Évolutions de la linguistique outillée : méfaits et bienfaits du TAL

Ludovic Tanguy et Cécile Fabre
CLLE-ERSS : CNRS & Université de Toulouse

Introduction

Cet article examine l'impact des technologies du traitement automatique des langues (TAL) sur les pratiques de la linguistique outillée, au sens de Habert (2004). Cette linguistique qui s'est dotée de nouveaux instruments d'observation et de calcul est largement alimentée par les méthodes définies dans le champ du TAL, héritant ainsi des évolutions de cette discipline pour le meilleur et pour le pire. En effet, si la linguistique a l'opportunité d'intégrer dans son outillage des instruments plus sophistiqués basés sur des approches quantitatives et des données plus diverses et plus nombreuses, elle a néanmoins des difficultés à se les approprier et à contribuer en retour aux avancées du TAL. Nous rendons compte dans cet article de ce bilan contrasté.

Nous présentons tout d'abord un rapide panorama des évolutions récentes, en évoquant notamment la façon dont le développement des méthodes quantitatives du TAL a modifié la place dévolue à la linguistique dans cette discipline. Nous exposons les problèmes que nous percevons actuellement : un TAL à l'outillage trop complexe et opaque pour être appréhendé pleinement par des linguistes, des linguistes dont le rôle se cantonne à l'annotation de données d'entraînement ou d'évaluation, des applications qui simplifient la définition des objets langagiers en misant sur la massification des données pour les traiter.

La deuxième partie s'intéresse à l'impact, cette fois plus positif, qu'ont eu ces évolutions majeures du TAL sur la méthodologie des recherches en linguistique. Elles se traduisent par la possibilité de faire appel à de nouvelles méthodes d'expérimentation et de reconsidérer les phénomènes langagiers à la lumière de données à grande échelle, assorties de méthodes d'observation plus efficaces.

La dernière partie suggère une voie permettant à la linguistique ainsi outillée de jouer pleinement son rôle face à de nouvelles demandes exprimées par des acteurs externes (autres disciplines, entreprises, institutions, etc.). Ces demandes, de natures diverses, ne semblent en effet pas adaptées aux exigences du TAL quantitatif qui s'est concentré sur quelques applications centrales. La linguistique outillée garde alors sa spécificité et son avantage dans des situations qui requièrent des connaissances linguistiques et un traitement plus riche du matériau langagier. Dans ce genre de situation, parce qu'elle a bénéficié des avancées récentes du TAL, la linguistique est aujourd'hui à même d'intervenir efficacement sur de grandes quantités de données et de dialoguer avec d'autres disciplines et acteurs de terrain. Elle a alors l'opportunité de faire émerger des problématiques linguistiques sur des données souvent très riches, et de traiter des questions nouvelles, dont nous donnons plusieurs exemples.

1 Changements méthodologiques dans le traitement informatisé des données langagières

La plupart des nombreux auteurs qui se sont penchés sur l'évolution récentes du traitement automatique des langues n'hésitent guère à parler de changement de paradigme, voire de révolution (Abney, 2011; Church, 2011). L'argument principal consiste à remarquer que l'entrée en

scène des méthodes par apprentissage a modifié les composants essentiels de la culture scientifique disciplinaire des technologies du langage, que ce soit les approches concrètes pour aborder un problème, le rapport aux données, la notion même de modèle, les liens avec les connaissances linguistiques, les pratiques de publication, et bien entendu la formation des étudiants et le bagage de connaissances attendues d'un jeune chercheur de la discipline.

Sans rentrer dans les détails de l'ensemble des secousses produites, nous allons ici évoquer celles qui, selon nous, ont le plus d'impact sur la linguistique outillée, autrement dit sur les pratiques courantes d'un chercheur en linguistique qui se positionne à la fois sur le terrain de l'investigation empirique des mécanismes du langage et comme acteur dans le développement de nouvelles méthodes pouvant intéresser l'ingénierie linguistique.

1.1 Accroissement de la masse de données

La première évolution radicale évoquée ici est celle de l'accroissement des quantités de données textuelles disponibles ou créées, et en tout cas utilisées dans les travaux tant en TAL qu'en linguistique de corpus. La figure 1 montre ainsi (courbe bleue) l'évolution exponentielle des grands corpus de référence disponibles pour l'anglais. Les principaux corpus y sont indiqués (flèches pointillées) aux différents points d'inflexion de la courbe : depuis le million de mots du Brown Corpus du début des années 1960, on a récemment dépassé la centaine de milliards grâce au Web et aux activités de numérisation d'ouvrages de Google.

Cette croissance rapide des volumes de données disponibles a à la fois généré des besoins applicatifs pour gérer cette masse (notamment en recherche d'information, la première sur ce front (Grau et Bellot, 2014)) et permis le développement d'approches massives, en rendant possible des traitements purement quantitatifs efficaces. Cette évolution est présentée et quantifiée dans le précédent numéro de cette revue (Mariani, 2014) et plus spécialement pour les applications qui se sont les premières engagées sur cette voie : la reconnaissance de la parole (Adda-Decker et Estève, 2014) et la traduction automatique (Schwenk, 2014).

Du côté de la linguistique descriptive, les grands corpus de référence ont rapidement reçu une attention soutenue, et le développement de la linguistique de corpus en est une conséquence évidente. Comme on le verra dans la suite, cette disponibilité a eu un impact sur les méthodes comme sur les objectifs. Notons toutefois que dans ce schéma d'ensemble, l'évolution du volume s'est accompagnée de modifications importantes de la nature des corpus. Les premiers grands corpus (Brown, British National Corpus) étaient construits sur la base d'une réflexion préalable sur les types/genres de textes qu'il fallait y intégrer, en visant une bonne représentativité d'un état de langue donné (en l'occurrence l'anglais contemporain). Une seconde génération, surtout sensible en France (où aucun corpus équivalent aux précédents n'a été développé), a concerné la compilation de textes plus homogènes, comme c'est le cas de la base Frantext avec des textes essentiellement littéraires, ou encore l'utilisation massive de corpus journalistiques à partir des archives des grands quotidiens. La troisième et dernière étape a vu la constitution de corpus issus du Web, pour lesquels la plupart des considérations typologiques sur le contenu ont été balayées au profit de l'explosion quantitative.

