Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs

Catherine Fuchs

▶ **To cite this version:**

Catherine Fuchs. Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs. L'information grammaticale, 2014, 142, pp.8-13. hal-01382602

HAL Id: hal-01382602
https://hal.science/hal-01382602

Submitted on 17 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

[Version pre-print d'un article paru en 2014 dans :
L'Information Grammaticale 142, pp. 8-13.]

Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs [*]

Catherine FUCHS
ENS, PSL Research University,
CNRS (LATTICE, UMR 8094), Paris 3

Ce que l'on est convenu d'appeler le « traitement automatique du langage » (TAL) constitue un vaste domaine hétérogène, allant de la conception de modèles théoriques jusqu'à la fabrication d'outils opérationnels, en passant par toute une série de travaux intermédiaires, dont les objectifs, les méthodes et les démarches sont très divers (Fuchs & Habert, 2004).

Ainsi que l'a montré le précédent numéro thématique de l'*Information grammaticale*, le TAL est actuellement engagé dans le traitement de grandes masses de données. Une évolution assez comparable s'est produite, au cours des vingt dernières années, dans le domaine de la linguistique. Là aussi, privilège est désormais accordé aux grandes masses de données langagières, ce qui conduit à un changement radical des perspectives et des méthodes de travail.

Comme on va le voir, cette évolution reflète celle, plus générale, des sciences cognitives ; mais, malgré une apparente analogie, les enjeux du tournant quantitatif sont différents dans le cas du TAL et dans celui de la linguistique. C'est un point de vue – inévitablement partiel et partial – de linguiste ayant coopéré avec des spécialistes de TAL et œuvré dans le domaine des sciences cognitives, que je présenterai ici, en restreignant mon propos au traitement de la langue écrite.

1. Une évolution inscrite dans l'histoire des sciences cognitives

Commençons par le rappel de quelques jalons historiques de nature à éclairer la situation actuelle (Fuchs, 2011).

1.1. L'enthousiasme des débuts

Nées à la fin des années 1940 dans le contexte politique de la « guerre froide », les premières recherches en TAL portaient sur la traduction automatique. Dans ce domaine – tout comme dans ceux de la compréhension automatique de textes, puis de la génération automatique de textes, qui devaient occuper le devant de la scène quelques décennies plus tard –, l'ampleur des ambitions proclamées était considérable, alors que les informaticiens disposaient évidemment de matériels infiniment moins puissants et performants que les ordinateurs actuels, et (surtout) qu'ils ne mesuraient pas encore la complexité de l'objet et la difficulté de la tâche.

Parallèlement, émergeait en 1956 au M.I.T. (dans le cadre d'une importante conférence organisée au Dartmouth College par Mc Carthy, consacrée à l'intelligence artificielle, - et à

[*] Cet article a bénéficié des remarques de trois relecteurs (Sophie Rosset, Thierry Poibeau et un(e) anonyme), que je remercie vivement ; leurs commentaires ont permis de préciser certains points.

laquelle participait Noam Chomsky) un courant linguistique dit de *« linguistique computationnelle »* (c'est-à-dire calculatoire), fondé sur un rapprochement entre langues naturelles et langages formels. L'objectif commun aux linguistes, mathématiciens et logiciens de ce courant était de décrire le fonctionnement des langues à la manière d'une machine, en termes de calculs correspondant au traitement d'informations diverses — essentiellement syntaxiques au départ. D'où la recherche des « structures mathématiques du langage » (Harris, 1968) et l'élaboration de différents types de « grammaires formelles » (grammaire chomskienne ou autres formalismes syntaxiques), puis, des années plus tard, un certain nombre de tentatives visant à prolonger cette démarche formelle au plan de la sémantique, comme par exemple celle de Montague (Chambreuil & Pariente, 1990).

Cette branche de la linguistique s'est beaucoup diversifiée au fil des années et poursuit à l'heure actuelle ses avancées sous l'étiquette large de « *linguistique formelle* ». Elle recourt à des formalismes logico-algébriques pour construire des représentations métalinguistiques censées correspondre aux règles maîtrisées par les locuteurs de la langue. Cette démarche s'inscrit dans le paradigme *cognitiviste* classique, dit « computo-représentationnel-symbolique » (Fuchs, 2004 : 6-9) : il s'agit de *calculer* sur des *symboles* pour construire des *représentations*. La linguistique formelle a fourni aux spécialistes de TAL un certain nombre de modèles permettant de représenter des connaissances syntaxiques et sémantiques.

