Jacqueline Léon
CNRS, UMR7597 HTL, Université Paris Diderot, Université Paris Sorbonne Nouvelle, Sorbonne Paris Cité

Isabelle Tellier
UMR8094 LaTTiCe, Université Paris Sorbonne Nouvelle, Sorbonne Paris Cité

Le *data turn*. Des premiers traitements statistiques du langage (1950-60) à la fouille de textes

Introduction

On tient généralement les années 1990 pour le moment où le TAL a commencé à abandonner les méthodes symboliques au profit des méthodes statistiques ; ce mouvement apparaît lorsque le traitement de grandes masses de données devient possible grâce aux développements technologiques inédits des ordinateurs et des logiciels. Or si l'on examine les débuts du TAL, on s'aperçoit que, dès les années 1950, sont mis en œuvre des traitements du langage fondés sur des méthodes statistiques et probabilistes dans le sillage de la cryptographie et de la théorie de l'information. C'est particulièrement le cas en France où, dans les années 1950-60, sont utilisées et discutées les chaînes de Markov et la loi d'Estoup-Zipf, de sorte que l'on peut dire qu'en France les analyses quantitatives précèdent les premières expériences en traduction automatique et l'utilisation des méthodes logico-mathématiques dans le processus d'automatisation-mathématisation du langage[1]. Dans les années 1960-70, on voit se développer la lexicométrie (travaux de Moreau, Tournier, Lafon, Salem) et l'analyse des données (travaux de Benzécri dans les années 1970-80). Ces traitements sont contemporains des premiers analyseurs syntaxiques automatiques fondés sur les grammaires formelles, mais ils ont été partiellement éclipsés par l'analyse syntaxique une fois celle-ci devenue dominante dans les années 1960. Enfin, si l'on considère les unités linguistiques envisagées par ces traitements, on observera une partition assez nette : alors que la phrase est l'unité de l'analyse syntaxique automatique, le texte et le mot sont les unités privilégiées par les méthodes statistiques et probabilistes.

---

[1] Sur cette spécificité de la France voir Léon (2010) et Léon (sous presse).

Dans cet article, nous examinerons les deux moments « statistiques » du TAL, la période des débuts, dans les années 1950-60, et la période initiée dans les années 1990 où les méthodes statistiques et probabilistes se sont généralisées pour déboucher progressivement sur ce qu'on peut appeler le « *data turn* » (ou « tournant des masses de données »). Nous aborderons ces périodes de façon comparative en nous interrogeant sur les rapports entre TAL, méthodes statistiques et probabilistes, et linguistique. Nous nous poserons les questions suivantes :
- en quoi la généralisation du traitement des grandes masses de données remet-elle en cause le statut linguistique des unités « texte » et « mot »?
- en quoi la réorganisation des objectifs du TAL à partir de la notion de tâche déplace-t-elle ou oblige-telle à repenser le rapport du TAL avec la linguistique ?
- Quel est le statut de la linguistique elle-même lors de l'utilisation de ces méthodes : fournit-elle un cadre théorique pour le TAL, ou bien est-elle instrumentalisée et ses unités et ses méthodes utilisées comme simples ressources ? Constitue-t-elle encore un enjeu ou bien se retrouve-t-elle purement et simplement mise à l'écart ?

1. Unités linguistiques (mot et texte) et débuts du traitement automatique des langues

Quand on examine les débuts du traitement automatique des langues dans les années 1950-1960, le statut linguistique des unités est unanimement discuté par l'ensemble des acteurs du domaine, linguistes, ingénieurs, mathématiciens, même si les méthodes utilisées sont parfois inédites ou non conformes aux descriptions de la linguistique de l'époque. Les unités linguistiques sur lesquelles ont porté la plupart des traitements statistiques et probabilistes sont le mot et le texte. Ces deux unités sont souvent indissociables dans le traitement automatique et nous examinerons les définitions mises en œuvre par les différents modèles et théories au moment des débuts de l'automatisation du langage : chaînes de Markov, théorie de l'information, cryptographie, modèles probabilistes et modèles statistiques de distribution des fréquences. Deux points sont à noter cependant, et nous y reviendrons plus loin : le mot n'a pas été l'unité de traitement uniquement de méthodes statistiques, il a été aussi au cœur de certains travaux pionniers en traduction automatique. Inversement, d'autres unités que le mot ont fait l'objet de traitements statistiques. C'est le cas des phonèmes et des lettres (unités de longueur des mots).
Enfin il faut souligner que l'automatisation des traitements statistiques s'est faite progressivement. Le passage des comptages « à la main » aux machines mécanographiques puis aux ordinateurs n'a pas entraîné de grands changement ni dans les méthodes