Cette évolution de nature est celle qui a sans doute le plus d'impact sur la différenciation des pratiques entre le TAL et la linguistique : le besoin de données langagières supposées génériques pour le développement de méthodes de traitement automatique a entraîné des choix que nombre de linguistes (pourtant convaincus de la nécessité des corpus) refusent de suivre. On citera notamment Fuchs (2014, ce numéro) sur la sur-valorisation du quantitatif, Rastier (2005) à propos du Web, mais aussi dès la fin des années 1990 les mises en garde de Péry-Woodley (1995) ou Habert (2000) contre le slogan « gros, c'est beau ».

1.2 Suprématie des méthodes numériques

Sans chercher à déterminer quelle est la cause et quelle est la conséquence, on constate que l'accroissement des volumes de données s'est accompagné de changements radicaux dans la fa-

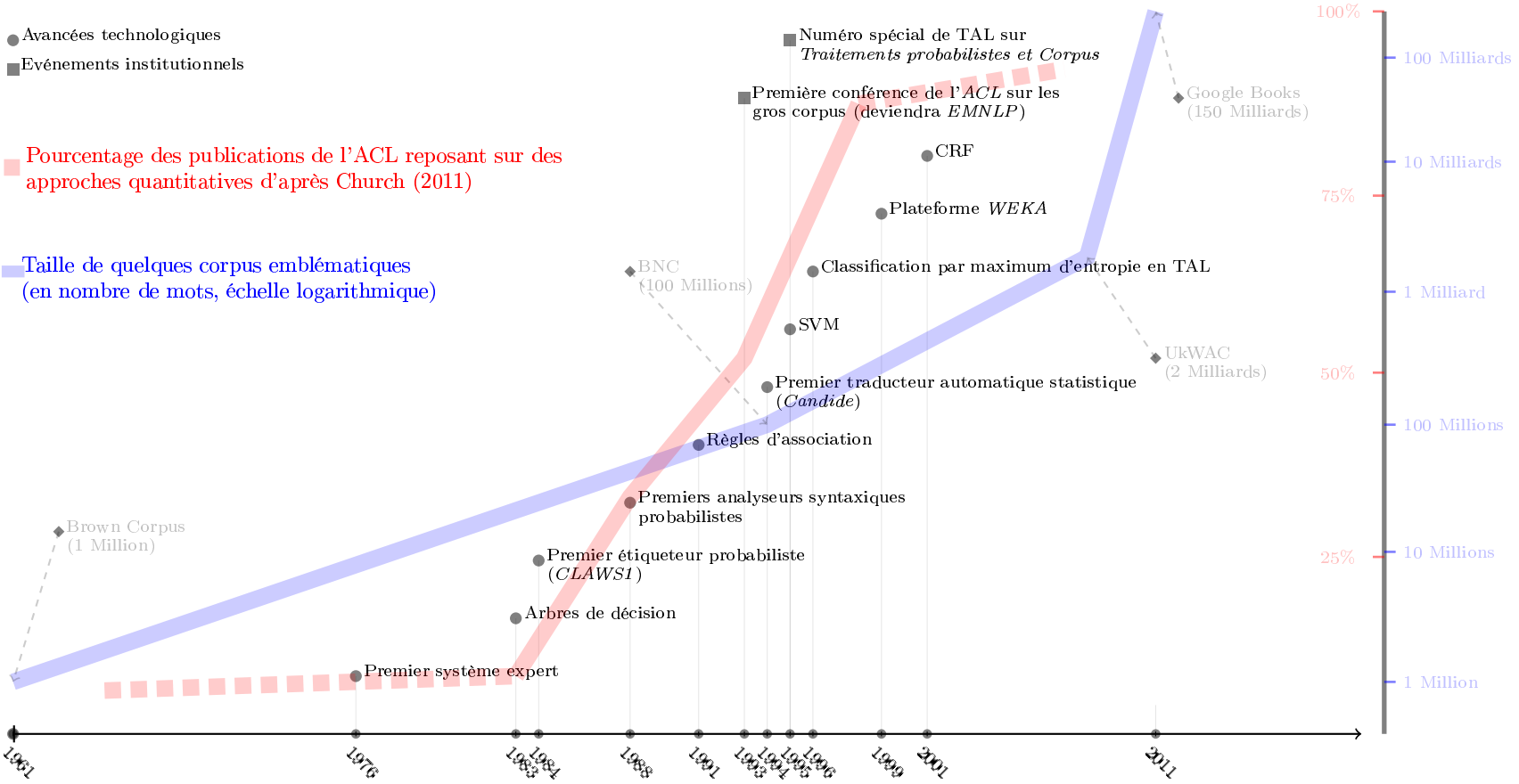

FIGURE 1 – Quelques repères chronologiques de l'évolution du TAL

çon d'aborder le matériau linguistique, à travers le développement des méthodes quantitatives et statistiques.

Résumons en quelques lignes le principe à l'œuvre. Pour réaliser un système de traitement automatique du langage, la méthode « traditionnelle » consiste à expliciter les opérations à effectuer sous la forme de règles à appliquer, et à développer des ressources (lexiques, grammaires, etc.) sur lesquelles ces règles s'appuient. Autrement dit, il s'agit de formaliser finement les étapes d'analyse et de traitement, de les tester sur des données et de corriger les règles (ou les ressources) afin d'obtenir un traitement convenable. Cette étape passe donc par une utilisation des connaissances disponibles sur les phénomènes langagiers, tout en nécessitant des compétences spécifiques permettant de trouver un compromis entre la théorie et les besoins d'un système efficace et robuste. Prenons un exemple simple, celui de l'étiquetage morphosyntaxique : si l'on souhaite annoter automatiquement les catégories grammaticales (par exemple les verbes) dans un texte, on établira une liste des formes verbales de la langue (la plus étendue possible), ainsi que des stratégies permettant de résoudre les cas ambigus (comme le mot *danse*). On peut par exemple établir comme règle simple que la présence d'un pronom personnel sujet à gauche (comme dans *il danse*) ou d'un adverbe à droite (*danse encore*) indique qu'il s'agit d'un verbe et pas d'un nom.

