Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée

Mathieu Valette[1]

[1]INALCO – ERTIM – mvalette@inalco.fr

Abstract

Our purpose in this article is to compare the Statistical Analysis of Textual Data and the Natural Language Processing by regarding several points of view: epistemological, methodological and applicative, and studying objects. The aim is to clarify the relations between these sub-disciplines in order to consider the possible conciliations: (i) automatics vs. hermeneutics; (ii) techne vs. episteme; (iii) tests vs. acceptability judgments; (iv) algorithms vs. ergonomics; (v) corpus vs. corpus as resources as sources.

Résumé

Notre propos dans cet article est de comparer le TAL et l'ADT selon des points de vue gnoséologique, méthodologique, applicatif et des objets d'étude (texte, corpus). Il s'agit d'expliciter, au moyen de cinq positions antagonistes, les relations entretenues par ces deux sous-disciplines pour envisager les terrains de conciliation possibles : (i) automatisation *vs* herméneutique ; (ii) tekhnè *vs* épistémè ; (iii) test *vs* jugement d'acceptabilité ; (v) algorithmique *vs* ergonomie ; (v) corpus comme ressources *vs* corpus comme sources.

Key words: NLP, Statistical Analysis of Textual Data, methods, discussion.

1. Introduction[1]

Bien que parfois associés, voire partiellement confondus, notamment par les linguistes, l'analyse statistique des données textuelles (ADT) et le traitement automatique des langues (TAL) entretiennent aujourd'hui des relations fort ténues et parfois circonspectes sinon hostiles. Leurs quelques points communs (l'algorithmique, les corpus numériques) et des affinités intercommunautaires ponctuelles suffisent pour que l'on se pose la question de leur relations possibles, impossibles et souhaitables. Notre propos dans cet article est de caractériser ces sous-disciplines selon des points de vue gnoséologique, méthodologique et applicatif. Il s'agit, au moyen de quelques positions *a priori* antagonistes, d'expliciter les rapports entretenus pour envisager les terrains de conciliation possibles. On peut résumer les antagonismes soulevés de la manière suivante : (i) *Automatisation* vs *herméneutique* : le TAL vise l'automatisation des processus tandis que l'ADT repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes. (ii) *Tekhnè* vs *épistémè* : le TAL est aujourd'hui essentiellement *utilitariste* et a pour finalité la production d'applications ; l'ADT a des objectifs épistémiques. (iii) *Test* vs *jugement d'acceptabilité* : à la différence du TAL où la mise en place d'un protocole d'évaluation est indispensable, l'évaluation et la reproductibilité ne sont pas problématisées par l'ADT. (v) *Algorithmique* vs *ergonomie :* les

[1] J'ai plaisir à remercier François Stuck et Egle Eensoo pour leur aide dans la réalisation de cette étude et les relecteurs des JADT pour leurs remarques constructives.

praticiens de l'ADT ne sont pas des informaticiens mais, en majorité, des utilisateurs finaux de logiciels dotés d'interfaces graphiques. (iv) *Corpus comme ressources* vs *comme sources* : l'ADT porte un soin particulier à la description philologique des corpus de textes entiers tandis que le TAL identifie des ressources générales permettant de produire des corpus considérés comme des réservoirs d'objets linguistiques infratextuels.

L'étude s'appuie ponctuellement sur l'analyse textométrique d'un corpus diachronique de textes académiques : la totalité des actes de deux conférences communautaires francophones emblématiques : les *Conférences en Traitement Automatique de la Langue Naturelle (TALN),* et les *Journées internationales d'Analyses statistiques des Données Textuelles (JADT)*. Le corpus couvre toute la période disponible électroniquement, c'est-à-dire de 1997 à 2015. La périodicité des deux conférences n'est pas la même : TALN est annuelle et JADT bisannuelle. Sur un ensemble de 1 462 articles, on compte 474 articles issus du sous-corpus JADT et 988 articles issus du sous-corpus TALN. Le nombre total d'occurrences de formes du corpus est de 6 057 610. S'agissant d'une réflexion épistémologique générale et non d'une étude de corpus à proprement parler, nous n'expliciterons nos résultats d'analyse textométrique que de façon marginale[2].

