Vers une description formelle des traitements textométriques

Cédric Lamalle, Serge Fleury, André Salem

EA2290 SYLED - Centre de textométrie, Université de la Sorbonne nouvelle – Paris 3

Ilpga, 19 rue des Bernardins, 75005 Paris

> /…/ Le second, de diviser chacune des difficultés que j'examinerais en autant de parcelles qu'il se pourrait et qu'il serait requis pour les mieux résoudre.
>
> René Descartes, *Discours de la Méthode*, 1637.

Abstract

Various software dedicated to automatic text analysis apply different combinations of statistical methods to analyse text corpora stored with specific format. Results provided by such packages are generally very uneasy to be compared. We try here to provide a generic description of data structures used by such softwares (segmentation, partition, lexical tables) in order to allow accurate descriptions of these structures and easier exchanges between packages.

Résumé

Différents logiciels de textométrie applicables aux corpus de textes informatisés mettent en œuvre des combinaisons variables de méthodes statistiques à partir de formats d'entrée qui leur sont propres. Ces logiciels produisent des résultats dans des formats qui constituent souvent un obstacle à leur comparaison. On tente ici d'esquisser une description générique des objets intermédiaires manipulés par ces logiciels (segmentations, partitions, tableaux de décomptes) qui permettrait à la fois : de mieux décrire ces objets, de permettre leur transmission d'un logiciel à l'autre, de comparer les séquences de traitement et de permettre une meilleure confrontation des résultats finaux.

1. Introduction

Dans les dernières décennies, le recours à l'informatique a permis d'automatiser de nombreuses procédures utilisées depuis longtemps par les spécialistes des études textuelles. L'activité qui consiste à explorer des corpus de textes en s'appuyant sur des méthodes formelles s'est rapidement étendue à l'ensemble de plus en plus volumineux des données textuelles disponibles sur support numérique (Habert et al., 1997). Parallèlement, des procédures de calcul dont la mise en œuvre n'aurait pu être envisagée sans le secours de l'outil informatique se sont banalisées pour devenir de simples fonctionnalités de la plupart des logiciels de textométrie (analyses factorielles, classifications automatiques, analyses arborées, etc.) (Habert et al., 1998) [1].

L'introduction de ces différentes méthodes à la fois *reproductibles* et *objectivantes* a donné un souffle nouveau aux études textuelles. Cependant, dans notre discipline plus qu'ailleurs,

[1] Cet article se veut une contribution à la discussion engagée sur les traitements textométriques dans le cadre du réseau ATONET de l'UQAM (http://www.ling.uqam.ca/forum/reseau.html). Les propositions avancées dans cette étude n'engagent que les auteurs qui tiennent cependant à remercier l'ensemble des membres du réseau et tout particulièrement F. Daoust, S. Heiden, M. Jacobson, A. Lelu pour les fructueuses discussions qu'ils ont eu avec eux à ce sujet et dont ils espèrent avoir tiré profit.

l'introduction du principe de division temporaire entre la forme et le sens afin de permettre des traitements à grande échelle portant sur la seule forme du matériau textuel continue d'interpeller les spécialistes de la langue et du discours. Ces derniers souhaitent légitimement introduire dans les analyses automatisées les catégories de description qu'ils ont élaborées à propos des textes et des unités qui les composent. Pour les spécialistes des sciences exactes habitués à manipuler des traits et des propriétés dont la description est explicite et ne souffre que peu d'exceptions, les catégories manipulées par les linguistes (lemmes, catégories grammaticales ou sémantiques) peuvent paraître un peu floues au premier abord. Cependant, le chercheur qui travaille au contact des textes est inévitablement amené à en reconnaître la pertinence et l'utilité pratique.

Cette dernière exigence a été à l'origine de l'élaboration dans les dernières décennies de toute une série d'objets informatisés, de plus en plus disponibles pour les chercheurs de toutes disciplines, que nous appellerons des *ressources textométriques informatisées* (procédures de traitement ou bases de données textuelles ou grammaticales). Ces objets peuvent être des dictionnaires (unilingues : une entrée lexicale pourvue d'une définition ou de l'instanciation d'un système de traits, plurilingues : donnant la traduction de chaque unité dans une série de langues, etc.). Ils se présentent en général sous la forme de procédures faciles à mettre en œuvre mais qui réalisent cependant des traitements relativement complexes. Ces objets informatisés ont en commun d'avoir été élaborés à partir d'un important travail humain[2] qui n'est pas toujours intégralement formalisable et dont la description exhaustive nécessite parfois l'énumération complète des actions qu'ils effectuent.