Du côté du TAL, l'essor de *l'intelligence artificielle* (Sabah, 1988/89) a constitué un moment privilégié de réflexion théorique interdisciplinaire, autour de la notion de *connaissances* sur la langue (indispensables à intégrer dans les traitements) et *d'architecture fonctionnelle* de ces connaissances (concernant le mode de mise en œuvre effective de ces connaissances). Pour faire traduire, comprendre ou produire un texte par une machine, il était nécessaire de lui donner les moyens de faire intervenir au bon moment et à bon escient les divers types de connaissances nécessaires (morphologiques, syntaxiques, sémantiques, …) : fallait-il *hiérarchiser* ces connaissances et les mobiliser *séquentiellement* ou bien *en parallèle*, ou encore les faire *interagir* ? Cette problématique (centrale pour la cognition, tant artificielle qu'humaine), rejoignait — au moins partiellement — les questions posées par la linguistique et la psycholinguistique (Fayol ed., 2002 ; Le Ny, 2005) concernant, d'une part, les formalismes de représentation des connaissances, et d'autre part les niveaux de connaissances. (On notera au passage que cette question du traitement séquentiel ou parallèle tend actuellement à redevenir d'actualité à propos de l'élaboration de systèmes complexes mêlant la parole, le signal, la transcription, la vidéo, etc.).

En matière de *formalismes* de représentation, la remise en question de l'isomorphie postulée entre langues dites « naturelles » et langages formels logico-algébriques a permis d'élaborer des formalismes considérés comme plus adéquats au traitement du langage – qu'il s'agisse, par exemple, de logiques non classiques (Kayser, 1990), ou de formalismes d'inspiration topologique (Victorri, 1994). Concernant les *niveaux* de connaissances, le besoin de représenter le sens a conduit à intégrer, par-delà la syntaxe, des connaissances sémantiques et pragmatiques. D'où un intérêt croissant pour la dimension du contexte et pour l'idée d'une pluralité de niveaux de sens (Kayser, 1987 ; 1991) : la notion de « profondeur variable » en est une illustration particulièrement représentative.

La perspective adoptée (celle de la mise en œuvre de *connaissances* expertes sur la langue) supposait évidemment le recours à la notion de *règles* : tel était l'édifice conceptuel sur lequel reposaient alors les travaux en intelligence artificielle – d'où, notamment, la vogue des systèmes dits « experts ».

Progressivement, les recherches en TAL se sont démarquées, plus ou moins fortement selon les cas, du strict paradigme cognitiviste des débuts. Certains travaux ont alors pris appui sur des théories du langage élaborées hors du cadre de la linguistique computationnelle « orthodoxe », telles les grammaires « cognitives » de Lakoff ou de Langacker (Fuchs, 2004 : 10-12). Et un certain nombre de recherches se sont résolument inscrites dans une perspective cognitive dite « *constructiviste* ». Mettant en jeu les notions d' « émergence » et d'« auto-organisation », ces travaux ont adopté une approche *connexionniste non-symbolique* – certains allant jusqu'à récuser l'idée même de « représentation » (Varela, 1989, 1996[2] : ch. 4 et 5).

L'intelligence artificielle se fondait explicitement sur l'analogie entre l'esprit(-cerveau ?) et la machine – certains chercheurs filant même la métaphore jusqu'à vouloir fabriquer des programmes de traitement automatique dont, non seulement les résultats (les produits de sortie), mais également les processus de traitement, auraient vocation à reproduire ceux de l'humain. En d'autres termes, des outils qui pourraient, non seulement *émuler* mais plus fondamentalement *simuler*, le comportement langagier des êtres humains.

1.2. Le tournant quantitatif

Toutefois, en raison de la difficulté de la tâche, nombre de réalisations de cette époque n'ont guère dépassé le stade de petites maquettes. A l'enthousiasme des débuts, devait alors succéder la désillusion : peu à peu, les chercheurs en TAL ont pris la mesure des problèmes posés par la recherche de modèles globaux de traitement du langage et réduit leurs ambitions. Les recherches se sont donc progressivement tournées vers des applications moins ambitieuses mais plus réalistes : fabriquer des outils (cantonnés à des sous-domaines fermés) limités mais opérationnels, plutôt que vouloir construire des artefacts quasi indiscernables de l'humain. On est également passé de l'objectif d'un traitement entièrement « automatique » (censé remplacer l'humain) à celui d'un traitement « assisté par ordinateur » (réputé aider l'humain en le déchargeant de certaines tâches fastidieuses et longues à exécuter pour lui). Ce sera l'essor des « industries de la langue » ou « ingénierie des langues » (Pierrel ed., 2000), suivi à présent par celui des « techno-sciences de la langue ».