statistiques ni dans la conception du rapport entre statistiques et linguistique. C'est le *data turn* amorcé dans les années 1990 qui a provoqué un véritable bouleversement.

1.1. Phrase, mot, texte

Il faut tout d'abord préciser que la définition des unités « phrase », « mot » et « texte », voire leur statut même d'unité linguistique, est loin de faire consensus pour tous les linguistes, et dépend largement de l'approche théorique adoptée. Néanmoins, ces entités ont un caractère empirique indéniable dès lors qu'il s'agit d'accomplir une tâche, qui, on le sait, constitue un des principaux objectifs du TAL.

La phrase est l'unité par excellence de la syntaxe. Comme il ne convient pas de faire ici l'histoire de l'analyse syntaxique automatique peu concernée, surtout à ses débuts, par les méthodes statistiques, disons simplement que le premier qui ait envisagé une analyse syntaxique automatique est Yehoshua Bar-Hillel (1915-1975). Philosophe du langage, ayant fait une thèse sur la syntaxe logique de Carnap, Bar-Hillel introduit la récursivité en linguistique et élabore une « syntaxe opérationnelle » pour la traduction automatique, fondée sur une grammaire catégorielle « A Quasi-arithmetical notation for syntactic description », associant la méthode de Harris et la notation d'Ajdukiewicz [Bar-Hillel 1953]. Pour Bar-Hillel, la syntaxe constitue la question principale à résoudre pour la traduction automatique, c'est pourquoi il conçoit sa syntaxe opérationnelle comme une machine capable de découvrir de façon automatique la structure syntaxique d'une chaîne d'une langue source donnée. Cette syntaxe opérationnelle peut être considérée comme le premier analyseur syntaxique automatique au fondement même de ce qui deviendra la linguistique computationnelle dans les années 1960 [Cori et al. 2002].

En linguistique, le mot est une notion complexe et hétérogène, d'ailleurs souvent controversée, dont les différentes dimensions, graphique, phonétique, syntaxique ou sémantique coïncident rarement et n'ont pas de propriétés constantes[2]. Pour le traitement automatique des langues, que ce soit en linguistique computationnelle ou en linguistique de corpus, une définition semble faire toutefois consensus : le mot est une suite de caractères délimités par des séparateurs[3]. On verra que cette définition, en vigueur dès les débuts du traitement automatique, est historiquement située et a fait l'objet de vifs débats.

---

2 Pour le mot, voir Tamba et Luzzati dans ce numéro, Léon 2001, Baratin et al. 2004.

3 Sur le plan empirique et pour le TAL, on peut distinguer deux types de mots : (i) les mots-formes, *tokens*, formes fléchies ou *running words*. Ce sont des « unités perceptibles de texte écrit qui peuvent être reconnues selon les espaces ou d'autres marques de séparation » [voir Antié et al. 2006] ; (ii) les

Comme le remarque Kyheng [2005], bien que le texte constitue l'un des objets les plus anciens des sciences du langage, il n'est envisagé comme concept par les linguistes qu'à partir de la seconde moitié du XX^{e} siècle. Aux trois écoles qui, selon elle, ont contribué à l'établissement du texte comme objet pour la linguistique, à savoir l'école sémiotique de Tartu, l'école sémiotique de Paris, et la *Textlinguistik* allemande, il faut ajouter les travaux des empiristes britanniques, en particulier de J.R. Firth dans les années 1950, pour lequel le texte intégral et authentique est une unité essentielle pour l'élaboration d'une théorie sémantique. C'est cette conception du texte qui sera à la base de la *Corpus Linguistics* telle qu'envisagée par John Sinclair au début des années 1960 (voir [Stubbs 1993], [Léon 2007]) – voir plus loin §1.4.3.