2. Automatiser ou décrire

Comme son nom l'indique, le TAL a pour objectif l'automatisation, c'est-à-dire l'élimination de la part de l'humain dans les traitements linguistiques, tandis que l'ADT est une linguistique assistée par ordinateur. Nombre de travaux en TAL consistent à mettre en place une chaîne de traitement permettant, à partir d'un ensemble de données injectées en entrée, d'obtenir un résultat distant en sortie. Parmi les améliorations des systèmes proposées par les talistes, réduire l'intervention humaine apparaît souvent prioritaire, après l'amélioration des performances brutes. Les tâches effectuées par les humains sont par tradition qualifiées de *manuelles*[3]. L'antonyme en est certes *automatique,* mais peut s'entendre également comme *intellectuel*, auquel cas il n'est pas exclu que *manuel* soit connoté de manière dépréciative.

Le travail produit par la pensée humaine nue constitue en effet une pierre d'achoppement du TAL. Une tâche *manuelle* par excellence est la lecture du corpus. Le taliste ne lit pas son corpus, non seulement parce que cela n'apparaît pas nécessaire à la réalisation de son projet, mais plus encore, il en fait un point de méthode : ne pas lire son corpus signifie s'en tenir à distance. L'effacement du lecteur-interprète le conduit à élaborer des stratégies d'objectivation. Par exemple, les normes déontologiques relatives à l'annotation des corpus (indispensable pour les méthodes validées par apprentissage supervisé, très majoritaires aujourd'hui[4]), imposent que les annotateurs ne participent pas à la mise en place des méthodes. La connaissance des corpus induirait un biais dans les choix algorithmiques et fausserait les résultats. Ainsi, les tâches d'interprétation sont-elles externalisées[5]. Certes,

[2] Les analyses ont été réalisées au moyen du logiciel Lexico 3 (Salem *et al*., 2003).

[3] Indice de ce clivage constitutif, 'manuel'- (*manuel*, *manuellement*) est un des morphèmes lexicaux les plus caractéristiques du sous-corpus TALN en comparaison avec le corpus JADT.

[4] Lire (Yvon 2006) pour une analyse approfondie et (Church, 2011) pour une étude sur l'opposition entre empirisme et rationalisme dans l'histoire du TAL.

[5] Sur l'annotation de corpus, on se réfera au travail approfondi de (Fort, 2012).

l'annotation de corpus pourrait s'apparenter à de l'analyse de corpus dans une perspective ADT, mais on verra ci-dessous à quel point les méthodologies et les pratiques sont distantes.

L'annotation comme description ?

Souvent, les talistes confient les tâches d'annotation à des tiers académiques variés. Selon les moyens mis en œuvre et l'enjeu considéré, ces travaux d'annotation (donc de lecture) peuvent être délégués à des linguistes en poste, ce pourquoi on assiste aujourd'hui à un dangereux glissement de sens selon lequel, pour les talistes, l'activité scientifique des linguistes serait l'annotation. Lorsque les linguistes ne sont pas disponibles ou disposés à annoter, on fait appel à des stagiaires, des étudiants, des prestataires (ELDA, *Amazon Mechanical Turk*), en bref – et les intéressés excuseront ce raccourci peu aimable – on *déqualifie* le travail d'annotation[6].

On pourrait certes penser que ces tâches sont de haut niveau, notamment quand les annotations sont dites « fines »[7], mais elles sont, de fait, reléguées au rang de sous-traitance. En témoigne la division du travail dans les campagnes d'annotation rapportée par (Bontcheva *et al.*, 2010). Les auteurs distinguent trois rôles : le chef de projet, l'éditeur et les annotateurs. Le chef de projet sélectionne les corpus, définit l'enchaînement des opérations, contrôle leur bon déroulement. Il est également en charge du guide d'annotation (parfois avec l'éditeur), c'est-à-dire les consignes et les jeux d'étiquettes mis à disposition des annotateurs. Il arrête également des choix méthodologiques comme le nombre d'annotateurs, le nombre de documents à traiter pour chaque annotateur, les prétraitements et notamment d'éventuelles pré-annotations automatiques destinées à réduire la part de travail manuel en aval. L'éditeur, quant à lui, supervise l'annotation et fait l'interface entre le chef de projet et les annotateurs. Il les forme au guide d'annotation, contrôle la qualité des annotations et fait office d'*expert*. Enfin les annotateurs, qui n'ont pas de statut d'expert, se conforment au guide d'annotation. L'annotateur correspond à l'ouvrier non qualifié de la chaîne de traitement et est suspecté structurellement d'en être l'élément faible : il peut faire des erreurs d'inattention ou ne pas comprendre le guide d'annotation. C'est la raison pour laquelle on multiplie les annotateurs pour un même corpus, comptant sur le calcul d'un accord inter-annotateurs. L'accord inter-annotateurs est un point de méthodologie crucial dans le TAL. Il s'agit d'une mesure statistique de la congruence d'annotations obtenues indépendamment les unes des autres (*kappa* de Cohen). Jamais remis en question, l'accord inter-annotateurs apparaît en effet comme un paradoxe, au moins dans les termes, car il doit être obtenu sans que les annotateurs communiquent entre eux, c'est-à-dire s'accordent sur leurs annotations. Autrement dit, il s'agit d'un système de vote à bulletin secret amélioré. Or, on pourrait très bien imaginer une méthode dialectique visant la recherche d'un consensus entre les annotateurs, mais cela reviendrait à leur donner la possibilité de mettre en défaut un guide d'annotation lacunaire ou imprécis. En somme, les seuls échanges possibles, pour les annotateurs, le sont avec l'éditeur. Le guide d'annotation est produit en amont, sans retour possible de la « base » pourtant en contact direct avec les documents. Ainsi, la parcellisation des tâches et la division du travail à