Désormais, les approches en TAL se proclament volontiers empiriques : le bricolage et l'éclectisme théorique ne semblent plus tabous. On ne cherche pas à explorer les textes en profondeur, à l'aide de règles, pour donner une représentation du sens ; on revendique une analyse « light », qui en écume *quantitativement* et *statistiquement* la surface. Les nouveaux besoins en matière d'accès rapide à l'information à partir de documents électroniques, ainsi que la puissance décuplée des machines, ont sans conteste favorisé cette évolution, symptomatique de ce tournant de la fin des années 1990. Il s'agit de pouvoir traiter rapidement de très grandes quantités de données langagières, qui se comptent en milliards de mots. C'est pourquoi le TAL actuel recourt très largement à des techniques d'*apprentissage automatique*, permettant de mettre en évidence certaines *régularités* présentes dans les données textuelles traitées. C'est ce que l'on appelle « l'acquisition de connaissances à partir de données » : les connaissances résultant du processus d'apprentissage par la machine à partir d'un échantillon représentatif des données sont utilisées par le système (qu'elles contribuent ainsi à enrichir et à améliorer) pour le traitement de grandes masses de données. Cette volonté de faire *émerger* statistiquement des configurations régulières n'est pas sans rappeler la démarche de la cybernétique (première manière) des années 1940 (Dupuy, 1994,

1999[2]), et rejoint à certains égards celle du néo-connexionnisme des années 1980 évoquée plus haut.

L'une des caractéristiques les plus frappantes de ce tournant quantitatif en TAL est la disparition du type de dialogue *interdisciplinaire* qui prévalait dans les années 1980, lorsque l'intelligence artificielle fédérait informaticiens, linguistes et psychologues autour d'un programme de recherche commun ambitieux. Pourrait-on voir, de nos jours, un débat d'égal à égal aussi vigoureux et passionné que celui qui opposa, plusieurs années durant, l'informaticien Kayser et le linguiste Kleiber, pour savoir s'il convenait d'appréhender le sens comme multiple et à profondeur variable (Kayser, 1987), ou bien comme unique et soumis à des mécanismes pragmatiques de dérivation (Kleiber, 1990) ? Aujourd'hui, la distribution des rôles a considérablement changé. Certains (rares) chercheurs dotés d'une réelle double compétence de linguistique et d'informatique peuvent assumer par eux-mêmes les deux faces du travail de TAL. Mais, dans la majorité des cas, les linguistes coopérant dans des équipes de TAL se trouvent de fait réduits au rôle de simples annotateurs, pourvoyeurs ou vérificateurs de connaissances expertes locales (Fabre & Tanguy, dans ce numéro).

Du côté de la linguistique, l'un des faits marquants de l'évolution des dernières décennies concerne le type de données étudiées : progressivement, les recherches linguistiques ont été conduites à dépasser le niveau (traditionnel) de la phrase au profit de celui du *texte*. Ce dépassement est observable dans deux secteurs différents de la discipline.

D'une part, au sein même de la linguistique formelle, de nouveaux formalismes ont été élaborés, afin de modéliser l'interprétation du *discours* en termes de sémantique « dynamique » : ainsi la « Discourse Representation Theory » (Kamp, 1981), conçue pour représenter les relations anaphoriques et temporelles entre les phrases, puis la « Segmented Discourse Representation Theory » (Lascarides & Asher, 2007), pour modéliser l'interface sémantique-pragmatique à l'aide de relations rhétoriques.

D'autre part, à côté de cette linguistique formelle, s'est développée une linguistique que l'on pourrait appeler « de l'usage » (ou des usages). Elle se fonde sur l'étude de *corpus* (Habert & *al*., 1997) et se démarque ainsi de la construction introspective des données de langue, telle que la pratiquait la grammaire générative. Le recours aux corpus n'est certes pas une nouveauté en linguistique : il était déjà attesté du temps du structuralisme. Mais ce qui est nouveau c'est l'ampleur du phénomène – et aussi les vertus quasi-miraculeuses dont la notion de corpus semble auréolée aux yeux de certains. Cette valorisation du *quantitatif* se retrouve d'ailleurs chez bien des évaluateurs et experts, qui considèrent qu'une étude linguistique ne saurait avoir de valeur si elle ne s'appuie sur l'observation de corpus (supposés) représentatifs et numériquement importants. A tel point qu'à la question « quelles sont tes hypothèses et dans quel cadre théorique travailles-tu ? » semble s'être substituée celle-ci : « quel est ton corpus et combien de millions/milliards de mots brasses-tu ? ».