1.2. Chaînes de caractères et probabilités

1.2.1. Le modèle de Markov

Les premiers travaux probabilistes sur les textes concernent les successions de voyelles et consonnes et non directement les mots. Il s'agit des travaux du mathématicien russe Andrei Andrejevitch Markov (1856 - 1922), dont l'objectif est la recherche de constantes de probabilités liées en étudiant la succession des voyelles et des consonnes (vv, cc, vvv, vcv, cvv, ccv) –plus tard appelées digrammes et trigrammes - dans un chapitre et demi *d'Eugene Onéguine*, roman en vers d'Alexandre Pouchkine [Markov 1913]. Son modèle généralisé sous le terme de chaîne de Markov est un automate à états finis dont les transitions d'un état à un autre sont réglées par des probabilités. En 1948, dans le cadre de la théorie de l'information, Shannon a proposé un modèle probabiliste des séquences de lettres et de mots en anglais, fondé sur les chaînes de Markov. Par exemple, dans le cas où les n-grammes sont des lettres (ce peut être des mots), si un texte comporte 100 occurrences de "th", dont 60 occurrences de "the", 25 occurrences de "thi", 10 occurrences de "tha", et 5 occurrences de "tho", le modèle de Markov prédit que la prochaine lettre qui suit le 2-gramme « th » est « e » avec une probabilité de 3/5, elle est « i » avec une probabilité de ¼, « a » avec une probabilité de 1/10, et « o » avec une probabilité de 1/20.

Dans ce type de recherche, les unités sont des caractères graphiques, les consonnes et les voyelles, considérées selon leur probabilité d'apparition dans un extrait de texte. Les objectifs sont ceux d'un mathématicien : il s'agit pour Markov de trouver des constantes de probabilités. Nulle préoccupation stylistique ou littéraire ne l'anime. Comme l'a fait

---

formes vides (*types*, lemmes) qui ont déjà fait l'objet d'une abstraction. Cette distinction comporte des enjeux théoriques importants (voir ci-dessous §2.2.2).

remarquer Mandelbrot (1961, p.191), le texte selon Markov est appréhendé en tant que séquence de lettres résultant de tirages au hasard, suivant en cela la « règle d'urne »[4]. Il ne comporte aucune structure grammaticale, seule est conservée la dimension séquentielle.

1.2.2. Le mot « groupe codique »

Les n-grammes et les chaînes de Markov ont tout d'abord été utilisés en cryptographie. Ainsi, Moreau[5] [1961] montre comment en cryptographie et en télécommunications, lettres, syllabes et mots sont des unités pouvant travailler pour le même objectif, à savoir coder et décoder un message. En termes de la théorie de l'information, Moreau définit le mot ou « groupe codique » comme l'unité la moins coûteuse en termes d'entropie. En cryptographie comme dans le traitement automatique, quand le mot identifié dans le texte ne se trouve pas dans la mémoire de la machine on descend au niveau de la syllabe (ou plutôt des « psyllabes », pseudo-syllabes définies selon des critères mi-phonétiques mi-graphiques pour la machine). En dernière instance, on descend au niveau des lettres, mais le traitement devient plus coûteux. C'est bien en tant qu'unité empirique linguistique stockable dans un dictionnaire que le mot est utile en cryptographie comme il le sera en traitement automatique des langues. Dans ce cadre toutefois, il est très important que le mot soit considéré en dehors de tout recours au sens. Mandelbrot [1957] insiste sur le fait qu'utiliser des méthodes statistiques en linguistique suppose de travailler sur des lois purement formelles (entendues ici comme distinctes de la logique) à savoir des relations entre formes linguistiques considérées indépendamment de leur sens. Pour cela, il suffit, dit-il, d'utiliser les outils mis au point pour les télégraphistes, notamment par Shannon, en traitant de l'information, à savoir des signaux vides de sens.