[6] Lire (Sagot *et al.*, 2011) à propos des campagnes de *crowdsourcing* et du *Mechanical Turk* d'*Amazon* plus particulièrement.

[7] « Annotation fine » est la mauvaise traduction de *fine-grained annotation*, soit « annotation *à grain fin* ». L'adjectif *fin* porte sur le grain, c'est-à-dire sur l'unité mot, et non l'annotation. Celle-ci, de fait, manque souvent singulièrement de finesse (voir *infra*, la note 12).

l'œuvre dans les campagnes d'annotation relève-t-elle d'un discret taylorisme en parfaite adéquation avec les présupposés épistémologiques du TAL.

L'ADT constructiviste

Si l'annotation de corpus en TAL correspond à une analyse du corpus, le profond fossé qui sépare le TAL de l'ADT n'aura pas échappé au lecteur. Au prisme de ce qui a été dit, l'ADT relève d'une tradition descriptive artisanale. Elle ne se focalise pas sur l'amélioration de traitements automatisés mais se fonde sur l'utilisation d'instruments d'observation et une interprétation itérative allant du texte aux résultats de traitements statistiques. Si lesdits traitements sont souvent peu élucidés, le linguiste praticien de l'ADT (le textomètre notamment) doit *connaître son corpus* pour pouvoir l'interpréter, à l'inverse du taliste. Le corpus est donc ici investi subjectivement, personnellement, quand celui du taliste au contraire *ne doit pas l'être*. En somme, le corpus en ADT est une *construction*, l'observateur est inclus dans l'observation, suivant en cela le paradigme néo-cybernétique (ou constructiviste) du *système observant* imaginé par (Von Foerster, 1981) tandis que le TAL demeure fortement attaché au paradigme traditionnel de la séparation de l'objet et du sujet. Il n'est pas impossible que le va-et-vient méthodologique entre vues macroscopiques[8] et focus sur le texte (concordances, recherche de patrons) ne soit l'indice de cette co-construction de l'objet et de l'observateur.

Le vif débat relatif à la lemmatisation qui anima la communauté ADT au début des années 2000 (Brunet, 2000, 2002, Labbé, éd., 2003) est selon nous un indice de cette relation particulière qui lie l'observateur à son objet. Complètement absent dans la communauté TAL, il portait sur l'altérité de l'objet et posait deux questions conjointes : d'une part, la question de la sauvegarde de l'objet car la lemmatisation corrompt le texte en en donnant des représentations qui seront appauvries si l'on n'en sélectionne qu'une (par exemple les seuls lemmes sans parties du discours) ; d'autre part, la question de la sauvegarde de l'observateur-constructeur dans la mesure où les lemmatiseurs sont assimilables à des observateurs tiers. Ainsi, en lemmatisant son corpus, le linguiste en perd partiellement le contrôle et le met à distance, embrassant de la sorte une partie du paradigme TAL de l'objectivation des données. Indubitablement, l'enrichissement des corpus, devenu banal en TAL, progresse en ADT suivant le principe selon lequel, si c'est possible, autant le faire. La dernière génération de logiciels de textométrie (TXM, Le Trameur, Iramuteq, etc.) lemmatise les corpus par défaut.

Tout se passe comme si l'historique déférence philologique de l'ADT vis-à-vis de la *forme*, évoluait aujourd'hui, vraisemblablement sous l'influence du TAL. Les possibilités d'enrichissement ou d'appauvrissement subséquent s'accroissent[9]– et bientôt, les ressources sémantiques courantes en TAL mais encore rares en ADT pourraient se développer[10].