Plusieurs facteurs internes à la linguistique ont sans doute contribué à ce tournant quantitatif. Citons, d'une part, le développement des travaux en pragmatique, en socio-linguistique et sur l'oral, qui ont mis au premier plan la notion d'usage et la collecte de *données attestées*. Et, d'autre part, l'évolution même des *théories* de linguistique cognitive, qui se sont de plus en plus tournées vers les formes. A la vogue des grammaires cognitives – elles-mêmes issues de la « sémantique générative », qui, à la fin des années 1960, prônait déjà la prise en compte des structures de surface pour le calcul du sens, en réaction à la « sémantique interprétative » chomskienne, qui ne s'appliquait qu'aux structures profondes –, a ainsi succédé celle des

« grammaires de construction » de Fillmore (Goldberg, 1995 ; 2006) (Fortis, dans ce numéro). Par ailleurs, le tournant quantitatif pris par l'informatique, le prestige institutionnel associé au TAL, et ce que l'on pourrait appeler « l'empirisme ambiant », ont certainement aussi influencé en ce sens la pratique de nombreux linguistes.

2. Des enjeux différents

L'évolution récente du TAL et celle (d'une partie) de la linguistique semble donc analogue : les deux témoignent d'un glissement vers le quantitatif et vers des traitements ciblés sur les formes de surface. Pourtant, les enjeux de ce tournant ne sont pas les mêmes dans les deux disciplines.

2.1. Du côté du TAL

En TAL, il s'agit de *traiter* automatiquement des *données* qui sont des expressions langagières (mots, séquences, textes, dialogues — écrits ou oraux), à des fins applicatives : par exemple pour les traduire ou pour en extraire des informations. Le critère essentiel est celui de la *performance* des systèmes de traitement : il faut trouver les moyens de construire le système le plus efficace, celui qui conduit aux meilleurs résultats, tout en étant le plus rapide et le plus fiable. Les évaluations concernant les performances des systèmes semblent montrer deux choses (Mariani, dans le précédent numéro). D'une part, *l'influence du quantitatif sur la performance* : plus les données traitées sont quantitativement importantes, plus les résultats s'améliorent. D'autre part, la suprématie des approches fondées sur l'*apprentissage* automatique avec des méthodes *statistiques* (opérant sur les données appréhendées en tant que formes de surface) aux dépens des approches fondées exclusivement sur des connaissances expertes et des règles. De ce double point de vue, le tournant quantitatif orienté vers les données peut apparaître, au premier abord, comme un tournant positif pour le TAL. Pourtant, certains problèmes de fond demeurent (comme le souligne l'article de Mariani déjà cité).

La plupart des *tests* évaluant la performance des systèmes sont eux-mêmes de nature *quantitative* : la mode du « benchmarking » qui envahit tous les secteurs de notre vie quotidienne (Bruno & Didier, 2013) s'inscrit en effet exclusivement dans le quantitatif [1].

En dehors de ces mesures quantitatives, la question de l'évaluation *qualitative* des systèmes ne semble pas avoir encore reçu de réponse satisfaisante, malgré certains efforts en ce sens. Que signifie d'arriver à des scores de réussite avoisinant les 90% (pour des cas triviaux, dans des domaines restreints !), sachant que ce sont souvent les quelques pourcents restants qui, qualitativement, font toute la différence ? En effet, ces quelques pourcents correspondent aux faits de langue échappant aux régularités statistiques superficielles : loin de recouvrir des

[1] Les critères traditionnels de mesure des performances des systèmes de TAL par rapport aux attentes sont ce qu'on appelle le « rappel » et la « précision ». Le *rappel* est défini par le nombre de réponses pertinentes fournies par le système au regard du nombre de réponses pertinentes attendues : plus le nombre de réponses pertinentes fournies se rapproche des attentes, plus le taux de rappel est élevé ; à l'inverse, on parle de *silence* pour les réponses pertinentes attendues qui ne figurent pas dans les résultats du système. La *précision* se définit comme le nombre de réponses pertinentes fournies par le système par rapport au nombre total des réponses qu'il fournit : toutes les réponses non pertinentes données sont considérées comme du *bruit* ; moins il y a de bruit, plus le taux de précision est élevé. Signalons, pour être tout à fait complet, qu'à côté de ces deux mesures de rappel et de précision, il existe également certaines mesures « d'erreur » ; et que, par ailleurs, certaines évaluations font appel à des juges (humains), voire même à des utilisateurs.