Les approches qui s'appuient sur des méthodes par apprentissage visent à contourner cette phase d'explicitation par un humain expert. À sa place, des méthodes statistiques sont utilisées pour faire émerger les régularités à partir des données, sans intervention d'un expert de la question. Cela ne veut pas pour autant dire que l'intervention humaine n'est pas requise. Le principe de l'apprentissage artificiel supervisé repose sur l'analyse de données annotées, fournies au système comme base d'exemples à partir desquels il va généraliser et construire (automatiquement) un modèle. Pour reprendre l'exemple précédent de l'étiquetage morphosyntaxique, cette opération consiste à fournir au système des textes dans lesquels toutes les formes verbales ont été marquées (à priori manuellement). L'étape suivante consiste à définir un ensemble de caractéristiques génériques des unités à traiter sur lesquelles le système va pouvoir se baser (par exemple, la catégorie des mots qui le précèdent), en se basant sur une connaissance très générale du phénomène visé. Le système va alors exploiter la répétition de configurations dans les données annotées, par exemple le fait qu'une séquence *Déterminant-Adjectif* précède bien plus souvent un nom qu'un verbe. Ce type de règles peuvent être très différentes de celles vues plus haut et produites par un linguiste, mais elles sont tout aussi applicables pour prendre des décisions sur de nouvelles données. De nombreuses méthodes, rappelées dans la figure 1, ont ainsi été développées, suivant une progression de complexité croissante depuis les méthodes symboliques (dont les modèles sont représentés par des ensembles de règles comme les arbres de décision) vers des méthodes purement numériques d'une opacité totale (entropie maximale, SVM, CRF).

En TAL, les différents champs d'applications ont vu s'imposer ces méthodes par apprentissage. Venues du traitement de la parole (Adda-Decker et Estève, 2014), elles ont transformé le monde de la traduction automatique (Mariani, 2014; Schwenk, 2014), pour couvrir l'essentiel des besoins, de l'annotation de bas niveau (étiquetage, analyse syntaxique) aux besoins plus spécifiques (classification de documents, extraction d'information). Le développement de ces méthodes passe nécessairement par la disponibilité de grandes quantités de données pour pouvoir réaliser un apprentissage efficace. C'est ce changement que reflète notamment la courbe rouge de la figure 1 qui indique le taux de publications en TAL dans les supports de l'ACL (*Association for Computational Linguistics*) faisant usage de méthodes statistiques (calcul effectué sur 15 ans par Church (2011) et extrapolé par nos soins), et qui transparaît également dans d'autres études portant sur le contenu des publications, visant par exemple l'émergence des thématiques majoritaires d'une base de documents (Hall et collab., 2008).

En linguistique, les méthodes sur corpus se sont elles aussi dotées d'outils d'investigation et de mesure qui font un fort usage de méthodes quantitatives et d'appareillage statistique. S'il est moins facile d'en mesurer l'impact dans les publications comme cela a été fait pour le TAL, il semble que l'utilisation de la quantification et de méthodes statistiques (mesures de liaison, tests statistiques) est devenue une pratique courante, sans doute influencée par les exigences de certaines communautés et revues. Un autre indice est le succès que remportent des ouvrages méthodologiques sur l'exploitation quantitative des données langagières, comme Baayen (2008) et Gries (2009).

1.3 Conséquence : un éloignement réciproque de la linguistique et du TAL

On présente souvent l'âge d'or de la collaboration entre linguistes et informaticiens du TAL comme étant celui où les premiers proposaient des connaissances formalisées aux seconds qui les intégraient dans des systèmes automatisés. Idéalement, la connaissance linguistique ainsi injectée permettait d'accroître les performances des applications, et en retour cette interaction fournissait aux linguistes une confrontation à large échelle de leurs propositions théoriques pour améliorer les modèles. Cette image d'Epinal a été sans doute justifiée dans le domaine de l'analyse syntaxique basée sur différents formalismes issus de la grammaire générative. Dans ce cas, le linguiste pouvait à juste titre se considérer comme positionné au centre du système, même si les tâches étaient parfois séparées. Mais les techniques par apprentissage ont largement changé la donne, et attribué aux linguistes un rôle secondaire, quand elles ne l'ont pas tout simplement évincé.

On se souvient longtemps après de la fameuse phrase de Jelinek [1] : « *Whenever I fire a linguist our system performance improves.* » [2]. Appliquée au domaine de la reconnaissance automatique de la parole, elle traduisait le rôle et le succès croissants des méthodes statistiques à la fin des années 1980. Si ce genre d'ostracisme ne doit pas être pris au pied de la lettre (Jelinek dit regretter de ne pas avoir réussi à utiliser efficacement les connaissances produites par les linguistes), il n'en est pas moins vrai que le rôle principal attribué au linguiste a été largement modifié. Dès lors, la place qui lui est laissée comme participant au développement de systèmes de TAL est limitée à l'annotation manuelle des données d'apprentissage lorsque la tâche s'y prête, ou lors de l'évaluation des sorties des systèmes (qui se base généralement sur la comparaison avec des résultats produits manuellement). Pire encore, il semble que même cette place se voie progressivement réduite, puisque les systèmes par apprentissage ont de plus en plus tendance à se passer de phases intermédiaires impliquant une annotation et à se contenter de résultats ne nécessitant pas une expertise en linguistique descriptive (comme la traduction, la retranscription ou la classification de documents) (Fabre, 2010, chapitre 8) ; de plus, les méthodes d'évaluation qui guident les avancées favorisent les mesures quantitatives au détriment de l'examen qualitatif (Mariani, 2014).

Les succès notables rencontrés par ces méthodes s'accompagnent ainsi d'un appauvrissement des représentations et des connaissances langagières exploitées : en traduction automatique comme dans d'autres tâches, les « modèles de langue » (Schwenk, 2014) construits à partir de données sont souvent limités aux seules configurations de formes de surface (n-grammes de mots). Dans d'autres tâches comme l'extraction d'information ou la classification de documents, on peut même noter un déficit dans l'utilisation des techniques pourtant performantes d'annotation syntaxique automatique au profit du traitement des seules chaînes brutes de caractères. Cette décision est argumentée par le surcoût des méthodes syntaxiques en temps de calcul et en ressources, lequel est encore multiplié par le besoin de traiter plusieurs langues. Dès lors, l'augmentation constante des volumes de données est présentée comme la solution pour établir une marge de progression, quitte à faire table rase des connaissances accumulées par l'étude minutieuse des mécanismes du langage, à défaut de pouvoir sélectionner celles qui pourraient être utiles, et à identifier le moyen de les intégrer dans des systèmes de plus en plus complexes (Tanguy, 2012, chapitre 8).