Cette mise à distance du corpus s'observe également par le biais d'initiatives émanant cette fois davantage de la linguistique de corpus que du TAL. Elles visent à rendre les corpus

[8] Par exemple l'analyse factorielle des correspondance (Benzécri, 1973), la classification descendante hiérarchique (Reinert, 1983), l'analyse arborée (Barthélémy *et al.*, 1987), etc.

[9] Iramuteq, par exemple, propose par défaut de ne sélectionner que les mots pleins (qualifiés de « formes actives » : noms, verbes, adjectifs, adverbes, mots inconnus) et de supprimer les mots outils (qualifiés de « formes supplémentaires »).

[10] Cf. les logiciels Prospero ou Tropes qui en utilisent depuis de nombreuses années déjà.

partageables, c'est-à-dire établis suivant des formats et des normes d'échange (XML TEI) et disponibles sur des plateformes académiques. Cette philologie numérique, fortement patrimoniale, est globalement étrangère au TAL. Dans notre corpus de conférences, la TEI (*Text Encoding Initiative*) fut signalée précocement dans le cadre des conférences TALN mais elle ne connaît un véritable essor, que dans les colloques JADT, en particulier de 2002 à 2010. Le format d'échange XML, qui relève davantage de la structuration que de l'édition, est adopté par les deux communautés, même si les tenants de l'ADT en font plus souvent mention positivement. Très fortement soutenue, voire commanditée institutionnellement, la patrimonialisation des corpus textuels numérisés laisse ouverte la question de leur réutilisabilité effective dans la mesure où, produits pour des tâches et par des observateurs spécifiques, leur pertinence pour d'autres tâches n'est pas assurée. Quoi qu'il en soit, la philologie numérique cantonne pour l'heure les linguistes dans le rôle de producteur de corpus plus que d'inventeur de modèles pour leur analyse.

3. Tekhnè *vs* épistémè

Au début des années 2000 encore, les épistémologues opposaient le TAL « robuste », « statistique » ou encore « applicatif » au TAL « formel », « linguistique » ou « théorique » (Cori *et al.*, 2002). Ce dernier visait à simuler au mieux des phénomènes linguistiques en s'adossant à des formalismes théoriques (par exemple les grammaires de dépendance formelles) et relevait d'une épistémologie confirmatoire. Mais depuis une dizaine d'années, le TAL statistique a accompagné la demande sans cesse croissante en applications liées à internet et à la dématérialisation des supports. Cette demande implique le renouvellement des problématiques de recherche applicative : il y a 15 ans, l'extraction d'informations lexicales ou syntagmatiques destinées à alimenter des bases de connaissances (mémoires de traduction, terminologies de métier, systèmes de question-réponse, etc.) structurait le champ. Puis, avec l'essor des réseaux sociaux sur le web, des applications en fouille d'opinion, analyse des sentiments, analyse du *buzz*, etc. se sont développées. La traduction automatique, historiquement liée au TAL symbolique, connaît également un regain d'intérêt motivé par l'efficacité des méthodes statistiques.

Devenu utilitariste et ayant aujourd'hui pour finalité essentielle des applications, le TAL doit composer avec un impératif d'efficacité qui implique une recherche systématique de performance et d'optimisation. La mise en place d'un protocole d'évaluation est indispensable à toute recherche en TAL (lire *infra* § 5)[11]. Les expériences doivent être reproductibles, ce pourquoi la mise à disposition des jeux de données devient aujourd'hui une obligation éditoriale dans la communauté. Mieux encore, des jeux de données (déjà annotées, c'est-à-dire *interprétées*, voir § 2), sont parfois distribués en amont d'une conférence. Les chercheurs sont invités à appliquer leurs algorithmes sur ces données partagées. C'est le cas de campagnes telles que SemEval (*Semantic Evaluation*) ou DEFT (*Défi en Fouille de Textes*). On comprend bien l'intérêt de cette pratique de mise en commun des données. Elle est émulative mais également économique car l'annotation, comme on l'a vu, est coûteuse. Toutefois, elle ne préjuge en rien de la qualité des données. Aux côtés des corpus de référence (*gold standard corpus*), c'est-à-dire des corpus dont les annotations ont été validées pour servir consensuellement de référence à d'autres, on trouvera des données médiocrement

[11] Dans le corpus de conférences que nous avons analysé, sur 3381 occurrences de la forme *évaluation*, seules 650 appartiennent au sous-corpus JADT. C'est le deuxième item le plus spécifique du sous-corpus TALN (seuil de 5) après *TA* pour « Traduction Automatique ».

annotées, voire franchement erronées, en particulier quand une pré-annotation automatisée n'a pas fait l'objet d'un contrôle suffisant[12]. Ces données mal annotées posent alors un problème d'éthique scientifique : le chercheur est-il tenu d'assumer les données qu'il n'a pas produites mais qu'il exploite à des fins de traitement ? Là encore, la mise à distance des données ressortit, *in fine*, à une déresponsabilisation du chercheur. Qu'importeraient les données pourvu que les algorithmes les traitent ?