1.2.3. Chaînes de Markov comme méthode de détermination des unités linguistiques

Les chaînes de Markov ont été utilisées par les linguistes distributionnalistes américains, pour segmenter la chaîne de phonèmes en unités supérieures (morphèmes). Hockett [1953] applique la méthode en phonémique et en morphologie où l'encodage d'un morphème

---

4 une règle d'urne (avec remise) est une règle qui reste invariable au cours du temps de l'expérience, comme l'est le tirage de boules dans une urne où pour chaque tirage les probabilités sont remises à zéro (voir [Moreau 1963]).

5 René Moreau (1921-2009), ingénieur, mathématicien et militaire de carrière, a fait du chiffre pendant la guerre d'Indochine. Directeur scientifique d'IBM-France et membre fondateur du Centre de Linguistique Quantitative, il a joué un grand rôle dans la diffusion de la théorie de l'information et des langages formels auprès des linguistes au début des années 1960 (voir [Léon 2010]).

s'effectue en fonction de son contexte: "*wife* is encoded into /*wayv*/ if the next morpheme is the noun-plural *-s* and *-s* is encoded into /*z*/ rather than /s/ when the preceding morphem is *wife*" [Hockett 1953, p.87]. Il signale des expériences qui montrent que la probabilité d'apparition d'un phonème augmente au milieu d'un morphème, et diminue aux frontières. Elle est la plus faible à la frontière entre constituants immédiats. En 1955, Harris mettra cette méthode en œuvre pour segmenter une chaîne transcrite en phonèmes. Par exemple, l'énoncé transcrit /hiyzkwiker/ *He's quicker* sera segmenté selon les points: /hiy.z.kwik.er/.
La méthode prend particulièrement sens lorsqu'il s'agit de décrire des langues amérindiennes non-écrites. Ces linguistes partent d'une transcription en phonèmes de données orales, à savoir des textes formés de récits élicités[6] auprès d'informateurs, au statut anthropologique. Leur première tâche est d'effectuer une analyse morphémique en déterminant les frontières de morphèmes à l'intérieur d'une chaîne de phonèmes. Les morphèmes sont issus d'une segmentation qui ne fait en aucune façon appel au sens. Toutefois cette méthode reste limitée sur le plan linguistique ; Harris précise que seule l'application de méthodes morphologiques traditionnelles est capable de déterminer si les segments obtenus sont des morphèmes ou des mots.
C'est ainsi le même modèle mathématique, les chaînes de Markov, qui s'applique à différents niveaux de la structure linguistique, phonèmes, lettres et mots. Toutefois, la validation des résultats reste l'apanage des linguistes et la linguistique reste l'enjeu essentiel. Depuis, les chaînes de Markov ont été utilisées en traduction automatique puis en TAL pour des tâches de désambiguisation à l'aide du contexte formé par les mots précédents, continuant ainsi à exploiter le caractère séquentiel du modèle. Des formes plus élaborées, les *Hidden Markov Models* (ou chaînes de Markov cachées) sont actuellement toujours largement utilisées en reconnaissance de la parole, de l'écrit, annotation, extraction d'information dans les textes, *data compression*, et *spam filtering* (voir §2.1.4).

1.3. Le mot, unité de mesure

Le mot comme unité de mesure pose un certain nombre de questions, qu'il s'agisse de la fréquence des lettres qui le composent, de la fréquence des mots dans un texte, chez un auteur particulier, dans une langue de spécialité, dans une langue donnée ou dans le langage en général. On peut se demander si la fréquence, et plus généralement la distribution statistique,

---

6 L'élicitation est une technique, utilisée notamment pour la description de langues non écrites, consistant à inciter un locuteur natif, ou informateur, à produire du matériel linguistique et à statuer sur différentes hypothèses.

est une propriété linguistique ou bien si elle n'est qu'une méthode ou un outil pour la linguistique. Autant d'hypothèses qui ont marqué les débuts de la statistique linguistique et qui font encore débat de nos jours, en particulier dans le traitement automatique des grands corpus.