Convoquée afin d'apporter dans le domaine des études textuelles un corps de méthodes fiables, reproductibles et susceptibles d'une description exhaustive, la démarche textométrique apparaît parfois, en dépit des progrès qu'elle entraîne, comme un maquis de procédures s'appuyant sur des savoirs qui restent hétérogènes.

Une mise à plat des opérations réalisées par les logiciels textométriques permettrait de mettre en évidence les similarités qui existent entre les différentes familles de traitements ainsi que les aspects sur lesquels elles divergent. Elle ouvrirait la voie à des échanges de données, de procédures de traitement et de produits d'analyse entre différents logiciels ou du moins à une comparaison plus aisée entre les résultats qu'ils produisent. Nous commencerons par faire le point sur la nature des traitements effectuées par les différents logiciels (§1) pour étendre ensuite cette classification aux principaux objets qu'ils manipulent (§2). La dernière partie (§3) sera consacrée à des propositions de modularisation des traitements permettant l'échange de résultats entre logiciels.

2. Traitements textométriques

Les différents logiciels de textométrie rassemblent en des combinaisons variables des méthodes directement élaborées dans le domaine de l'analyse quantitative des textes ou importées de domaines de recherche différents. Pour l'utilisateur, ces logiciels se présentent souvent comme des associations inextricables de procédures informatisées qui lui interdisent en fait de comprendre la nature des traitements utilisés, de percevoir les limites qui leur sont inhérentes ainsi que les artefacts qui peuvent résulter de leur utilisation simultanée.

[2] On peut par exemple souligner que l'annotation *automatique* par des catégories de description n'est pas exempte d'erreurs (Véronis, 2000).

La présentation sous forme *logicielle* des différentes séquences de traitements tend à masquer la distinction entre les traitements qui réalisent des opérations entièrement formalisées et ceux qui font intervenir des savoirs et des déterminations humains plus ou moins explicités. Elle complique l'identification à l'intérieur des différentes chaînes logicielles des étapes qui réalisent des traitements analogues, voire identiques.

La figure 1 propose un schéma général de l'analyse textométrique des corpus de textes. L'approche textométrique peut être sommairement divisée en trois phases : préparation des corpus, analyse statistique textuelle proprement dite et phase de commentaire des résultats obtenus. On a représenté en noir sur ce schéma les phases au cours desquelles interviennent des opérations entièrement formalisables (décomptes d'unités définies dans une phase précédente, analyses statistiques à partir de ces décomptes). Nous appellerons ces phases : des *phases de type 1*. Nous appellerons *phases de type 2* les séquences de traitement construites à partir d'opérations nécessitant une intervention humaine moins formalisée (sélection du corpus, détermination de la nature des unités de décompte et des partitions du corpus). Ces étapes de l'analyse textométrique ont été représentées en gris sur le schéma.

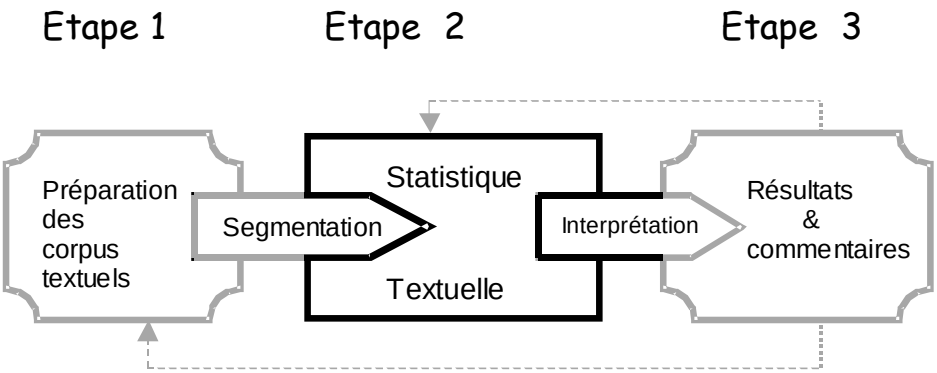

Figure 1 : *Schéma de l'analyse textométrique*

Les opérations qui permettent la transition entre les différentes phases présentent inévitablement un statut intermédiaire du point de vue de leur degré de formalisation. L'opération de *segmentation* commence par mettre en œuvre des règles et des savoirs parfois difficiles à expliciter de manière communicable. Elle aboutit cependant à un découpage suffisamment opérationnel pour être confié aux procédures de statistique textuelle.