Fondamentalement observatoire et descriptive comme on l'a vu précédemment, l'ADT, pour laquelle les applications ne sont que marginalement un enjeu, a des objectifs épistémiques : accroître les connaissances sur un corpus, participer à son interprétation. L'évaluation et la reproductibilité ne sont pas problématisées par l'ADT. Les analyses sont validées par homologation, c'est-à-dire par l'assentiment d'une communauté, qui peut ne pas être celle de l'ADT (cf. par exemple, le très bon accueil fait par les vingtièmistes aux travaux de (Kastberg, 2006) portant sur l'œuvre de Le Clézio). Attention toutefois que, très souvent, l'assentiment communautaire s'apparente à un simple avatar du « jugement d'acceptabilité » contre lequel s'est pourtant dressée – et même instituée – la linguistique de corpus jadis. On entend par là que la plupart des études en ADT sont irréfutables, c'est-à-dire qu'aucun protocole n'est prévu pour juger qu'elles sont vraies ou fausses. Or, l'irréfutabilité, sans contre-analyse possible dans la plupart des études textométriques, pose un problème de méthode sinon de déontologie. Pour une spectaculaire affaire Corneille/Molière (Labbé *et al.*, 2001) longuement débattue dans les communautés concernées, combien d'études d'ADT approximatives mais jamais récusées ? combien de sorties logicielles contrariant l'objectif analytique passées sous silence ? Il n'est pas anodin qu'une des formes les plus spécifiques du sous-corpus JADT soit *évidence*, constituant de syntagmes tels que « mettre en évidence » ou « mise en évidence ». L'ADT impose des évidences mais n'est jamais en mesure de prouver. Malgré toutes les réserves que l'on peut émettre concernant les méthodes d'évaluation du TAL, convenons qu'elles ont le mérite d'exister.

4. Algorithmique *vs* ergonomie

En matière d'innovation, le TAL et l'ADT suivent des pentes semble-t-il inverses. Dans le TAL théorique, qui établissait des relations entre des symboles – qu'ils fussent logiques ou qu'ils relevassent d'autres formalismes – l'épistémologie linguistique dominait. Bien qu'il traite des données *réelles*, le TAL statistique a significativement dévalué l'apport de la linguistique théorique. Bon nombre d'algorithmes sont plus ou moins indifférents à la nature des données et peuvent aussi bien traiter des données démographiques, bio-informatiques que textuelles. Seuls quelques algorithmes sont conçus pour le traitement de données textuelles (par exemple l'analyse sémantique latente, Landauer *et al.*, 1998). Par ailleurs, la discipline, portée, il est vrai, par le regain d'intérêt dont fait l'objet l'intelligence artificielle aujourd'hui et aidée par la montée en puissance constante des ordinateurs, s'avère très créative en termes d'algorithmes (cf. par exemple les récents modèles Word2Vec).

À l'inverse, dans l'ADT, les méthodes mathématiques utilisées, qui satisfont le plus grand nombre, évoluent peu depuis 30 ans, mais les heuristiques et les savoir-faire analytiques, eux,

[12] On lira à ce propos la critique des données utilisées pour la campagne DEFT 2015 dans (Eensoo *et al.* 2015, 2-4). Les auteurs relèvent notamment d'importants contresens comme par exemple, le tweet « *Elles font fureur... Leur toucher doux, leur couvercle cristal, leur respect d'environnement ! URL* » annoté comme exprimant un sentiment de colère, vraisemblablement parce que le mot *fureur* y est présent.