1.3.1. La longueur des mots

L'étude de la longueur des mots dans les textes comme indicateur de style, de comparaison entre auteurs ou de paternité des œuvres existe depuis le milieu du XIX^{e} siècle [Grzybek 2006]. Les distributions statistiques étant peu connues à l'époque, la mesure de la longueur des mots repose sur un simple comptage des lettres. Le mot est alors tenu pour une unité empirique non contestée pour l'étude des textes littéraires. Quant au texte, il est rapporté le plus souvent à un auteur ou un ensemble d'auteurs pour une langue donnée[7].

1.3.2. Distribution des fréquences des mots dans un texte

1.3.2.1. La loi d'Estoup-Zipf

Le premier à s'intéresser à la distribution des fréquences des mots dans un texte est le sténographe Jean-Baptiste Estoup (1868-1950). Ses travaux repris par Zipf ont abouti à la « fameuse » loi d'Estoup-Zipf. Comme le rappelle Petruszewycz [1973, p.42], tous les sténographes se forgeaient des sténogrammes abrégés pour les mots qui revenaient souvent, mais c'est Estoup, le premier, qui va fonder le vocabulaire de ses gammes sur les fréquences d'apparition des mots usuels.

George Kingsley Zipf (1902-1950), est un philologue (américain) et son objectif est d'ordre linguistique : démontrer qu'il existe un principe d'économie dans les langues, fondé statistiquement, ayant pour but de maintenir un équilibre entre le comportement et la forme des éléments du discours (phonèmes, syllabes, morphèmes, mots, phrases). En effectuant des expérimentations sur le latin, le chinois et l'anglais, il découvre que la fréquence relative d'une unité linguistique a une relation inverse à sa complexité : autrement dit, plus un phonème est difficile à prononcer, plus il est rare, et plus une unité linguistique est utilisée, moins elle devient complexe. Pour ce qui concerne les mots, plus ils sont fréquents plus ils sont courts (mesurés en nombre de lettres, de phonèmes ou de syllabes)[8]. Le principe

---

[7] Ces travaux sur la longueur des mots sont actuellement poursuivis à l'aide de méthodes statistiques très perfectionnées [Gryzbek 2006].

[8] « l'accent ou degré de difficulté de tout mot, syllabe, ou son est inversement proportionnel à la fréquence relative de ce mot, cette syllabe ou ce son, parmi les autres mots, syllabes ou sons dans le discours. Plus l'usage est fréquent, moins la forme est accentuée c'est-à-dire plus facile à prononcer, et vice versa » [Petruszewycz 1973, p.42].

d'économie est donc un principe réglant le discours des locuteurs que Zipf formulera ultérieurement comme le principe du moindre effort (1949). Un autre aspect de ce principe d'équilibre concerne la distribution de la fréquence des mots dans un texte, à savoir la relation entre la fréquence des mots et le nombre de mots qui ont la même fréquence. La loi de Zipf (ou loi d'Estoup et Zipf) peut s'énoncer de la façon suivante : quand les mots d'un texte sont rangés par ordre de fréquences décroissantes, la fréquence d'un mot est inversement proportionnelle à son rang, ou encore les mots étant rangés par ordre de fréquences décroissantes, le produit du rang (r) par la fréquence (f) est constant. Soit la formule : $f \times r = constante$

1.3.2.2. La loi de Zipf-Mandelbrot

La loi d'Estoup-Zipf a fait l'objet de nombreuses critiques, mais nous nous intéresserons ici à celle de Mandelbrot[9] dans la mesure où elle concerne le rapport entre mot et texte. Sa critique porte sur le fait que Zipf a pris comme base les « mots formes » (appelés aussi *tokens* ou « formes fléchies »), c'est-à-dire l'ensemble des mots tels qu'ils apparaissent dans un texte. En s'intéressant ainsi aux mots-formes, Zipf leur confère une propriété statistique intrinsèque, valide pour un texte donné dans une langue donnée mais indépendamment de leur usage. La loi d'Estoup-Zipf donne en effet le même résultat quel que soit le texte où elle est appliquée, ce qui implique que tous les textes (dans une même langue) sont tenus comme identiques entre eux. Zipf a apparemment travaillé sur des textes littéraires, quatre pièces de Plaute en latin [Zipf 1935] et *Ulysses* de Joyce [Zipf 1949], mais c'est aussi à partir de syllabes extraites de textes connectés (*connected texts*) en chinois parlé qu'il établit son principe de fréquence. Pour Zipf, ces « textes » sont des échantillons représentatifs d'une langue donnée et n'ont pas de spécificité. Comme le remarque d'ailleurs Mandelbrot [1957], le fait d'avoir choisi *Ulysses* a été un mauvais guide pour Zipf qui pensait avoir à sa disposition le meilleur échantillon possible (de grande dimension et très varié). Or « Ulysses », exceptionnellement long et au vocabulaire exceptionnellement varié, a faussé les résultats. Mandelbrot montre que la méthode ne peut en fait s'appliquer que sur des « formes vides » (appelées aussi *types* ou lemmes). Il propose ainsi une généralisation de la loi d'Estoup-Zipf, qui ne peut être valide que pour un texte donné revêtant ainsi un caractère empirique. Sous sa forme remaniée et

