimer aussi bien explicitement au fil du fichier de données, à l'aide de balises XML par exemple, qu'en étant déporté dans des fichiers annexes au sein desquels les codages désigneront précisément les données du fichier de base sur lesquelles ils portent au moyen de mécanismes de pointeurs entre fichiers de données, selon la technique d'annotation débarquée (Thompson H., 1997).
[23] pour *w*ord.

6.4. Propriétés des unités

Il est possible d'appliquer en amont de la lecture du corpus des outils d'enrichissement de la segmentation. Cet enrichissement peut correspondre au résultat d'appels d'outils de TAL comme les étiqueteurs en morpho-syntaxe (EF13[24] ou Cordial par exemple) ou les analyseurs syntaxiques, ou plus simplement à la projection hors-contexte de listes de propriétés a priori sur les unités (mots outils, synonymes, thésaurus, etc.).

Si les unités lexicales du corpus, ou de certaines œuvres du corpus, comportent des informations linguistiques (ou autres) supplémentaires, il faut expliciter leur existence et les moyens d'y accéder.

6.5. Délimitation des contextes

L'ensemble des contextes forme un espace de rencontre pour les modèles statistiques (collocations, cooccurrences, etc.). Cela peut aussi être un moyen de contraindre l'expression de ce qui est recherché dans les moteurs de recherche. Par exemple, pour la BFM, cela consiste en phrases et en paragraphes. On peut s'intéresser aux unités les plus cooccurrentes au sein de phrases. On peut par ailleurs limiter une recherche de collocations aux limites de paragraphes.

On peut distinguer trois moyens de segmentation en contextes :

- par intention : par le paramétrage d'outils de segmentation automatiques : caractères délimitant les phrases (ponctuation forte), les cas spéciaux (abréviations formées de points), etc. On obtient des phrases orthographiques ou des contextes spécifiques ;
- par fenêtrage : en faisant glisser une fenêtre de n unités, obtenus par intention ou par extension, dans un calcul de cooccurrences par exemple ;
- par extension : par l'encodage explicite, dans les données elles-mêmes, de la segmentation (typiquement `<S>` pour sentence [phrase], ou encore `<P>` pour paragraphe).

6.6. Analyse contrastive : partitions

La textométrie privilégie les approches de type contrastives. L'opposition sert à comparer différents ensembles de contenus selon des partitions obtenues à partir des métadonnées d'un corpus. Beaucoup d'outils de synthèse, comme les spécificités ou l'analyse factorielle (axe S), dépendent de la définition de partitions. Le contenu exprime les phénomènes ou les notions observés dans les données : mots, entités, etc. Les partitions, elles, sont construites - ou définies - à partir des métadonnées d'entête de chaque unité (œuvre ou transcription par exemple) et à partir des métadonnées situées dans le corps de ces unités[25].

6.7. Index

Il s'agit d'expliciter pour l'outil :

[24] (Prévost et al, 2003)

[25] Dans la Base de Français Médiéval (BFM, 2006) par exemple, les métadonnées trouvées dans les entêtes TEI des textes permettent de construire des contrastes entre : certains textes, certains auteurs, les genres de textes, les domaines, des combinaisons genre/domaine, la forme des œuvres : vers ou prose, des périodes chronologiques (ancien, moyen), des siècles ou des parties de siècles, etc. Pour les métadonnées situées dans le corps des textes, on peut par ailleurs contraster : le matériau textuel attribué à chaque locuteur dans certaines pièces de théâtre, les différentes sections d'un même texte, etc.

- les différents types d'index qu'il faut prendre en charge à la lecture du corpus : celui des mots du corps des textes, celui des mots d'autres langues, des mots des titres, des mots cités ou rapportés, des mots composant les notes, des mots distingués typographiquement par une marque (e.g. italique), etc. ;

- ce qui ne doit pas faire partie d'index particuliers (notes, commentaires éditoriaux, commentaires de codage, etc.) : le « hors texte ».

Les index serviront aux synthèses de vocabulaires et aux moteurs de recherche.

6.8. Alignement

Dans le cas de corpus multilingues alignés, il s'agit d'expliciter les portions de corpus à aligner et le niveau de structure sur lequel repose cet alignement (phrase, paragraphe, etc.).