sont très étudiés et, en quelque 10 ans, la discipline a été marquée par l'arrivée de nouvelles applications intégrées dans lesquelles sont particulièrement soignés tantôt l'ergonomie (par exemple TXM, Heiden *et al.*, 2010), tantôt la définition des unités (par exemple le Trameur, Fleury *et al.*, 2014) tantôt les outils de visualisation (par exemple Iramuteq, Ratinaud *et al.*, 2012). Si elles implémentent pour l'essentiel des fonctionnalités déjà présentes sur les applications phares du domaine (Lexico 3, Salem *et al.*, 2003 ; Hyperbase, Brunet, 2011), l'accent est mis sur les paramétrages fins, les menus contextuels, l'affordance et l'interopérabilité. C'est qu'aujourd'hui plus encore qu'hier, les praticiens de l'ADT ne sont pas des informaticiens mais, majoritairement, des utilisateurs finaux de logiciels dotés d'interface graphique permettant souvent la manipulation des outils que des informaticiens développent ou utilisent soit pour leurs propres tâches, soit à l'adresse de la communauté des utilisateurs. Ceux-ci cheminent alors de fonctionnalité en fonctionnalité de façon à élaborer des parcours interprétatifs. Lorsque leur pratique permet de constater une carence des outils, ils se tournent vers les informaticiens qui en prennent compte. Un exemple édifiant est l'analyse des cooccurents au moyen d'un test d'écart-réduit, élaborée par (Bourion, 2001) pour identifier les isotopies et les molécules sémiques (Rastier, 2001) dans un corpus. Un premier outil fut réalisé à l'INALF par Jean Maucourt, puis Étienne Brunet l'intégra dans Hyperbase sous le nom de fonction THEME. La méthode implémentée est restée relativement conservatrice mathématiquement mais n'a pas cessé d'être perfectionnée dans le temps de façon à répondre au mieux aux besoins des linguistes (cf. les paramétrages très fins proposés par TXM). Elle témoigne parfaitement du dialogue entre élaboration théorique, pratique d'analyste et conception du logiciel[13]. Aujourd'hui, la cooccurrence connaît, avec les graphes de cooccurrences, la poly-cooccurrence, la cooccurrence de deuxième ordre, l'asymétrie de la cooccurrence, etc. un renouveau des plus stimulants (cf. Mayaffre *et al*, 2012 pour un panorama récent).

5. Le corpus comme ressource *vs* le corpus comme source

Les pratiques de l'ADT et celles du TAL opposent les notions de *source* et de *ressource* : *auteur* – la source par excellence – est une des formes les plus spécifiques du sous-corpus JADT. D'une manière générale, les documents analysés en ADT sont variés et souvent caractérisés avec précision. À la fin des années 1990, les œuvres littéraires dominaient (*romans*, *poésie*, *théâtre*) mais on étudiait aussi des *enquêtes* ouvertes, des textes *politiques*, *syndicaux*, etc. Au milieu des années 2000, les nouveaux genres de l'Internet font leur apparition (mails, puis forums de discussion, *tweets*). On retrouve en partie ces types documentaires en TAL (très rarement les textes littéraires), mais les textes à vocation technique ou encyclopédique, tels que *Wikipédia,* apparaissent privilégiés. Surtout, davantage que des sources précises (*i.e.* des auteurs, des œuvres ou des éditeurs électroniques, des sites web, etc.), ce sont des ressources générales ou des portails qui sont désignés dans notre sous-corpus TALN : *Internet*, *Web*, *Google*, *Google Books*, *Facebook*, etc. Il faut dire que dans les applications TAL destinées à l'extraction d'informations, les corpus sont avant tout des réservoirs d'objets linguistiques infratextuels (*termes*, *phrases*, *structures prédicatives*, etc.). L'établissement philologique du corpus en TAL est souvent réduit à quelques valeurs quantitatives (nombre d'occurrences de mots, nombre de textes) quand les textomètres présentent – en général – leur corpus de manière plus qualitative (description des auteurs, des genres textuels, etc.).

[13] Lire (Pincemin, 2012) à qui nous empruntons cet exemple pour un développement.

Par ailleurs, l'inclination apprentiste qu'a suivie le TAL ces dernières années a profondément accentué les différences culturelles liées à l'utilisation et à la fonction du corpus. Les méthodes d'apprentissage automatique supervisé, privilégiées en TAL, consistent à créer un modèle reproduisant la configuration optimale des données du corpus, quelles qu'elles soient. Si, dans une tâche de classification de textes, par exemple, un corpus est composé de deux classes (*i.e.* deux sous-corpus exclusifs l'un de l'autre) et que la tâche consiste à classer des textes dans une classe plutôt que dans l'autre, l'entraînement du modèle consistera à sélectionner les critères qui caractérisent de façon appropriée les textes d'une classe par rapport à l'autre, quand bien même ces critères ne seraient nullement interprétables d'un point de vue linguistique. Pour une tâche d'annotation des parties du discours, le modèle, une fois entraîné, calculera la probabilité que le mot « ferme », par exemple, soit un substantif ou un verbe conjugué, compte tenu de ce qu'il aura appris dans la phase d'entraînement sur le corpus annoté.