À l'inverse, la phase d'*interprétation* s'appuie sur des résultats produits par des machines à l'aide d'algorithmes hautement formalisés. Cependant, les choix opérés parmi ces résultats par le chercheur et l'ordonnancement des résultats sélectionnés afin de produire un commentaire intelligible par un être humain et pertinent pour les spécialistes du domaine font de cette étape une opération moins formalisée que l'étape statistique précédente. Des flèches grises en pointillé indiquent la possibilité, largement utilisée dans la pratique par les chercheurs, de retourner vers le choix et le paramétrage des méthodes ou vers la redéfinition du corpus de départ au vu des résultats produits par les analyses statistiques.

2.1. Une tour de Babel logicielle

Pour évaluer la pertinence et l'apport réel de chaque méthode il serait indispensable de réaliser des comparaisons systématiques, en faisant varier : les corpus, les modes de

segmentation de la chaîne textuelle, les procédures d'identification des types, les méthodes de calcul et les visualisations réalisées à partir de ces méthodes.

Cependant, pour le chercheur désireux de comparer les démarches proposées par les différents logiciels, la difficulté principale réside dans le fait que ces derniers n'autorisent que très exceptionnellement l'accès aux objets intermédiaires (listes, tableaux, partitions) qu'ils construisent au cours des différentes étapes du traitement. Les chaînes de traitement se présentent la plupart du temps comme des *tunnels* méthodologiques qui imposent que les corpus de textes soient mis dans une forme particulière pour pouvoir y pénétrer et qui restituent des résultats dans des formes qui leur sont propres interdisant pratiquement toute comparaison et toute analyse sur la pertinence et l'apport de chacune des étapes du traitement. De ce fait, les tentatives de comparaison que l'on trouve dans la littérature se limitent le plus souvent à des appréciations *globales* sur l'apport de chacun des logiciels utilisés, pris comme un tout indissociable.

Dans ce qui suit, nous proposons de modulariser les chaînes logicielles et d'ouvrir des possibilités de comparaisons plus fines, portant à la fois sur les différentes méthodes de segmentation des corpus de textes et sur les éléments de procédures qui entrent dans la composition des chaînes de traitement.

2.2. Analogies entre traitements

Au-delà de ces particularités, les logiciels de la famille *textométrique* présentent de nombreux traits communs qui concernent à la fois les objectifs qu'ils affichent, les méthodes qu'ils convoquent pour atteindre ces objectifs et les résultats qu'ils produisent. Parmi les objectifs communs, on remarque tout d'abord la volonté de réaliser des rapprochements, des typologies, à partir des textes réunis en corpus. Dans certains cas, la visée typologique concerne avant tout les unités textuelles dont on peut recenser les occurrences dans la chaîne textuelle (formes graphiques, familles morphologiques, lemmes, segments répétés, etc.).

Pour avancer vers ces objectifs, les logiciels textométriques réalisent simultanément deux opérations distinctes. La première que nous appelons *microsegmentation* consiste à segmenter la chaîne textuelle en *occurrences* (angl. *token*), puis à regrouper ces dernières, à partir de critères d'identification variables, en des ensembles d'occurrences considérées comme identiques que l'on appellera des *types* (angl. *types*). La seconde opération que nous appelons *macrosegmentation* permet de repérer les frontières des parties ou fragments du corpus que l'on désire comparer. À partir de cette division du corpus en parties il est ensuite possible de calculer la ventilation des différents types dans les parties du corpus.

Parmi les résultats produits par les logiciels de la famille textométrique, on retrouve la plupart du temps :

- des listes d'unités textuelles (unités caractéristiques pour un ensemble de textes) ;
- des tableaux présentant les décomptes de ces unités dans les parties d'un corpus ;
- des matrices de distances entre parties du corpus pouvant ensuite servir de base à différentes visualisations (typologies, classifications, représentations arborées, etc.) portant sur les parties du corpus ;
- des délimitations de zones remarquables dans le texte de départ (sélections de zones, partitions, partitions hiérarchisées).