« exceptions » inexplicables ou des « irrégularités » préjudiciables, ils peuvent révéler certaines propriétés essentielles de la langue. Par exemple, en matière de traduction, la qualité ne dépend pas seulement de la conservation du sens (l'absence de contre-sens) : elle est aussi fonction de la plus ou moins grande idiomaticité et de la valeur stylistique de l'équivalent proposé – comme le savent bien les enseignants de langue chargés de corriger des exercices de thème ou de version.

Plus généralement, la question du *qualitatif* surgit dès lors qu'il s'agit de faire effectuer par un système automatique une tâche un peu *complexe* : on constate en effet que plus la tâche est complexe, moins les performances progressent. On aurait donc affaire à une sorte de « plafond de verre » invisible, mais bien réel. En somme, l'augmentation de la quantité des données d'observation sur lesquelles on entraîne un système d'apprentissage ne permet d'en améliorer significativement les performances que s'il s'agit de faire émerger des régularités prenant appui sur certains points stables et fiables des formes en présence. Mais cela cesse d'être le cas quand on a affaire à une instabilité globale des relations de cooccurrence entre formes de surface. Ainsi une forme potentiellement ambiguë sera-t-elle désambiguïsée dans un contexte univoque, mais ne le sera pas dans un environnement immédiat lui-même plurivoque. De même, l'augmentation de la taille des données constitue un facteur favorable à l'amélioration des performances s'il s'agit de traiter des données assez similaires à celles qui ont servi à l'apprentissage ; mais cela ne semble plus être le cas face à des données d'une autre nature, de plus « haut niveau » - voire même simplement en cas de changement de genre de document.

Les systèmes reposant sur l'apprentissage sont donc relativement robustes (à condition de ne pas changer de domaine ou de genre), mais connaissent des *limites* importantes : ils ne fonctionnent que dans des cas relativement simples et homogènes (lorsqu'il s'agit de tâches de *reconnaissance de formes*) et fournissent des résultats d'une qualité assez moyenne. En pratique, cela peut être tout à fait suffisant pour répondre à certains objectifs. Mais pour les cas plus difficiles (sans doute plus intéressants du point de vue théorique), on cherchera à construire des systèmes hybrides mêlant connaissances acquises par apprentissage statistique à partir de données d'observation et connaissances expertes préalables. En effet, dès que le traitement exige la prise en compte de phénomènes autres que des régularités de surface simples, la nécessité se fait sentir de recourir à des connaissances expertes de haut niveau (sémantiques et pragmatiques). Toutefois, cette réinjection de connaissances linguistiques préalables ne conduisant pas immédiatement à une amélioration notoire de la performance des systèmes, le scepticisme des informaticiens concernant l'apport de la linguistique reste assez largement répandu – d'autant que les ressources linguistiques mises à leur disposition sont, tantôt trop complexes (et risquant de conduire à de mauvaises solutions), soit trop partielles.

Au plan cognitif, la notion centrale qui est ici en jeu est celle d'*apprentissage*. En quoi les connaissances *acquises* par apprentissage statistique à partir des données sont-elles différentes des connaissances *expertes* ? La boîte noire que constitue le système reliant les données d'entrée et les résultats de sortie ne permet pas de répondre à cette question. L'apprentissage statistique par la machine fait, à certains égards, penser à l'apprentissage humain. Cette problématique intéresse les enseignants et apprenants d'une langue étrangère, à la recherche du dosage optimal entre règles de grammaire et « bain de langue ». Elle intéresse également les psychologues et les enseignants de maternelle, soucieux de comprendre les mécanismes de l'acquisition du langage (par le biais la langue maternelle). Comme on peut l'observer, l'enfant cherche à reproduire certaines données entendues, à partir desquelles il procède à des tentatives de généralisation fondées sur l'analogie – ce qui donne lieu à des séries d'essais-

erreurs et d'ajustements progressifs : « on apprend en observant », comme le remarque Schwenk (dans le précédent numéro).