Les bonnes pratiques en matière d'évaluation des résultats évoquées ci-dessus imposent de scinder le corpus annoté en (au moins) deux parties distinctes : le corpus d'apprentissage et le corpus de test sur lequel on a masqué à la machine les annotations[14]. L'algorithme de classification est entraîné sur le corpus d'apprentissage et le corpus de test permet de mesurer la capacité du modèle produit par entraînement à bien restituer les annotations initiales. L'évaluation des performances du système repose ainsi sur les mesures de congruence entre le résultat de la classification et le corpus de test annoté, au moyen de différentes mesures de performance, par exemple les très célèbres mesures de *précision* (qualité), de *rappel* (couverture) – ou leur combinaison, la *f-mesure*. Or, comme l'observe très justement (Yvon, 2006, 41), d'autres évaluations seraient possibles (analyse sémantique des valeurs discriminantes sélectionnées par l'algorithme, adéquation avec une théorie linguistique, plausibilité cognitive, etc.) mais les alternatives sont rares et peu valorisées en termes académiques. De toute évidence, les données linguistiques sont jugées encombrantes et, pour des raisons éditoriales sans doute, mais aussi par manque d'outils intellectuels pour les appréhender, on ne les montre guère (Hall *et al.*, 2008).

Le corpus en ADT n'est pas conçu comme une ressource ni dans une perspective évaluative, mais comme un mode de contextualisation à échelle multiple des phénomènes observables, de la cooccurrence, « forme minimale du contexte » (Mayaffre, 2008) au corpus intégral qui objective l'intertexte (Rastier, 1998) et qui, à mesure qu'il s'élargit, tend vers le contexte extralinguistique qu'il simule (Mayaffre, 2002). Ainsi, les sous-corpus construits ont toujours une fonction différentielle. (Rastier *et al.*, 1999, 84) distinguent « (i) un *corpus existant*, correspondant aux textes accessibles dont [on] peut disposer, (ii) un *corpus de référence*, constituant le contexte global de l'analyse, ayant le statut de référentiel représentatif, et par rapport auquel se calcule la valeur de paramètres (pondérations...) et se construit l'interprétation des résultats, (iii) un *corpus de travail*, ensemble des textes pour lesquels on veut obtenir une caractérisation, et le cas échéant (iv) un *corpus d'élection*, sous-corpus du corpus de travail, contrasté par rapport à celui-ci ». Alors que les corpus en TAL sont scindés par rapport à l'opposition analyse / évaluation, ceux de l'ADT sont inclusifs et entièrement dédiés aux tâches descriptives et analytiques (le concept de *sous-corpus* apparaît spécifique à l'ADT dans le corpus que nous avons analysé).

[14] Dans le sous-corpus d'étude « TALN », les cooccurrents spécifiques du mot-pôle *corpus* (fenêtre : paragraphe) relèvent de cette méthodologie : *test*, *apprentissage*, *entraînement*, *évaluation*.

6. Conclusion

Admettons que le TAL – c'est notre credo – offrirait aujourd'hui de sérieux débouchés sociaux et économiques à la linguistique si le rôle de celle-ci ne s'y réduisait pas considérablement depuis quelques années. Il s'apparente en effet, de plus en plus, à un rôle de sous-traitance dans le TAL applicatif. Les applications en fouille de textes, par exemple, ont peu recours à la théorisation linguistique et encore moins aux théories du texte (linguistique textuelle, analyse du discours, sémantique interprétative, etc.). Le TAL actuel transforme peu à peu les linguistes en *annotateurs experts* et écarte tacitement leurs propositions théoriques du domaine. En élaborant de puissantes méthodes d'extraction des connaissances et de classification de textes, c'est lui qui, aujourd'hui, est confronté à la complexité des textes, à la textualité. Les fameux linguistes en chaise longue (*« armchair linguists »*, Fillmore, 1992), en négligeant les données empiriques, sinon à des fins de vérification *a posteriori*, ont évidemment une part de responsabilité dans cette situation mais les choix épistémologiques de l'ADT ont également une incidence sur la pauvreté du dialogue entre les deux communautés. Pourtant, les linguistes de corpus et de l'ADT, sensibles au texte et soumis à son principe de réalité, sont sans doute plus à même de résister à cette prolétarisation rapide de la discipline, à condition toutefois qu'ils prennent la mesure des enjeux applicatifs du TAL et de son exigence, très perfectible mais indubitablement vertueuses, d'évaluation.