2.3. Classer les traitements

Les descriptions des traitements textométriques effectués à partir de grands corpus de textes que l'on rencontre dans la littérature révèlent un certain embarras terminologique lorsqu'il s'agit de qualifier la nature des traitements effectués sur les données textuelles. Face à des corpus comportant plusieurs millions d'occurrences, les qualificatifs *informatisé* et *automatisé*, largement employés dans la période précédente, sont devenus peu informatifs dans la mesure où personne n'imagine plus que des traitements aussi volumineux puissent être effectués sans le secours de l'ordinateur. Pour décrire les opérations qu'il n'est pas possible de confier à une machine et qui nécessitent une intervention humaine (ex : désambiguïsations portant sur le lemme, la catégorie grammaticale, etc.) se sont répandus des termes comme : *manuel, assisté par ordinateur* ou *requérant une intervention humaine* sans que ne s'installe l'habitude de préciser à chaque fois et de manière explicite la nature des opérations effectuées. Cette imprécision complique l'évaluation des résultats produits et leur comparaison avec des résultats issus d'autres méthodes.

L'évolution actuelle des logiciels de textométrie complique encore la catégorisation explicite des traitements. À côté des traitements affichant de manière exhaustive les opérations qu'ils font subir aux textes sont apparus des traitements appuyés sur des ressources extérieures (dictionnaires de formes, de traits sémantiques, etc..). Ces traitements se présentent sous une forme double : procédures informatisées d'une part, ils mettent en œuvre des déterminations sélectionnées et agencées par des utilisateurs humains, sans que les actions qu'ils opèrent ne donnent toujours lieu à une description exhaustive pour l'utilisateur.

Traitements *intrinsèques* et traitements *assistés*

Pour décrire un ensemble *A*, on utilise couramment la distinction entre *description en intension* (ou *en compréhension*) et *description en extension*[3]. Cette distinction fournit une analogie utile pour classer les traitements textométriques. Nous distinguerons les traitements pouvant être décrits par l'énoncé des *tâches* qu'ils sont censés réaliser par opposition aux traitements dont la description, pour être exhaustive, nécessite une énumération des traitements élémentaires réalisés dans un grand nombre de cas particuliers.

Deux tâches textométriques nous permettrons d'illustrer cette distinction[4]. La *segmentation automatique* d'un texte en formes graphiques s'opère à partir d'une courte liste de délimiteurs de formes qui serviront de point d'appui au découpage automatique de la séquence textuelle. Cette opération fournit l'exemple d'une tâche du premier type. Dans ce cas, la méthode accompagnée d'une courte liste de paramètres permet à elle seule de prédire les résultats du processus à partir d'un fichier texte donné. Nous appellerons les méthodes de ce type des *méthodes intrinsèques* pour souligner qu'elles contiennent en elles-mêmes l'ensemble des données nécessaires à leur mise en œuvre.

À l'opposé, le rattachement de chacune des occurrences d'un fichier texte à un *lemme* s'appuie en général sur la mobilisation de ressources textuelles volumineuses constituées à la fois de dictionnaires (de formes, de locutions) et de procédures adaptées à la levée des

[3] Pour un ensemble ***A***, la description en extension réalise l'énumération de tous les éléments appartenant à l'ensemble (ex : ***A*** = {2, 4, 6}). Une description en intension (ex : ***A*** = {ensemble des x tels que x est un entier pair, strictement positif inférieur à 7} définit le même ensemble en utilisant des propriétés communes à tous les éléments de l'ensemble et à eux seuls.

[4] Notons qu'il ne s'agit nullement de porter ici une appréciation sur l'utilité ou la pertinence de chacune des tâches présentées mais plutôt d'insister sur la différence de nature que ces tâches présentent au plan méthodologique.

nombreux cas d'ambiguïtés. Ces ressources qui résultent de savoirs mis en forme par des spécialistes humains au fait des particularités de la langue concernée ne peuvent en général être décrites autrement que par une énumération de traitements adaptés à des cas particuliers. Nous appellerons ces dernières méthodes : des *méthodes assistées*[5]. Notons que, dans la pratique, les traitements textométriques combinent souvent les tâches du premier type et celles du second[6].

3. Corpus textuels

Les corpus de textes que l'on réunit à des fins textométriques se présentent sous des formats extrêmement variables qui résultent des contraintes et des traditions diverses ayant présidé à leur constitution, leur encodage, leur archivage et leur présentation. Cependant, ces corpus présentent, la plupart du temps, des caractéristiques générales communes. On distingue souvent des *zones textuelles* proprement dites qui contiennent le texte lui-même et des *zones péri-textuelles* qui contiennent des informations sur le corpus (date de création, auteur, etc.) et/ou des informations sur des éléments contenus dans la zone textuelle (appartenance à une partie, catégorie grammaticale, lemme de référence, etc.). Des encodages particuliers permettent en général de délimiter ces différentes zones dans le texte de départ à l'aide de *jalons textuels* (ou balises).