Malgré ce panorama qui met en évidence un profond fossé entre la linguistique et le TAL, plusieurs points positifs se dégagent. Le premier est que les exigences d'un travail systématique sur des données massives permettent le développement de méthodes expérimentales en linguistique, qui bénéficient doublement de ces avancées. Le second est la grande disponibilité des données, qui peut également bénéficier aux linguistes, et ce d'autant plus qu'ils disposent de moyens d'accès efficaces à la masse, comme nous allons le voir dans la seconde partie. De plus, il est important de relativiser cette évolution qui est essentiellement circonscrite à certains aspects applicatifs du TAL, limités aux grands besoins de traitement comme la recherche d'information, la reconnaissance de

1. Dans son allocution *Applying Information Theoretic Methods : Evaluation of Grammar Quality Workshop on Evaluation of NLP Systems* en 1988. Cette phrase est surtout provocatrice quand elle est tirée de son contexte, comme l'a commenté avec humour son auteur dans un exposé intitulé *Some of my best friends are linguists.*

2. *Chaque fois que je vire un linguiste, notre système s'améliore.*

la parole et la traduction automatique. Il existe fort heureusement de nombreux terrains plus spécifiques dans lesquels la large couverture d'un traitement robuste n'est pas le seul objectif, et qui mobilisent bien plus facilement des savoirs plus pointus sur le langage.

2 Bénéfices du changement d'outillage et d'échelle pour l'analyse linguistique

Dans la continuité des évolutions que nous venons de décrire, la linguistique dispose à son tour des moyens de faire évoluer ses procédures d'observation et d'expérimentation. Elle peut ainsi mettre pleinement en œuvre les principes qui ont motivé la constitution du champ de la linguistique de corpus, à savoir la possibilité d'examiner des données langagières attestées numérisées, permettant de quantifier les faits, d'accéder à des régularités invisibles à l'œil nu, et d'étudier le langage à travers la variété de ses usages. Le vœu de John Sinclair de pouvoir considérer le langage à la lumière de données en grand nombre se trouve exaucé :

> "Analysis of extended naturally occuring texts, spoken and written, and, in particular, computer processing of texts have revealed quite unsuspected patterns of language (...) [The] major novelty was the recording of completely new evidence about how language is used (...) The language looks different when you look at a lot of it at once."[3] Sinclair (1991)

De nombreux auteurs ont salué l'avènement de ces nouvelles conditions d'expérimentation en linguistique. Girault et Victorri (2009) considèrent que le changement d'échelle et la disponibilité d'outils de TAL favorisent l'émergence "d'un nouveau dispositif (...) qui mérite pleinement le nom d'observatoire de la langue". Abney (2011) affirme également que la collecte de données à grande échelle associée à la mise en œuvre de techniques expérimentales systématiques fonde une nouvelle linguistique – expérimentale et "data-intensive" – qui aurait toutes les caractéristiques d'une discipline scientifique : la linguistique a désormais la capacité de tester ses hypothèses en opérant un traitement rigoureux des données.

Nous présentons et illustrons dans ce qui suit les caractéristiques de cette nouvelle méthodologie linguistique, basée sur des procédures automatiques de traitement des corpus.

2.1 Nouvelle méthodologie expérimentale

Le rapport aux corpus a changé. Considéré initialement plutôt comme un réservoir d'exemples servant à illustrer des faits et vérifier des hypothèses, le corpus est aujourd'hui également manipulé comme une source de données d'où l'on cherche à faire émerger des régularités dans une approche plus exploratoire, inductive. Cette opposition entre une linguistique *corpus-based* et une linguistique *corpus-driven* a été popularisée en particulier par Tognini-Bonelli (2001). D'un côté, les travaux dits *corpus-based* utilisent les corpus pour obtenir des illustrations et des décomptes de catégories préétablies ; la modélisation des phénomènes langagiers est préalable à l'exploration des corpus, elle est seulement éventuellement amendée par la découverte de données nouvelles, complétée par des éléments de quantification, richie par la prise en compte de la variation selon les genres de textes. De l'autre, les travaux de type *corpus-driven* placent les corpus au cœur du processus d'élaboration des catégories ; la modélisation est le résultat d'un examen systématique et sans à priori des corpus, qui bouscule des distinctions préalables et façonne de nouveaux découpages.

Les méthodes numériques, quantitatives, basées sur l'examen des fréquences des formes linguistiques, facilitent précisément ce second type d'approche. Elles trouvent leur origine dans les travaux de lexicométrie (Lafon et Salem, 1983) et sont aujourd'hui popularisées par la disponibilité d'un outillage statistique adapté (Gries, 2009). Par exemple, les travaux menés en lexicologie

3. *L'analyse de grandes quantités de textes, écrits et oraux, et en particulier le traitement informatique de ces textes, a révélé des schémas langagiers inattendus (...) La principale avancée a été l'établissement de nouveaux faits concernant les usages du langage (...) Le langage apparaît sous un jour nouveau quand on en observe beaucoup à la fois.*

montrent comment une approche inductive du matériau linguistique amène à considérer de nouvelles unités de description. Des mesures statistiques mettent au jour des motifs lexicaux constitués de plusieurs mots qui cooccurrent régulièrement dans des corpus examinés. Ces observations ont un impact pour la description lexicographique : elles permettent d'identifier de nouvelles unités phraséologiques et d'enrichir les informations contextuelles (en identifiant les expressions et les constructions dans lesquelles les mots apparaissent de façon privilégiée). Toutes sortes d'objets linguistiques sont ainsi considérés : unités terminologiques (ex : *prélèvement bactériologique*), collocations (ex : *prendre une option*), patrons lexico-syntaxiques (ex : adjectif de couleur + *de* + nom d'émotion – *rouge de honte*), voire tout bloc de mots récurrent (ex : *ne serait-ce que pour*, *le problème c'est que*). Ces travaux font évoluer les catégories existantes, dans la mesure où les séquences de mots mises au jour ne correspondent pas toujours à une segmentation lexicale ou syntaxique classique.

Cette approche inductive est donc une des premières caractéristiques de l'analyse linguistique quantitative. Elle n'oblige cependant pas à faire table rase de la connaissance linguistique et à se priver de toute modélisation préalable, comme le déplore Fuchs (2014, ce numéro). Des informations linguistiques de diverses natures (étiquettes grammaticales, fonctions syntaxiques, traits sémantiques, marques textuelles, etc.) peuvent être insérées dans les textes, pour accéder à des généralisations de plus haut niveau. L'approche quantitative peut alors s'appuyer sur une phase préalable d'annotation, réalisée de façon automatique ou manuelle. De fait, la production de corpus annotés est devenue une composante majeure de la linguistique outillée. Ces annotations peuvent combiner le recours à des outils de TAL (lemmatiseurs, étiqueteurs, analyseurs syntaxiques) devenus désormais faciles d'accès et standardisés, et l'appel à des procédures plus ciblées visant à coder manuellement ou automatiquement des informations linguistiques particulières qui peuvent varier d'une étude à l'autre.