Mais, pour séduisante qu'elle puisse sembler au premier abord, l'analogie entre apprentissage statistique automatique et apprentissage humain ne va pas de soi. Peut-on vraiment espérer simuler, à l'aide d'un système d'apprentissage automatique, l'acquisition du langage par l'enfant, sans tenir compte des interactions avec le monde extérieur ? Par ailleurs, comment rendre compte du fait que, chez l'humain, certaines constructions linguistiques peuvent être acquises plus rapidement que d'autres, sans pour autant être plus complexes, et qu'un même apprenant peut avoir des niveaux de compétence différents pour des phénomènes différents ? Comme le souligne Blache (2011 : 94) : « il n'est donc pas possible de rendre compte de ce type de phénomènes en termes de raffinement progressif d'un système homogène, la grammaire ».

La question des *rapports entre le traitement du langage par la machine et le traitement du langage par l'humain* est donc loin d'avoir reçu une réponse définitive. A côté des travaux consacrés à l'évaluation des performances respectives de la machine et de l'humain, d'autres se sont intéressés aux *stratégies* respectives mises en œuvre dans divers types de tâches, par exemple dans le raisonnement (Pitrat, 2011). Ces études ont montré que l'humain n'est pas toujours aussi performant qu'on voudrait le croire : il a une capacité (de travail et de mémoire) moindre que la machine, il est sujet à des erreurs, à des défauts d'attention et à des oublis. Mais, par ailleurs, il est capable de recourir à stratégies variées et hétérogènes, et sa faculté d'oubli lui permet d'éviter de s'enfermer dans la répétition et de s'ouvrir à la nouveauté (Kipman, 2013). On pourrait dire, schématiquement, que l'humain supplée son infériorité au plan quantitatif par une supériorité qualitative. C'est pourquoi, pour dépasser les limites inhérentes aux systèmes actuels de TAL, peut-être faudrait-il chercher ailleurs que dans la pure et simple augmentation de la quantité des données à traiter, et admettre qu'un traitement complexe ne se fait pas par simple addition de tâches simples et homogènes.

Pour aller dans ce sens, certains chercheurs s'orientent actuellement vers des modèles permettant de prendre en compte la *variabilité* et l'*hétérogénéité* des types de données et de leurs traitements possibles. Ils proposent des approches alternatives, qui s'appuient sur les acquis des sciences cognitives en matière de traitement de la variabilité (Lautrey *& al.*, 2002) : ainsi, par exemple, les modèles fondés sur des « systèmes de contraintes » (Blache, 2011). Plus largement, des questions qui avaient été au centre des relations entre intelligence artificielle et sciences cognitives se trouvent actuellement posées à nouveaux frais dans un contexte technologique totalement renouvelé : d'où une nouvelle forme d'intelligence dite « ambiante », qui développe des formes inédites de couplage entre l'humain et son environnement (Garbay & Kayser, 2014).

2.2. Du côté de la linguistique

Sur les différentes questions qui viennent d'être évoquées, on ne peut que regretter que les linguistes – à la différence de certains informaticiens – ne se soient guère penchés. C'est qu'en effet le tournant quantitatif orienté vers les données semble avoir eu, paradoxalement, pour conséquence de détourner de la réflexion théorique les linguistes engagés dans cette voie, en les entraînant prioritairement dans une course effrénée à la constitution de ressources.

Dans un premier temps, les linguistes ont vu, à juste titre, dans les nouveaux instruments informatiques, la promesse *d'outils* opérationnels, susceptibles de leur faciliter le travail sur

leurs propres données. Là où le grammairien d'autrefois passait plusieurs années de sa vie à faire des fiches à la main pour consigner certains observables, un linguiste « outillé » est devenu capable d'entrer, de trier et de gérer dans un système de gestion de bases de données une quantité importante d'observables en un temps infiniment plus court. Dans une telle perspective, le simple utilisateur qu'est alors le linguiste peut se livrer sur ses données de langue à une pratique empirique assistée par ordinateur. Et il est de fait que l'on dispose déjà, à l'heure actuelle, d'une palette appréciable d'outils et d'instruments électroniques de nature à faciliter le travail du linguiste (Habert, 2005). Une symbiose est ainsi rendue possible entre l'artefact et l'humain, basée sur leur complémentarité : capacité de stockage, vitesse de travail, fiabilité du traitement (du côté de la machine) ; souplesse, adaptabilité, diversité des parcours (du côté de l'humain).