3.1. Formats propres et formats d'échange

La mise en œuvre de chaque logiciel textométrique suppose, en règle générale, que les données lui soient présentées sous un *format propre*, incontournable pour la prise en charge du texte par le logiciel. Les résultats restitués par les différents traitements (listes, tableaux statistiques, etc.) se présentent également sous des *formats propres* interdisant pratiquement leur utilisation immédiate par d'autres procédures. Il n'est pas abusif de dire que chaque logiciel consomme, manipule et produit des *données captives* difficilement utilisables par d'autres logiciels.

Cette situation n'est pas propre au domaine textométrique. Les traitements informatiques modernes réalisés à partir des données produites dans la plupart des activités de recherche s'orientent de plus en plus vers la création de formats d'échanges utilisant un langage de balisage commun, le langage XML[7] (Harold, 2001). Pour ce qui la concerne, la communauté des études textuelles travaille depuis plusieurs années à mettre en œuvre des normes d'échange et d'archivage des textes. La tentative la plus aboutie est sans doute la *Text Encoding Initiative*[8] (TEI) (Ide et al,. 1996) qui propose, en conformité avec les normes XML, des normes de présentation et d'échange des textes particulièrement adaptées à leur traitement sur ordinateur.

[5] Cette distinction vise avant tout à permettre une meilleure description des procédures textométriques du point de vue de leur conception. Dans la pratique, cette tâche peut s'avérer délicate, les éditeurs de textes s'adjoignant de plus en plus des ressources orthographiques voire syntaxiques, les langages de programmation intégrant des fonctions capables de réaliser des tris lexicographiques, à partir de données unicodes, etc.

[6] Notons, en outre, que la description minutieuse d'un traitement sous forme algorithmique ne permet pas de percevoir, sans une expérience pratique réitérée du traitement sur des données variées, les possibilités et les limites des analyses qu'il permet de faire.

[7] Extensible Markup Language (XML) est un ensemble de règles, de lignes directrices, de conventions pour la conception de formats texte permettant de structurer des données (http://www.w3.org/XML/).

[8] La *Text Encoding Initiative* est un projet international visant à mettre au point des directives pour l'élaboration et l'échange de documents électroniques à des fins de recherche (http://www.tei-c.org/).

Dans le domaine textométrique le réseau ATONET[9] tente d'élaborer à la fois : des passerelles permettant de convertir les textes d'un format propre à chaque logiciel textométrique vers d'autres formats propres et des normes d'entrée compatibles avec les normes TEI et XML.

3.2. Avatars d'une base textométrique

Derrière ces formats propres spécifiques à chaque logiciel, on constate que les entités que l'on est amené à traiter sont souvent analogues (unités textuelles, parties de corpus, zones particulières, etc.).

Le tableau 1 présente, à titre d'exemple, trois formes différentes d'une même séquence textuelle, préalablement soumise à un étiquetage grammatical et à une lemmatisation. On voit sur cet exemple qu'il serait possible de transformer chacun des états dans l'autre, à l'aide d'une procédure très simple. Nous appellerons *avatars* (d'une même base ou d'un produit textuel) des états d'une base textuelle entièrement reconstituables l'un à partir de l'autre à l'aide de procédures intrinsèques.

- B_1 est un *avatar* de B_2 si chacune des bases textuelles peut être entièrement reconstituée à partir de l'autre[10].
- B_1 *contient* B_2 si B_2 peut être entièrement reconstitué à partir de B_1.

a) données grammaticales *embarquées* dans la séquence textuelle

Les [DETDEF LE] représentants [NOM REPRESENTANT] du [PREP-DET DE-LE] peuple [NOM PEUPLE] français [ADJ FRANCAIS]

b) données textuelles utilisant un balisage de type XML

```
<mot><forme>Les</forme><cat>Detdef</cat><lemme>Le</lemme></mot>
<mot><forme>représentants</forme><cat>NOM</cat><lemme>REPRESENTANT</lemme></mot>
<mot><forme>du</forme><cat>PREP-DET</cat><lemme>DE-LE</lemme></mot>
<mot><forme>peuple</forme><cat>NOM</cat><lemme>PEUPLE</lemme></mot>
<mot><forme>français</forme><cat>ADJ</cat><lemme>FRANCAIS</lemme></mot>
```

c) étiquettes grammaticales présentées en colonnes d'un tableau

Forme	*Catégorie*	*Lemme*
Les	DetDef	LE
représentants	Nom	REPRESENTANT
du	PREP-DET	DE-LE
peuple	NOM	PEUPLE
français	ADJ	FRANCAIS

Tableau 1 : *Trois avatars d'une segmentation à partir d'une même séquence textuelle*

Plusieurs classes d'objets manipulés par les logiciels textométriques (textes, listes, tableaux) se présentent également sous formes d'avatars qu'il est aisé de convertir à l'aide de méthodes intrinsèques.