À titre d'exemple, le projet d'annotation discursive de corpus Annodis (Péry-Woodley et collab., 2011) a permis de concevoir un dispositif expérimental inédit pour l'étude de certaines structures de discours, afin d'observer la manière dont les textes s'organisent, et comment ces modes d'organisation diffèrent selon les genres de discours. Un des objectifs du projet était la production d'un corpus annoté discursivement, pour permettre l'étude de structures discursives variées, parmi lesquelles les structures énumératives qui constituent un mode privilégié d'organisation du discours et de hiérarchisation de l'information. Il s'agissait d'asseoir cette étude sur l'annotation préalable d'un corpus de plus de 500 000 mots, comportant des textes de genres différents (articles scientifiques, articles d'encyclopédie, articles journalistiques). Le dispositif expérimental mis au point pour l'annotation et l'étude des structures énumératives a consisté à articuler plusieurs phases :

- une phase de marquage automatique d'indices linguistiques correspondant à des traits de surface susceptibles d'accompagner les structures énumératives (adverbiaux circonstanciels, connecteurs, séquenceurs, mais aussi indices de mise en forme comme les titres ou les puces) ;
- une annotation manuelle des structures énumératives réalisée au moyen d'une interface d'annotation dédiée. L'annotation manuelle est assistée par différents modes de visualisation des textes (zooms), et par la mise en évidence des indices prémarqués, dont la densité peut guider l'annotateur vers des zones susceptibles de contenir des structures. Cette phase d'annotation a été validée par le calcul d'un accord inter-annotateurs qui a démontré la faisabilité de la tâche d'annotation (structures bien identifiées par les annotateurs) ;
- une analyse quantifiée des structures annotées (plus de 700 au total), permettant de décrire leur anatomie (longueur, nombre d'items d'énumération, présence ou pas d'une amorce, répartition dans les trois sous-corpus, etc.). Cette étude est alimentée par la présence des traits linguistiques prémarqués, qui permettent par exemple de distinguer les structures basées sur des indices de mise en forme (puces, titres), de celles qui sont signalées par des indices lexicaux (du type *Premièrement*, *d'abord*, *Au XIX^{e}siècle*, etc.).

Cet exemple illustrant l'étude à grande échelle d'un objet linguistique complexe met en évidence les trois ingrédients principaux d'une démarche expérimentale en linguistique facilitée par le recours à des méthodes de TAL : la constitution d'un corpus de taille conséquente [4], diversifié, et préparé

4. 500 000 mots, cela peut paraître peu, au regard des corpus gigantesques évoqués précédemment. Mais il est

(marqué, annoté) ; la définition d'un protocole d'annotation précis, s'appuyant sur l'utilisation d'une plate-forme d'annotation et contrôlé par l'évaluation de l'accord inter-annotateurs ; la mise en œuvre de procédures de quantification permettant de dresser une typologie multi-dimensionnelle de l'objet linguistique étudié, c'est-à-dire combinant des indices linguistiques de différents ordres. La combinaison d'une phase d'annotation à large échelle et d'une phase d'analyse quantitative trouve son modèle dans les travaux de Biber (1988), qui a étudié les genres textuels à partir de leurs spécificités linguistiques grâce à la projection de dizaines de traits linguistiques sur les textes, suivie d'analyses factorielles faisant émerger les principales dimensions de variation.

Un outillage avancé de l'analyse linguistique à l'aide de méthodes de TAL ouvre donc le champ des méthodes disponibles en linguistique et permet de mettre en place des dispositifs rigoureux d'analyse qui augmentent considérablement les capacités d'observation des linguistes en les dotant d'une boîte à outils très élaborée.

2.2 Apport des données massives

L'accroissement de la masse de données, dont nous avons décrit les manifestations dans la première section, est une réalité pour la linguistique comme pour le TAL, à des degrés bien sûr très différents. Les facilités offertes au linguiste pour travailler sur de grands corpus sont nombreuses. Les données sont devenues beaucoup plus simples à collecter. En particulier, la grande masse des données issues du Web devient accessible selon diverses modalités. De larges corpus dérivés du Web sont disponibles dans plusieurs langues. Ainsi, en français, le corpus frWac (Baroni et collab., 2009) comprend 1,6 milliard de mots issus du Web (domaine .fr), lemmatisés et étiquetés. Des outils facilitent leur exploitation, en fournissant par exemple la possibilité d'interroger en ligne ces données à l'aide d'un concordancier [5], d'extraire des listes de formes, des collocations, etc.

On peut objecter à cet attrait des données volumineuses que la taille du corpus n'a pas lieu d'être, pour le linguiste, le critère prépondérant dans la phase de constitution de ses données. L'intérêt de disposer d'un corpus bien défini pour permettre l'étude d'une variété langagière particulière doit prévaloir. Néanmoins, il devient possible d'allier ces deux critères que sont le volume et la spécificité, comme dans le cas du corpus Scientext [6] constitué d'écrits scientifiques échantillonnés par genre, langue et discipline, et qui atteint 33 millions de mots pour la partie anglaise, près de 5 millions pour la partie française. De même, si l'on a besoin de mieux contrôler la nature des données, il est également possible de constituer rapidement un corpus spécifique à partir du Web en utilisant une liste d'amorces (termes associés au domaine recherché) grâce à un outil comme Bootcat (Baroni et Bernardini, 2004). Des langages de recherche performants, conçus pour traiter de grands volumes de données annotées et faciliter des recherches plus précises dans les textes (en combinant formes, étiquettes et autres descripteurs), sont désormais intégrés à de nombreux programmes, par exemple le langage de requête CQP, dans le logiciel TXM [7].

Des données linguistiques en grand nombre sont désormais disponibles. Mais quel bénéfice peut-on véritablement tirer de cette masse de données ? Comme le dit Fuchs (2014, ce numéro), les linguistes se sont-ils lancés en vain dans cette course aux données ? Perdent-ils leur temps dans ce long travail de constitution et d'annotation de corpus ou s'agit-il véritablement d'un moyen de faire émerger des connaissances nouvelles sur la langue ? On peut citer deux types de recherche illustrant les perspectives offertes par la possibilité de mener des études à large échelle.

Le premier apport des grands corpus a été la possibilité de mener des études quantifiées sur la variation linguistique. Les travaux de Douglas Biber, déjà cités, ont montré l'importance de disposer de vastes ensembles de données pour étudier la variabilité dans toutes ses dimensions, et en particulier pour mettre au jour des corrélations significatives entre des traits linguistiques et des paramètres situationnels :

question cette fois d'un corpus annoté en grande partie manuellement, dont l'enrichissement a pris plus de 500 heures.