---

9 Benoît Mandelbrot (1924-2010), polytechnicien et mathématicien, surtout connu pour sa théorie des fractales, a mené une carrière franco-américaine (au Caltech, à l'IAS de Princeton et au MIT). Il a effectué de nombreux travaux en théorie de l'information. Plus particulièrement, en 1954, il publie dans la revue *Word* un article sur la loi de Zifp dont il propose une généralisation

généralisée par Mandelbrot, sous le nom désormais de Zipf-Mandelbrot, la loi de Zifp est toujours utilisée à l'heure actuelle par un certain nombre de linguistes statisticiens qui lui reconnaissent un caractère universel et y voient un modèle du langage (voir [Grzybek 2006]). La loi de Zipf pose ainsi la question des lois statistiques comme propriétés intrinsèques d'unités linguistiques, mot et texte. Cette loi s'applique indépendamment du contexte textuel ou discursif dans lequel les mots apparaissent, et indépendamment du sens des mots et des structures grammaticales du texte.

1.3.2.3. Lois statistiques comme propriété linguistique

L'idée que les statistiques sont une propriété du langage est largement partagée parmi les linguistes des années 1950-60. Certains, comme Yule, et à sa suite Guiraud [1954], font une distinction entre le vocabulaire (l'ensemble des mots apparaissant dans un texte) et le lexique (ensemble des mots du stock mémoriel d'où sont tirés les mots du texte). Le vocabulaire d'un texte n'est que le reflet de son lexique. Ils tiennent la distribution rang-fréquence des mots - la loi d'Estoup-Zipf - non comme la propriété d'un texte mais comme une caractéristique du lexique, c'est-à-dire des mots en puissance. Toutefois, Guiraud modère son propos en disant qu'on ne connaît pas le lexique d'une langue dans la mesure où il n'est pas possible de définir la langue en général ; il n'existe que « des » langues : langue parlée, langue littéraire, langue des sciences etc. Autrement dit, les lois statistiques sont plus des caractéristiques des langues de spécialités ou des genres que de la langue en général.

Charles Muller (né en 1909) reprend la discussion en adoptant la distinction de Herdan[10] entre statistique du style et statistique de la langue [Muller 1968]. Cette distinction lui permet d'opposer la contrainte linguistique, définie par ses coordonnées aléatoires (*chance* = hasard), à la liberté individuelle qui est un choix (*choice*). Muller est ainsi conduit à postuler l'existence de la fréquence comme propriété de « la langue », même si celle-ci n'est que potentielle et doit être actualisée dans le discours.

Dans les années 1980, le débat se poursuit entre Muller et Tournier [1980, 1985]. Ce dernier dénonce le fait qu'il y ait des propriétés statistiques de la langue. Pour lui, un discours n'est pas un échantillon statistique de la langue, et il est nécessaire de prendre en compte les

---

10 Gustav Herdan (1897-1968), philologue statisticien d'origine roumaine, s'est attaché à définir des constantes stylistiques à base statistique. En particulier le ratio *type/token* serait une constante pour tous les textes.

conditions d'énonciation du discours. Ces questions sous-tendront les travaux en lexicométrie poursuivis dans l'équipe dirigée par M.Tournier [Lafon 1981, Lafon et Salem 1983].