[9] Le réseau ATONET tente de développer les conditions favorisant une mise en commun de ressources et de méthodes à des fins d'enseignement et de recherche dans le domaine de l'analyse de corpus textuels (http://www.ling.uqam.ca/forum/atonet/).

[10] Du sanscrit *avâtara* « descente ». Dans la religion hindoue, chacune des différentes incarnations de la divinité Vichnou.

4. Modulariser les traitements

La nécessité de comparer les résultats produits par les différentes étapes des traitements textométriques conduit à proposer des stratégies de traitement plus modulaires. Dans la partie supérieure du tableau 2 on a représenté la situation actuelle de la pratique des traitements textométriques à partir d'un même corpus de texte. Le corpus est d'abord soumis à un pré-traitement pour aboutir à un *format propre* (à chaque logiciel). Cette prise en charge par le logiciel permet d'appliquer ensuite une série de traitements pour aboutir à la production de *résultats* qui se présentent eux-mêmes dans des formats propres à chaque logiciel.

La partie inférieure de ce même tableau représente la situation vers laquelle nous proposons d'évoluer. Dans ce schéma, on conserve la possibilité d'entrer dans chaque logiciel à partir d'un format propre à chaque logiciel ; mais la partie centrale du tableau montre cette fois une zone normalisée contenant textes et produits de traitements. Cette zone communique par des passerelles avec les différents formats propriétaires[11]. Les différents modules qui constituent chacun des traitements textométriques permettent cette fois de recevoir et d'exporter des résultats structurés qui peuvent être réutilisés par d'autres chaînes de traitement.

4.1. Exemples de résultats échangeables

On trouvera ci-dessous, à titre d'exemple, quelques objets textométriques génériques pour lesquels il est envisageable d'organiser des procédures d'échange entre modules de traitement appartenant à des logiciels différents.

4.1.1. Texte segmenté

Comme on l'a vu plus haut, les différentes procédures de segmentation, intrinsèques ou assistées, produisent différents découpages de la chaîne textuelle en *occurrences*. Certaines des procédures de segmentation découpent la chaîne textuelle en unités graphiques, d'autres effectuent un rattachement des occurrences à des *types* par des procédures d'identification spécifiques. Quelle que soit la procédure de segmentation utilisée, le résultat de cette opération se présente comme une suite d'occurrences que l'on peut donc transmettre en tant que telle à des procédures externes. Le numéro d'occurrence peut alors constituer une *cadence textométrique* permettant de créer un système de coordonnées sur le texte, de définir différentes zones et partitions du texte et de rattacher des annotations à chacune des occurrences[12].

4.1.2. Liste d'unités textuelles

À partir d'un texte segmenté en occurrences, on peut constituer des listes d'unités textuelles (formes graphiques, lemmes etc.) en fonction du type de segmentation initiale ainsi que des listes de séquences ou de cooccurrences relatives à ces mêmes unités.

4.1.3. Tableau à double entrée

À partir d'un texte segmenté en occurrences et d'une partition du corpus, repérable à partir de jalons permettant de découper le texte en parties ou de tout autre avatar de l'objet partition, on

[11] Les participants au réseau ATONET proposent des solutions pour cette partie normalisée. F. Daoust a réalisé des passerelles entre les formats de données requis par les principaux logiciels de textométrie. (*Cf.* Daoust, 2006). On lira également sur ces questions la contribution de S. Heiden (2006).

[12] Notons qu'une telle segmentation peut être pratiquée sur une chaîne incluant des zones péritextuelles. Certains des types référeront alors à des unités péritextuelles (ouverture de balise, contenu de balise, attribut, valeur d'attribut, fermeture de balise). Ces occurrences distinctes des occurrences de la zone proprement textuelle.

peut constituer un tableau à double entrée dont chaque case *(i, j)* contient le nombre des occurrences de l'unité textuelle *i* attestées dans la partie *j* du corpus.