Mais, après ce stade de la « linguistique outillée », un nouveau pas a été franchi. Désormais, il ne s'agit plus simplement de recourir à des outils de bureautique avancée pour gérer certains observables de langue préalablement construits, mais de constituer, d'annoter et de valider des *ressources langagières*. L'intérêt des linguistes peut, on le sait, se porter vers des corpus de nature très diverse (corpus de linguistique historique, corpus spécialisés pour l'étude de terminologies techniques, corpus de langue parlée pour l'étude des interactions, corpus de consultations médicales, pour n'en prendre que quelques exemples). Dans cette course aux grandes masses de données langagières, le travail de *constitution* même des données (recueil de corpus, suivi d'une longue et fastidieuse phase d'annotation et de validation) a relégué au second plan – voire purement et simplement occulté – la question du traitement qualitatif de ces données. Ainsi semblent se rejoindre les préoccupations des linguistes désireux de travailler sur de *très gros corpus* et de les traiter automatiquement, et les préoccupations des informaticiens à la recherche de très grandes masses de données langagières pour alimenter et tester leurs systèmes d'apprentissage.

Le tournant quantitatif a donc bouleversé le cœur même de la pratique des linguistes. Mais il a aussi contribué à brouiller les rapports entre le TAL et la linguistique. Car l'objectif du linguiste n'est pas, en droit, le même que celui du TAL. Pour le linguiste, il s'agit d'améliorer, non pas la performance d'un système technologique, mais nos connaissances sur les langues. Même si la poursuite de cet objectif passe par l'observation de masses importantes de données attestées (préalablement soumises à des procédures d'étiquetage et d'annotations variées), il n'en demeure pas moins que ces ressources constituent un moyen, et non une fin en soi. Au plan des moyens, l'apprentissage automatique peut effectivement rendre des services aux linguistes, en opérant sur des sous-ensembles représentatifs (étiquetés à la main), qui servent ensuite de base à l'apprentissage : des masses considérables de données textuelles pourront ainsi être étiquetées automatiquement – ce qui réduira considérablement le travail manuel préparatoire.

Mais ce n'est pas le traitement des données par la machine (sans modélisation préalable) qui, sur le fond, fera surgir des connaissances inédites concernant le système de la langue : tout au plus permettra-t-il de mettre en évidence certaines régularités statistiques ayant trait à un certain type de pratique langagière dans des circonstances particulières.

Là précisément peut résider le piège, pour le linguiste tourné quasi-exclusivement vers les ressources. Le travail sur corpus n'est évidemment pas incompatible avec la réflexion théorique, il peut même être bienvenu pour étayer un raisonnement de linguistique. Mais à la condition de respecter les contraintes d'une démarche scientifique et de savoir construire une véritable problématique théorique. Or, dans la phase actuelle, le linguiste semble bien souvent

transformé en un travailleur de force qui n'en aurait plus ni le temps ni les moyens. A ce compte, la technicisation risque fort de renvoyer aux oubliettes le trésor de descriptions et de théories (non « outillées ») accumulées depuis des siècles (Lazard, 2013), en donnant l'illusion qu'un traitement de surface accompagné de quelques décomptes serait susceptible de révéler *proprio motu* les propriétés de la langue. Or les données langagières (même enrichies d'annotations diverses) ne se confondent pas avec le système de règles – aussi variable et labile soit-il – constitutif de la langue.

Point n'est besoin de rappeler que l'on ne saurait faire œuvre scientifique en présentant des résultats statistiques sur des « objets théoriques non identifiés », faute d'avoir réfléchi au statut des observables et émis de véritables hypothèses. Sachons donc nous préserver d'un possible effet pervers du tournant quantitatif en linguistique, qui risquerait de faire « prendre le Pirée pour un homme » !

Références

BLACHE, Philippe (2011). Vers une approche cognitive du traitement automatique des langues. In C. Garbay & D. Kayser (eds.). Pp. 73-98.

BRUNO, Isabelle & Emmanuel DIDIER (2013). *Benchmarking. L'Etat sous pression statistique*. Paris : Zones/La Découverte.

CHAMBREUIL, Michel & Jean-Claude PARIENTE (1990). *Langue naturelle et logique. La sémantique intensionnelle de Richard Montague*. Bern, Frankfurt/M., New York, Paris, : Peter Lang.

DUPUY, Jean-Pierre (1994, 1999[2]). *Aux origines des sciences cognitives*. Paris : La Découverte.