1.3.2.4. Mots disponibles

La loi d'Estoup- Zipf en tant que propriété des textes (du langage en général) a été critiquée également à partir de la notion de mot disponible. Les auteurs du *Français fondamental* [Gougenheim et al. 1956], destiné à l'enseignement du français, ont montré que tous les mots n'obéissent pas aux mêmes règles statistiques. Ils distinguent les mots « fréquents » (parmi lesquels les mots grammaticaux et les verbes), des mots « disponibles » (noms concrets liés au thème traité dans un texte donné). Ces mots disponibles n'ont pas nécessairement une fréquence élevée au sein d'une langue donnée, certains ont même une fréquence assez faible. C'est le cas par exemple de « fourchette », mot concret disponible pour tous les locuteurs du français, mais qui, pourtant, a une fréquence faible dans la plupart des textes. Pour Moreau [1962] les fréquences des noms concrets sont non seulement faibles mais ne sont pas constantes, ce qui revient à dire que les noms concrets-noms disponibles n'ont pas de fréquences propres et qu'ils ne sont pas probabilisables[11].

1.4. Mot et traitement automatique des langues

1.4.1. Dictionnaires électroniques et linguistique pour l'ingénieur

Les premiers travaux en traduction automatique (TA) aux Etats-Unis ont très largement été fondés sur une analyse syntaxique automatique. Ils mettent en œuvre des grammaires formelles partant du niveau phrase. D'autres travaux ont opté pour la confection de dictionnaires bilingues, première étape de l'automatisation, où le mot est une unité. Conçue comme une technologie de guerre froide destinée à produire des traductions en série de textes scientifiques russes, la TA est une affaire d'ingénieurs. En développant une linguistique pour la machine (*machine translation linguistics)*, les ingénieurs dénient à la linguistique toute légitimité dans l'automatisation de la traduction. Ils fabriquent des dictionnaires de racines et de terminaisons qui n'obéissent pas aux critères fonctionnels d'origine phonétique ou historique des grammairiens. Ils produisent de 'faux' radicaux, appelés bases de mots, de 'fausses' désinences et redéfinissent les affixes. Par exemple, pour le verbe français *saisir*, on choisit le radical *saisi* de préférence à *sai-* qui a des parties communes avec *savoir*. Aussi le

---

11 René Moreau a lui-même effectué des travaux de lexicométrie. Dans son étude du vocabulaire du Général de Gaulle [Cotteret et Moreau 1969], il s'attache à montrer que la longueur des phrases (en mots) est un indice du style d'un auteur.

mot, suite de caractères entre deux séparateurs, entre-t-il en conflit avec les analyses morphologiques des grammairiens, en particulier avec les descriptions morpho-phonémiques chères aux linguistes bloomfieldiens de ces années 1950.

1.4.2. Traduction automatique et nouveaux objets pour la lexicographie

En France, où existe une forte tradition d'études du vocabulaire (voir [Chevalier et Encrevé 2006]), et de statistiques lexicales (cf. les débats relatés §1.3), la possibilité d'automatiser la traduction a ravivé un intérêt pour le mot chez les linguistes et a conduit à la prise en compte des unités lexicales complexes. Il s'agissait de faire coïncider forme graphique (mot-code), unité syntaxique et unité sémantique (unité de traduction). Aussi le premier colloque organisé en 1962 par l'ATALA[12] a-t-il eu pour thème « le mot ». Les travaux de Bernard Pottier, un des pionniers de la TA en France, sont sur ce plan exemplaires. Pottier [1962] soutient que le mot ne peut être défini uniquement par sa forme (une suite de signes graphiques séparés par des séparateurs), par sa signification ou par le rapport forme/signification. Le mot ne peut être qu'une *unité de comportement*. En effet, un élément formel (mot-graphique par exemple) n'est une unité que s'il peut fonctionner librement ; *prendre la mouche* est soit une suite de trois mots (chaque forme gardant son autonomie fonctionnelle et sémantique) soit une suite formant une unité dans la langue. Pottier propose de nommer cette unité de langue une *lexie*. Ce sont les lexies qui doivent être introduites dans les vocabulaires fondamentaux et les dictionnaires de TA.