5. http://nl.ijs.si/noske/index-en.html
6. http://scientext.msh-alpes.fr/scientext-site/
7. http://textometrie.ens-lyon.fr/

> "[...] [A]dequate descriptions of variation and use must be based on empirical analyses of natural texts. Further, such analyses should be based on multiple texts collected from many speakers, so that conclusions are not influenced by a few speakers' idiosyncrasies. Finally, such analyses must simultaneously consider the influence of a range of contextual factors on linguistic variability." [8] (Reppen et collab., 2002)

Ces travaux débouchent en particulier sur de nouvelles descriptions de la grammaire des langues, où les structures linguistiques sont systématiquement décrites en relation avec leurs usages dans différents genres de discours.

Dans une perspective cette fois non variationniste, la disponibilité de très larges corpus de textes peut permettre de faire "sauter le goulot d'étranglement [...] de la collecte des formes" (Plénat et collab., 2002), lorsqu'il s'agit d'étudier des formes rares ou des configurations de traits complexes. Les travaux en morphologie ont particulièrement bénéficié de l'alliance entre vastes corpus et mesures quantitatives, donnant lieu à ce que Hathout et collab. (2009) nomment la morphologie extensive, définie comme "la collecte du plus grand nombre possible d'attestations du procédé étudié en vue de faire apparaître des régularités nouvelles". Ces travaux se sont particulièrement focalisés sur les phénomènes de créativité lexicale, sur les paramètres phonologiques conditionnant l'emploi des suffixes, et sur les configurations sémantiques qui leur sont attachées. Cette approche extensive a permis de compléter voire d'invalider les descriptions existantes.

L'utilisation de grands volumes de données a donc d'ores et déjà permis de mettre au jour des faits nouveaux, de dégager des généralisations jusque là inaccessibles. Elle suppose la maîtrise d'outils avancés de traitement du langage, ce qui constitue actuellement un obstacle important à l'extension de ces méthodes à la communauté linguistique. À titre d'exemple, les techniques automatiques d'analyse sémantique distributionnelle, inspirées de la démarche harrissienne d'analyse sémantique, calculent automatiquement des classes de mots sur la base des contextes qu'ils partagent. Ce principe donne lieu depuis plusieurs années à des implémentations nombreuses, qui fournissent une masse considérable d'informations sur le fonctionnement sémantique des mots dans différentes langues et différents corpus, sous la forme de graphes sémantiques reliant les mots qui partagent les mêmes contextes [9] (Fabre, 2010; Baroni et Lenci, 2010). Ce type de résultat reste encore sous-utilisé hors de la communauté du TAL, malgré son importance pour les études en sémantique lexicale.

La linguistique outillée a donc commencé à intégrer les méthodes quantifiées et automatisées d'analyse de corpus pour explorer ses propres objets de recherche. Nous allons voir dans la dernière partie que ces aptitudes peuvent être mises à profit pour développer un versant plus applicatif de la discipline qui ne soit pas nécessairement tributaire du TAL.

3 Une boîte à outils qui offre au linguiste de nouvelles perspectives d'intervention

Dans une période où de nouveaux terrains applicatifs existent, particulièrement en lien avec le traitement des documents et des échanges dans les milieux professionnels, l'outillage moderne de la linguistique peut lui permettre d'accroître substantiellement ses capacités d'intervention.

3.1 Des terrains et des problématiques qui se sont fortement renouvelés

On connaît l'importance actuelle des besoins des organisations en termes de traitement de l'information textuelle ; elle est liée au fait que cette information pléthorique, hétérogène, non struc-

8. *La description de la variation et de l'usage doit s'appuyer sur l'analyse de textes attestés. De plus, de telles analyses doivent se fonder sur des textes nombreux, produits par une grande variété de locuteurs, pour qu'elles ne soient pas influencées par des idiosyncrasies. Enfin, ces analyses doivent prendre en compte l'influence des facteurs contextuels sur la variabilité linguistique.*

9. On peut consulter ce type de données sur le site de l'équipe CLLE-ERSS : `http://redac.univ-tlse2.fr/applications/vdw.html`

turée, constitue un gisement de connaissances largement sous-exploité. Ces besoins sont multiples, parfois très ciblés. Parmi ces demandes émanant de la sphère professionnelle, on peut évoquer en particulier :

- l'amélioration de l'accès aux bases de documents. Des solutions passent par la construction de référentiels terminologiques, l'aide à la catégorisation des documents, le repérage d'informations importantes dans les textes. En particulier, des besoins importants concernent l'exploitation des données textuelles issues des procédures de retour d'expérience dans les entreprises (rapports d'incident, de pannes, d'accidents), voir par exemple (Pimm et collab., 2012) pour l'aéronautique ;
- l'analyse des interactions et des documents professionnels pour repérer des écarts à la norme, des sources d'ambiguïté et de malentendus, des évolutions terminologiques. Il s'agit d'identifier les sources de "risque langagier" (Condamines, 2008), qui se manifestent par des problèmes de compréhension des documents ou dans les échanges oraux ;
- l'amélioration de la lisibilité et de l'utilisabilité des documents, particulièrement des textes de nature procédurale, ou la conception de langages standardisés dans les situations à risque. On peut citer par exemple les travaux de Bouffier (2009) visant à faciliter la consultation de guides de bonnes pratiques dans le milieu médical. L'objectif est de mettre au jour la structuration des textes pour la rendre explicite, à partir d'indices à la fois typo-dispositionnels et linguistiques.

Ces besoins ne peuvent pas toujours être satisfaits par les solutions de TAL existantes, pour deux raisons principales.

Tout d'abord, les phénomènes impliqués sont spécifiques aux objectifs visés, et doivent être abordés par le biais d'observables à définir pour chaque cas. Ces observables (structures phrastiques, discursives ou typo-dispositionnelles par exemple) ne peuvent pas toujours se réduire aux unités classiques du TAL (mots ou séquences de mots) et ne sont définis que par une observation fine des données. Une fois identifiés, leur traitement et même leur observation au travers de méthodes quantitatives doit souvent faire appel à un outillage *ad hoc*.

Ensuite, les données à traiter sont propres à un domaine d'activité spécifique et présentent des caractéristiques (de taille, de forme, de vocabulaire) qui peuvent les rendre récalcitrantes aux approches de TAL usuelles. Typiquement, des textes professionnels présentant des spécificités d'expression requièrent une phase préalable d'examen, voire de normalisation. Leur contenu est bien souvent très différent de celui des données qui ont été utilisées pour établir ou étalonner des outils d'annotation automatique. Il est alors nécessaire d'adapter ou de contourner ces solutions génériques.