FAYOL, Michel (ed.) (2002). *Production du langage*. Paris : Hermès.

FUCHS, Catherine (2004). Pour introduire à la linguistique cognitive. In C. Fuchs (ed.) : *La linguistique cognitive*. Paris : Ophrys/MSH (Coll. Cogniprisme). Pp. 1-24.

FUCHS, Catherine (2011). Pertinence de l'informatique pour les sciences cognitives vue par une linguiste : informatique et langage. In C. Garbay & D. Kayser (eds.). Pp. 287-294.

FUCHS, Catherine & Benoît HABERT (2004). Le traitement automatique des langues : des modèles aux ressources. *Le Français Moderne*, *LXXII : 1*. Paris : CILF. Pp. 1-13.

GARBAY, Catherine & Daniel KAYSER (eds.) (2011). *Informatique et sciences cognitives : influences ou confluences ?*. Paris : éditions de la MSH (Coll. Cogniprisme).

GARBAY Catherine & Daniel KAYSER (2014). Addendum à 'Informatique et sciences cognitives'. In C. Fuchs (ed.) : *L'homme, l'animal et la société au prisme de la cognition*. Paris : éditions de la MSH (Hors-Série, Coll. Cogniprisme).

GOLDGERG, Adele (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago : University of Chicago Press.

GOLDGERG, Adele (2006). *Constructions at Work: the nature of generalization in language*. Oxford : Oxford University Press.

HABERT, Benoît (2005). *Instruments et ressources électroniques pour le français*. Paris : Ophrys (Coll. L'Essentiel Français).

HABERT, Benoît, Adeline NAZARENKO & André SALEM (1997). *Les linguistiques de corpus*. Paris : Colin/Masson.

HARRIS, Zellig (1968). *Mathematical Structures of Language*, New-York : Wiley (trad. fr. 1971 *Structures mathématiques du langage*. Paris : Dunod).

KAMP, Hans (1981). A theory of truth and semantic representation. In: J.A.G. Groenendijk, T.M.V. Janssen, and M.B.J. Stokhof (eds.), *Formal Methods in the Study of Language.* Mathematical Centre Tracts 135, Amsterdam. Pp. 277-322.

KAYSER, Daniel (1987). Une sémantique qui n'a pas de sens. *Langages*, 87. Pp. 33-45.

KAYSER, Daniel (1990). Adéquation et inadéquation de la logique au traitement sémantique des langues. *Modèles Linguistiques*, XII : 1. Pp. 119-136.

KAYSER, Daniel (1991). Meaning representation versus knowledge representation. In N. Cooper & P. Engel (eds.) : *New inquiries into meaning and truth*. New-York : St Martins Press. Pp. 163-186.

KIPMAN, Simon-Daniel (2013). *L'oubli et ses vertus*. Paris : Albin Michel.

KLEIBER, Georges (1990). Sur la définition d'un mot : les sens uniques conduisent-ils à des impasses ?. In J. Chaurand & F. Mazière (eds.) : *La Définition.* Paris : Larousse. Pp. 125-148.

LASCARIDES, Alex & Nicholas ASHER (2007). Segmented Discourse Representation Theory : Dynamic Semantics with Discourse Structure. In H. Bunt & R. Muskens (eds.) *Computing Meaning: Volume 3*. Springer. Pp. 87-124.

LAUTREY, Jacques, Bernard MAZOYER & Paul VAN GEERT (eds.) (2002). *Invariants et variabilité dans les sciences cognitives*. Paris : éditions de la MSH.

LAZARD, Gilbert (2013). Réflexions séculaires. *La Linguistique,* 49. Pp. 49-65.

LE NY, Jean-François (2005). *Comment l'esprit produit du sens*. Paris : O. Jacob.

PIERREL, Jean-Marie (ed.) (2000). *Ingéniérie des langues*. Paris : Hermès.

PITRAT, Jacques 2011). Raisonnement de l'homme et raisonnement de la machine. In C. Garbay & D. Kayser (eds.). Pp. 11-45.

SABAH, Gérard (1988/1989). *L'intelligence artificielle et le langage*, 2 vol. Paris : Hermès.

VARELA, Francisco (1989, 1996[2]). *Invitation aux sciences cognitives*. Paris : Le Seuil.

VICTORRI, Bernard (1994). La construction dynamique du sens. In M. Porte (ed.) *Passions des formes – à René Thom*. Paris : éditions de l'ENS Fontenay-St Cloud. Pp. 733-747.