4.1.4. Délimitations de zones dans le texte initial

Les procédures textométriques permettent parfois la délimitation de zones particulières à l'intérieur du corpus de départ (partitions du corpus, zones spécifiques, etc.). À partir d'un système de coordonnées textuelles comme le numéro d'occurrence, il est possible de transmettre ces zones d'un logiciel à l'autre pour les soumettre à des procédures spécifiques.

4.2 Structurer et gérer les résultats intermédiaires

Deux conditions se révèlent indispensables pour la transmission de ces résultats intermédiaires :

- La structuration des données transmises ;
- La description[13] des données produites aux différentes étapes (documentation, métadonnées…).

Les produits d'une analyse textométrique peuvent être ensuite organisés parmi l'ensemble des données manipulées ou produites au cours de l'analyse. Gérer et classer les produits d'une analyse textométrique, c'est organiser l'ensemble des données manipulées ou produites au cours de l'analyse. La transcription utilisée s'inscrit dans la démarche de normalisation des documents électroniques via XML.

Une analyse textométrique sur un texte T peut être complètement décrite si les éléments suivants sont clairement identifiables :

Objets	Type de description
Fichiers (Base textométriques, méthodes, produits), Url associée aux fichiers	Description des données et des méthodes utilisées lors de l'analyse, des produits construits au cours de cette analyse et enfin des identifiants permettant de localiser les fichiers disponibles
Classe des produits	Description des produits issus de l'analyse : données, paramètres et résultats construits
Commentaire	Textes complémentaires ou commentaires divers
Historique des traitements	Description séquentielle de l'historique des traitements
Liens sur les produits	Index général des produits et identifiants permettant de les localiser dans la base textométrique
Métadonnées	Ensemble des informations techniques et descriptives associées aux ressources numériques manipulées afin de décrire leur contenu et leur format

5. Conclusion

Pour dépasser la comparaison « qualitative » de stratégies d'analyse textométrique se confondant souvent avec l'application d'une série de méthodes appliquées de manière standardisée et qui aboutissent à des résultats hétérogènes, nous proposons d'avancer vers une

[13] On voit se développer d'importants travaux autour de la notion de *métadonnées*. Les métadonnées permettent de cataloguer les ressources électroniques, et donc de pouvoir les retrouver efficacement et les réutiliser de façon pertinente. Pour répondre à ce besoin de standardisation, plusieurs organismes ont proposé des schémas de métadonnées conçus pour être adoptés par le plus grand nombre. On retrouve notamment les normes du *Dublin Core* (DC) et de *l'Open Language Archive Community* (OLAC) (Bird, 2004).

description plus explicite qui porterait à la fois sur les données qui servent de base aux traitements textométriques et sur les différentes étapes qui composent ces traitements.

Une condition importante pour avancer dans ce sens consisterait en un découpage des logiciels d'analyse textométriques en modules clairement isolés et susceptibles d'échanger, à différents niveaux, des résultats de traitements intermédiaires produits par d'autres logiciels.

Les stratégies d'analyse capables d'exhiber une auto-description sur les objets et les données qu'elles manipulent devraient alors permettre de comparer avec beaucoup plus de profit les traitements qu'elles mettent en œuvre et les résultats auxquels elles aboutissent.