Les compétences des linguistes sont précieuses pour répondre à ces besoins. Malheureusement, ils sont rarement directement sollicités, du fait du manque de visibilité de la discipline dans ces milieux professionnels. La linguistique outillée est aujourd'hui mieux armée pour intervenir dans ce type de contexte.

3.2 Les atouts de la linguistique outillée

Une des caractéristiques importantes de la linguistique outillée est de pouvoir répondre rapidement à des demandes identifiées émanant des différents terrains évoqués. Si cela paraît évident au vu des différents besoins identifiés en termes de traitement de l'information, il est bon de rappeler que la mise en place d'une réponse rapide est bien souvent décisive, tant du point de vue de l'évolution des situations et de l'accroissement permanent des données, que pour asseoir la crédibilité de toute intervention concrète.

Le premier point critique est bien entendu l'accès aux données elles-mêmes, qui nécessite systématiquement de réaliser un ensemble d'opérations de sélection des éléments pertinents pour l'intervention, de convertir différents formats de données et, au besoin, d'ajouter des données empruntées à d'autres sources, que ce soit pour compléter ou pour fournir un support de comparaison. Dans tous ces cas, l'agilité nécessaire repose sur les techniques de gestion des textes électroniques disponibles à travers une panoplie d'outils dédiés.

Une fois les données identifiées ou réunies, les outils d'exploration de corpus permettent de dégrossir rapidement l'étude d'un phénomène particulier, dont on peut généralement élaborer une approche grossière sur la base de patrons lexicaux ou syntaxiques plus ou moins affinés. Les outils d'étiquetage et de recherche sur des données annotées qui font partie de l'arsenal du linguiste de corpus sont facilement déployables à moindres frais et pour un ensemble très varié de situations. C'est lors de cette phase de contact direct avec le matériau que le linguiste exerce véritablement ses compétences propres : identifier les phénomènes pertinents, repérer des configurations spécifiques, gérer les problèmes inhérents à la variation dans l'expression et bien entendu savoir formaliser ces différents aspects pour permettre leur gestion par des outils de repérage automatique.

Les outils génériques reposant sur des méthodes par apprentissage permettent d'accélérer de façon indéniable le processus de développement d'une première approche : qu'il s'agisse de mettre en place une première partition d'un corpus par apprentissage non supervisé (à la manière modernisée des outils d'analyse de contenu) ou d'identifier des caractéristiques saillantes d'une sous-partie du corpus, des plates-formes génériques d'exploration de données permettent d'obtenir rapidement un premier aperçu des régularités présentes dans les données.

Un autre aspect qui n'est pas à négliger est la nécessité pour la linguistique outillée de collaborer avec d'autres disciplines et/ou d'autres métiers, qu'il s'agisse d'autres acteurs du monde académique (psychologues, sociologues, historiens, etc.) ou des experts du domaine d'application (médecins, ingénieurs, etc.). Le dialogue interdisciplinaire peut s'établir à différents niveaux du processus, et peut largement dépasser la définition de catégories à mettre en œuvre et l'interprétation des résultats de l'analyse de corpus. Par exemple, la capacité pour la linguistique outillée à concevoir des descripteurs langagiers synthétiques permet d'incorporer ceux-ci dans des analyses multidimensionnelles. Cette démarche a notamment été menée dans le projet Intermede qui réunissait, autour d'un matériau constitué de consultations de médecine générale, des sociologues et des médecins, chaque expert étant capable de fournir des variables spécifiques aux différents éléments du corpus (Tanguy et collab., 2011). Le fait pour les linguistes du projet de pouvoir également produire sous forme de variables des mesures issues de leurs analyses (subjectivité, modalisation, répétition, etc.) a permis d'intégrer le tout dans une analyse multifactorielle qui prend une place centrale dans ce type de collaboration transdisciplinaire. Plus précisément, c'est le partage de compétences méthodologiques communes (en l'occurrence ici les analyses multivariées) qui a facilité le dialogue, lequel s'est prolongé sur d'autres aspects (vérification sur corpus des hypothèses formulées par les partenaires, extraction de passages spécifiques soumis à une analyse manuelle fine, etc.).

3.3 Vers une démarche d'intervention

Bien que la notion d'intervention linguistique outillée soit loin d'être aussi formalisée que celle d'intervention ergonomique par exemple, il nous semble possible d'en identifier les points principaux, et d'y associer les compétences techniques et méthodologiques correspondantes.

1. *Première observation des données* : circonscription des phénomènes ciblés et des spécificités du matériau. Cette phase, bien souvent ignorée par les approches classiques du TAL [10], est un préalable vital, durant lequel le linguiste peut rapidement cibler les difficultés et définir une marche à suivre. Le cas échéant, cette observation peut nécessiter l'obtention de données complémentaires par les moyens évoqués précédemment.
2. *Identification des besoins en termes de prétraitements et de ressources* (disponibles ou à développer). C'est ici que le linguiste fera l'inventaire des techniques pertinentes disponibles dans sa boîte à outils.
3. *Fouille linguistique des données.* En utilisant différents types de techniques classiques de la linguistique de corpus (recherche par patrons, mesures de fréquence, cooccurrences, etc.), les phénomènes linguistiques sont abordés de façon itérative, en affinant les descriptions.
4. *Caractérisation linguistique finalisée.* Une fois l'étape précédente aboutie, la description se stabilise sous la forme de marqueurs ou de mesures synthétiques calculées sur les données.

10. Dont on voit la trace dans la raréfaction des exemples dans les articles de TAL.

5. *Exploitation*. En fonction des besoins de l'intervention, elle peut prendre la forme d'un processus de traitement (classification), d'une finalisation de ressources (terminologiques par exemple) ou encore d'une confrontation des mesures dans une approche quantifiée avec les caractéristiques extra-linguistiques des données.

Comme on le voit, à chacune des étapes le savoir-faire s'appuie sur un contact direct et incontournable avec les données, et un outillage varié, souvent affiné et développé pour le besoin spécifique.