6.9. Références

La spécification des références consiste en l'explicitation du contenu (ce qu'on présente) et du format (comment on le présente) des références bibliographiques situant les occurrences des phénomènes recherchés. Ces références peuvent être construites à partir de n'importe quelle métadonnée d'entête ou de corps du corpus.

Dans le cas des concordances KWIC, par exemple, il s'agira des informations synthétiques situées sur chaque ligne permettant de situer précisément l'occurrence du pivot[26]. Ces informations peuvent influencer les différents tris de concordances disponibles et donc leur interprétation. Par exemple, le double tri d'une concordance par siècle puis par le contexte droit du pivot offrira une synthèse diachronique des contextes, alors qu'un double tri par genre puis par contexte droit offrira une synthèse générique des mêmes informations.

De manière générale, le contenu et la forme des références sont fortement liés aux partitions concernées. En outre, on distinguera souvent les références courtes (utiles aux KWIC) des références longues, bibliographiquement plus complètes (voir ci-dessous).

6.10. Édition

Il s'agit de l'édition des textes du corpus pour l'affichage ou l'impression d'une mise en forme de leur contenu logique. La mise en forme peut être spécifiée au moyen de feuilles de style générant l'édition. La mise en forme peut aussi, plus simplement mais avec moins de généralité, consister en la spécification d'éléments permettant de coder directement l'édition : comme la pagination, pour une édition papier, ou la typographie à utiliser, pour les mises en évidence au sein des contextes de concordances par exemple. Cette rubrique concerne également l'explicitation des différents niveaux de la macro-structure logique de l'œuvre pour établir le sommaire de son édition. Cet aspect est souvent lié au choix des références.

Si les outils de création de l'édition s'appuient sur des standards du W3C, il est possible d'envisager une interopérabilité à ce niveau. L'enjeu serait de pouvoir capitaliser les efforts de formatage des données textuelles indépendamment des outils qui les analysent. Dans le domaine XML, cela pourrait prendre la forme suivante. Les premières transformations de la

[26] Par exemple pour la BFM cela concerne pour les œuvres en vers : le nom du texte, le n° de page, le n° de vers et pour les œuvres en prose : le nom du texte, le n° de page, le n° d'ordre du paragraphe dans la page ou son numéro propre selon l'édition numérique TEI.

structure logique vers des structures de formatage pourraient reposer sur une articulation de feuilles de style de transformation XSLT (W3C, 2005a). Ensuite, le formatage définitif pour chaque type de support pourrait être assuré par des feuilles de style XSL-FO (W3C, 2001) ou CSS2 (W3C, 1998). Enfin, les données multimédia (images, vidéos) pourraient par ailleurs être synchronisées avec leur transcription au moyen du format SMIL (W3C, 2005b).

6.11. Descripteur bibliographique

Ceci concerne la description précise du corpus et de l'initiative à l'origine de sa création, des références bibliographiques complètes des œuvres qu'il contient, de publications éventuelles associées et des responsabilités (d'annotation, d'enrichissement, etc.). Il peut par ailleurs former un sommaire des points d'accès au corpus : partitions, index disponibles, enrichissements linguistiques éventuels, accès au sommaire des œuvres, etc. Enfin, le descripteur est le lieu privilégié pour rendre compte des métadonnées d'entête.

7. Conclusion : des feuilles de style d'entrée plutôt qu'un format pivot ou des passerelles ad hoc

Après avoir décrit les enjeux liés à l'effort d'interopérabilité des données et des fonctionnalités des outils de textométrie et de TAL, nous avons proposé un modèle fonctionnel de synthèse composé de 4 axes et un modèle de données de synthèse composé de 11 rubriques minimales à renseigner lors de l'intégration d'un corpus de données par un outil en vue de son traitement. Nous espérons que ces axes et rubriques alimenteront la discussion sur les points de convergence et de partage entre les outils. Dans le cas précis de la redéfinition de l'architecture de l'outil Weblex, nous avons déjà réinterprété les fonctionnalités d'après ces axes et ces rubriques. Le travail porte maintenant sur les moyens d'interprétation de ces rubriques dans les données standardisées. Nous étudions la possibilité d'apparenter cette interprétation à la mise en œuvre de feuilles de style de lecture : comme pour l'édition, quand il s'agit de rester dans l'infoset XML par exemple, il s'agit d'associer à des règles de lecture des instructions de création de la structure de données correspondant au modèle de données de l'application. Les feuilles de style d'entrée seraient alors le pendant symétrique des feuilles de styles de formatage de sortie XSL-FO (W3C, 2001), avec une structure logique d'intégration en lieu et place d'une structure logique de formatage.