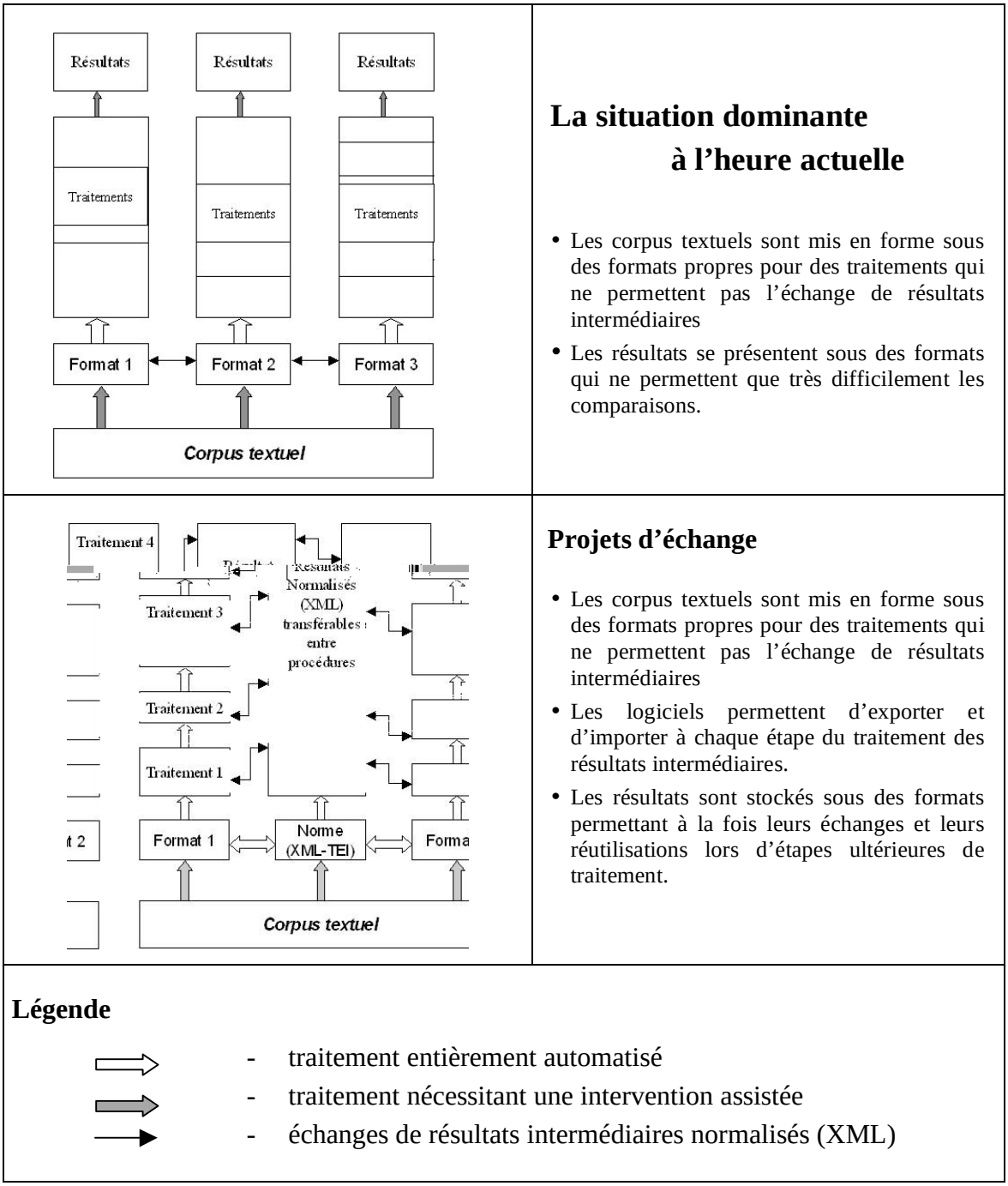

Figure 2 : *Possibilités actuelles et stratégies futures pour la comparaison des résultats textométriques*

Références

ATONET - Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur : http://www.ling.uqam.ca/forum/atonet/

DUBLIN CORE - Norme de métadonnées pour le codage des ressources informatiques: http://dublincore.org/

OLAC - Open Language Archives Community : http://www.language-archives.org/

TEI - Text Encoding Initiative : http://www.tei-c.org/

XML - Extensible Markup Language : http://www.w3.org/XML/

Bird S. and Simons G. (2004). Building on Open Language Archives Community on the DC Foundation. In Hillman and Westbrooks (editors), *Metadata in Practice: A Work in Progress*, ALA Editions : 203-222.

Bonhomme P. (2000). *Codage et normalisation de ressources textuelles.* Ingénierie de langues. J. M. Pierrel. Paris, Hermès.

Daoust F. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés. In *Actes des 8es journées d'analyse statistique des données textuelles*, Besançon.

Habert B., Fabre C. et Issac F. (1998). *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*. Paris, InterÉditions/Masson.

Habert B., Nazarenko A., Salem A. (1997). *Les linguistiques de corpus.* Paris, Armand Colin/Masson.

Harold E.R., Means W.S. (2001). *XML in a nutshell,* O'REILLY.

Heiden S. (2006). Modèles de données et formats d'échange pour l'interopérabilité des outils en textométrie. In *Actes des 8es journées d'analyse statistique des données textuelles*, Besançon, 2006.

Ide N., Véronis J. (1996). Une application de la TEI aux industries de la langue : le Corpus Encoding Standard. In *Cahiers GUTenberg 24*.

Ide N., Véronis J. (1996). Présentation de la TEI : Text Encoding Initiative. In *Cahiers Gutenberg 24* : 4-10.

Véronis J. (2000). *Annotation automatique de corpus : panorama et état de la technique*. Ingénierie de langues. J. M. Pierrel. Paris, Hermès.