Conclusion

Nous avons voulu dans cet article prendre pleinement la mesure des évolutions récentes des techniques informatiques de traitement du langage, dans la grande variété qu'elles recouvrent. Les bouleversements observés depuis les deux dernières décennies ont clairement placé sur le devant de la scène les méthodes numériques sur des données massives, créant un fossé que nous déplorons entre les linguistes et les informaticiens, et établissant une distance entre le versant plus finalisé de l'ingénierie linguistique d'une part et la connaissance du matériau langagier d'autre part. Combler ce fossé ne pourra se faire qu'avec une prise de conscience des avantages que ces évolutions ont néanmoins sur les pratiques des linguistes, qui se voient dotés d'outils et de ressources qu'il ne tient qu'à eux d'exploiter au mieux. Nous avons voulu montrer comment les nouvelles approches méthodologiques permettent de (re)positionner la linguistique empirique sur des bases expérimentales, dont les principaux succès sont dus au changement d'échelle des données utilisables pour requestionner certains phénomènes langagiers. De plus, la maturité des techniques informatiques et statistiques permet au linguiste de jouer un rôle de premier plan dans des situations concrètes, en lui donnant une réelle capacité d'action face à des questionnements qui sont à la fois passionnants et bien souvent hors d'atteinte des seules approches massives mises en œuvre par le TAL au niveau de ses applications centrales (recherche d'information, traduction automatique, etc.). Mais le chemin à parcourir est encore semé d'embûches : outre la nécessité d'une meilleure visibilité, et de l'établissement de méthodologies plus abouties, se pose la question de la formation des nouveaux linguistes, en cherchant à concilier l'affinité avec les données, les connaissances linguistiques à large spectre et les exigences de savoir-faire techniques toujours plus complexes.

Références

Abney, S. 2011, «Data-intensive experimental linguistics», *Linguistic Issues in Language Technology*, vol. 6.

Adda-Decker, M. et Y. Estève. 2014, «Reconnaissance automatique de la parole», *L'Information Grammaticale*, vol. 142, n° 1.

Baayen, R. H. 2008, *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*, Cambridge University Press, Cambridge.

Baroni, M. et S. Bernardini. 2004, «BootCaT : Bootstrapping Corpora and Terms from the Web.», dans *Proceedings of LREC*, Lisbon, p. 1313–1316.

Baroni, M., S. Bernardini, A. Ferraresi et E. Zanchetta. 2009, «The WaCky wide web : a collection of very large linguistically processed web-crawled corpora», *Language resources and evaluation*, vol. 43, n° 3, p. 209–226.

Baroni, M. et A. Lenci. 2010, «Distributional memory : A general framework for corpus-based semantics», *Computational Linguistics*, vol. 36, n° 4, p. 673–721.

Biber, D. 1988, *Variation across Speech and Writing*, Cambridge University Press, Cambridge.

Bouffier, A. 2009, «A textual approach for the analysis of health practice guidelines», *TAL*, vol. 50, n° 1, p. 35–59.

Church, K. 2011, «A pendulum swung too far», *Linguistic Issues in Language Technology*, vol. 6.

Condamines, A. 2008, «Peut-on prévenir le risque langagier dans la communication écrite ?», *Langage et société*, vol. 3, p. 77–97.

Fabre, C. 2010, *Affinités syntaxiques et sémantiques entre les mots : apports mutuels de la linguistique et du traitement automatique des langues*, Mémoire d'habilitation à diriger des recherches, Université de Toulouse le Mirail.

Fuchs, C. 2014, «Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs», *Information Grammaticale*, vol. 142, n° 1.

Girault, S. et B. Victorri. 2009, «Linguistiques de corpus et mathématiques du continu», *Histoire Epistémologie Langage*, vol. 31, n° 1, p. 147–170.

Grau, B. et P. Bellot. 2014, «Recherche d'information et fouille de texte», *L'Information Grammaticale*, vol. 142, n° 1.

Gries, S. T. 2009, *Quantitative corpus linguistics with R : a practical introduction*, Routledge.

Habert, B. 2000, «Des corpus représentatifs : de quoi, pour quoi, comment», dans *Linguistique sur corpus. Etudes et réflexions*, vol. 31, édité par M. Bilger, Presses universitaires de Perpignan, p. 11–58.

Habert, B. 2004, «Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs», *Revue française de linguistique appliquée*, vol. IX, n° 1, p. 5–24.

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

Hathout, N., F. Namer, M. Plénat et L. Tanguy. 2009, *Aperçus de morphologie du français*, chap. La collecte et l'utilisation des données en morphologie, Presses Universitaires de Vincennes, p. 267–288.

Lafon, P. et A. Salem. 1983, «L'inventaire des segments répétés d'un texte», *Mots*, vol. 6, n° 1, p. 161–177.

Mariani, J. 2014, «Ressources et évaluation, histoire et utilisation», *L'Information Grammaticale*, vol. 142, n° 1.

Péry-Woodley, M.-P. 1995, «Quels corpus pour quels traitements automatiques ?», *Traitement Automatique des Langues*, vol. 36, n° 1-2, p. 213–232.

Péry-Woodley, M.-P., S. D. Afantenos, L.-M. Ho-Dac et N. Asher. 2011, «La ressource ANNODIS, un corpus enrichi d'annotations discursives», *Traitement Automatique des Langues*, vol. 52, n° 3, p. 71–101.

Pimm, C., C. Raynal, N. Tulechki, E. Hermann, G. Caudy et L. Tanguy. 2012, «Natural language processing tools for the analysis of incident and accident reports», dans *Proceedings of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero)*, Brussels.

Plénat, M., S. Lignon, N. Serna et L. Tanguy. 2002, «La conjecture de Pichon», *Corpus*, vol. 1, n° 1, p. 105–150.

Rastier, F. 2005, «Enjeux épistémologiques de la linguistique de corpus», dans *La Linguistique de corpus*, édité par G. Williams, Presses Universitaires de Rennes, p. 31–46.

Reppen, R., S. M. Fitzmaurice et D. Biber. 2002, *Using corpora to explore linguistic variation*, vol. 9, John Benjamins.

Schwenk, H. 2014, «Y a-t-il un sens pour la traduction automatique ?», *L'Information Grammaticale*, vol. 142, n° 1.

Sinclair, J. 1991, *Corpus, concordance, collocation*, Oxford University Press.

Tanguy, L. 2012, *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*, Mémoire d'habilitation à diriger des recherches, Université de Toulouse le Mirail.

Tanguy, L., C. Fabre, L.-M. Ho-Dac et J. Rebeyrolle. 2011, «Caractérisation des échanges entre patients et médecins : approche outillée d'un corpus de consultations médicales», *Corpus*, vol. 10, p. 137–154.

Tognini-Bonelli, E. 2001, *Corpus linguistics at work*, *Studies in Corpus Linguistics*, vol. 6, John Benjamins, Amsterdam.