Références

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

Benzécri, J-P. (1973). *L'Analyse des Correspondances*. Paris, Dunod.

Bontcheva, K., Cunningham, H., Roberts, I., Tablan V. (2010). « Web-based collaborative corpus annotation: Requirements and a framework implementation ». *Proceedings of the workshop on New Challenges for NLP Frameworks*, Workshop at LREC 2010, Valletta, Malta, May 22.

Bourion, É. (2001). *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Université de Nancy II.

Brunet, É. (2000). « Qui lemmatise dilemme attise », Lexicometrica, n°2.

Brunet, É. (2002). « Le lemme comme on l'aime », in A. Morin & P. Sébillot (dir.), 6e Journées d'analyse des données textuelles, vol. 1, Rennes, IRISA : 221-232.

Brunet, É. (2011). *Ce qui compte. Méthodes statistiques*, textes édités par Céline Poudat, préface de Ludovic Lebart, Paris, Champion, 2011.

Church, K. (2011). « A pendulum swung too far ». *Linguistic Issues in Language Technology*, 6, Submitted version, October 2011.

Cori, M., David, S. et Léon, J. (2002). « Pour un travail épistémologique sur le TAL », *TAL*, 43/3 : 7-20.

Eensoo, E., Nouvel, D., Martin, A., Valette, M. (2015) « Combiner analyses textométriques, apprentissage supervisé et représentation vectorielle pour l'analyse de la subjectivité », *Actes du 11e Défi Fouille de Texte (DEFT'2015)*, Caen (France).

Fillmore, Ch. J. (1992). « "Corpus linguistics" or "Computer-aided armchair linguistics" », *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, J. Svartvik (eds), Mouton de Gruyter : 35-60.

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*, thèse de doctorat, Université Paris Nord (Paris 13).

Hall, D., D. Jurafsky, et C. D. Manning (2008). « Studying the history of ideas using topic models ». In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)* : 363–371.

Heiden, S., Magué, J.-P. & Pincemin, B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », I. C. Sergio Bolasco (éd.)*, JADT 2010*, vol. 2 : 1021-1032.

Kastberg Sjöblom, K. (2006). *L'écriture de J.M.G. Le Clézio. Des mots aux thèmes*. Paris, Honoré Champion.

Labbé, C., Labbé, D. (2001). « Inter-Textual Distance and Authorship Attribution Corneille and Molière ». *Journal of Quantitative Linguistics*. 8/3 : 213-231.

Labbé, D., éd. (2003). *Autour de la lemmatisation*, *Lexicometrica*, n° spécial.

Landauer, T. K., Foltz, P.W., Laham, D. (1998). « Introduction to Latent Semantic Analysis », *Discourse Processes*, vol. 25 : 259-284.

Mayaffre, D. (2002). « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus*, 1 : 51-69.

Mayaffre, D. (2008). « De l'occurrence à l'isotopie. Les cooccurrences en lexicométrie ». *Textes, documents numériques, corpus. Pour une science des textes instrumentée,* M. Valette, éd., *Syntaxe et sémantique*, 9 : 53–72.

Mayaffre, D., Viprey, J.-M., éd. (2012). *La cooccurrence, du fait statistique au fait textuel*, *Corpus,* 11.

Pincemin B. (2012). « Sémantique interprétative et textométrie » (version française complète), *Texto! Volume XVII, n°3,* coordonné par Christophe Cusimano.

Rastier, F. (1998). « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, 129 : 97-111.

Rastier, F. (2001). *Arts et sciences du texte*, PUF, Paris.

Rastier, F., Pincemin, B. (1999). « Des genres à l'intertexte », *Cahiers de Praxématique*, I. Kanellos, éd., *Sémantique de l'intertexte*, 33 : 83-111.

Reinert, A. (1983). « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte ». *Les cahiers de l'analyse des données*, 8/2 : 187-198.

Sagot, B., Fort, K., Adda, G., Mariani, J. Lang, B. (2011). « Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé », *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2011)*, Montpellier (France)

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. & Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle, Manuel d'utilisation*, Sorbonne nouvelle – Paris 3.

Von Foerster, H. (1983). *Observing Systems*, Seaside, CA, Intersystems Publications.

Yvon, F. (2006). *Des apprentis pour le traitement automatique des langues*. Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.