Références

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français. *Actes de la conférence sur le Traitement Automatique du Language Naturel (TALN-99)*, Cargèse, Corse, France, 12-17 Juillet 1999 : 15-24.

BFM (2006). *Base de Français Médiéval* [En ligne]. Lyon, UMR ICAR / ENS-LSH, http://bfm.ens-lsh.fr.

Geffroy A., Lafon P. et Tournier M. (1974). L'indexation minimale. Plaidoyer pour une non-lemmatisation. ENS de Saint-Cloud. Communication au *Colloque sur l'Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale*, Strasbourg, 21-23 mai 1973.

Habert B. (2005). *Instruments et ressources électroniques pour le français*. Ophrys, Paris.

Haralambous Y. (2004). *Fontes et codage*. O'Reilly.

Hornick M. (2004). *Java Data Mining (JDM)*. JSR-73, Oracle Corporation, <http://jcp.org>

Hornick M. (2005). *Java Data Mining (JDM) 2.0*. JSR-247, Oracle Corporation, <http://jcp.org>

Issac I., Loiseau S., Pincemin B. et Chanove M. (2005). Repères et propositions pour l'intégration d'XML dans les analyseurs linguistiques de corpus. *Communication à la Journée ATALA, Articuler les traitements sur corpus*, Paris, 12 février 2005, http://www.limsi.fr/Individu/habert/Projets/JourneeATALA05ArticulerLesTraitements/ChanoveEtAl.pdf.

OASIS (2006). http://www.oasis-open.org.

Pierce B. C. (2002). *Types and Programming Languages*. MIT Press, Cambridge.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In *Le Poids des mots, JADT 04*, Louvain, UCL-Presses universitaire de Louvain, vol. 2 : 865-874.

Plotkin G. (1981). *A structural approach to operational semantics*. Tech. Rep. DAIMI FN-19, Computer Science Dept., Aarhus University, Aarhus, Denmark. <http://citeseer.ist.psu.edu/plotkin81structural.html>.

Poole J. (2003). *Java OLAP Interface (JOLAP)*. JSR-69, Hyperion Solutions, http://jcp.org.

Prévost S. et Heiden S. (2002). Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités. In *Romance Corpus Linguistics : Corpora and Spoken Language*, Gunter Narr Verlag, Tübingen.

Schmidt D. (1986). *Denotational Semantics, a Methodology for Language Development*. Allyn and Bacon. B.

Sun Microsystems, Inc. (1988). *Request For Comments : 1057*, RPC, Remote Procedure Call, Protocol Specification, Version 2, http://www.ietf.org/rfc/rfc1057.txt.

Sun Microsystems, Inc. (2004). *JCP 2 : Process Document*, http://jcp.org/en/procedures/jcp2.

Thompson H. et McKelvie D. (1997). Hyperlink semantics for standoff markup of read-only documents. *Actes de la conférence SGML Europe '97*, Barcelone, May 1997.

TEI (2006). *Text Encoding Initiative*, http://www.tei-c.org.

W3C (1998). *Cascading Style Sheets*. Level 2, CSS2 Specification, http://www.w3.org/TR/REC-CSS2.

W3C (2001). *Extensible Stylesheet Language (XSL)*. Version 1.0, section 6 “Formatting Objects”, http://www.w3.org/TR/xsl/slice6.html#fo-section.

W3C (2003). *SOAP Version 1.2 Part 1 : Messaging Framework*. http://www.w3.org/TR/soap12-part1.

W3C (2004). *XML Information Set* (second edition), http://www.w3.org/TR/xml-infoset.

W3C (2005a). *XSL Transformations (XSLT) Version 2.0*, http://www.w3.org/TR/xslt20.

W3C (2005b). *Synchronized Multimedia Integration Language (SMIL 2.1)*. http://www.w3.org/TR/2005/REC-SMIL2-20051213.