1.4.3. Collocations et *Corpus Linguistics*

Dans la tradition britannique firthienne [Firth 1951], le mot est d'emblée appréhendé au sein de collocations. Et ce sont les collocations que les premiers travaux en linguistique de corpus se sont donnés pour tâche de repérer au sein de textes (voir [Léon 2008]). Pour J.R. Firth, la notion de collocation est au centre de sa théorie sémantique (*meaning by collocation*). Au départ, elle désigne la cooccurrence de deux éléments linguistiques quelconques dans n'importe quelle étendue de texte. Pour des raison pratiques (enseignement des langues, traduction, fabrication de dictionnaires), Firth a limité l'étude des collocations aux mots en attente mutuelle (*mutual expectation*) permettant de déterminer le sens : dans « nuit noire », le

---

[12] Alors l'acronyme d'Association pour la Traduction Automatique et la Linguistique Appliquée, qui changera en 1992.

sens de « nuit » est qu'il peut entrer en collocation avec « noire » et inversement. En 1957, il ajoute la notion de colligation, où l'attente mutuelle ne concerne plus les mots mais les catégories grammaticales (*mutually expectant order*). Les colligations sont destinées à rendre compte du sens au niveau grammatical et non plus au niveau lexical[13]. Pour Firth, l'étude des collocations (et des colligations) ne peut se faire que selon certaines conditions. Elle doit être effectuée non dans le langage en général mais dans des langages restreints (langues de spécialité, genres etc.). Une deuxième contrainte concerne le texte : pour aborder le sens, il faut étudier des collocations de mots dans des textes authentiques et intégraux et non à partir de corpus échantillonnés. Une troisième contrainte exige que le corpus de textes soit a priori non fini. Bien que la méthode ait été élaborée dès le début des années 1960 [Sinclair 2004], elle n'a pu être mise en œuvre qu'à la fin des années 1980 quand de grandes masses de textes informatisés ont été rendues disponibles.

2. Le *data turn*, ses causes et ses conséquences

Il est désormais classique de repérer un tournant important dans l'évolution du TAL (Traitement Automatique des Langues), qui remonte aux années 1990. Ce tournant ne concerne d'ailleurs pas uniquement ce seul domaine : on l'identifie aussi dans d'autres branches de l'informatique comme le traitement des images, la robotique, le raisonnement, la programmation des jeux de stratégies… La plupart de ce qui, d'une manière générale, relève de l'intelligence artificielle s'en est trouvé bouleversé. La tendance générale amorcée dans ces années tend à substituer aux traitements « fondés sur des modèles formels » opérant de façon *déductive* (automates, grammaires, formalismes logiques…) des traitements « fondés sur des données », opérant de façon *inductive* en faisant appel à des comptes statistiques plutôt qu'à l'expertise des modélisateurs. Nous allons évoquer dans cette partie les causes et les conséquences de ce changement de paradigme dans le champ du TAL, en expliquant comment il contribue à redéfinir les places respectives qu'y occupent les mathématiques et la linguistique, et en montrant que les unités qui nous occupent dans cet article, à savoir les mots et les textes, s'y trouvent redéfinies.

2.1. Le *data turn* en TAL

2.1.1. Les causes du changement

[13] Comme le signale Fortis dans ce même numéro, la linguistique américaine fondée sur l'usage se rapproche de la linguistique empiriste britannique, notamment avec la notion de *collostruction,* très proche des notions firthiennes de collocation et de colligation.

Les années 1980 ont vu l'apparition des premiers micro-ordinateurs à destination du grand public, mais ce n'est que dans les années 1990 que ces machines commencent vraiment à se diffuser. La technologie a alors atteint un certain niveau de maturité : les capacités de stockage et de calcul sont désormais suffisantes pour permettre d'enregistrer dans un disque dur plusieurs textes numérisés, et y appliquer des programmes efficacement. La revue *Computational Linguistics* consacre ainsi en 1993 (Computational Linguistics 93) deux numéros complets au thème « Using Large Corpora », tandis que la revue TAL [TAL 95] lui emboite le pas en 1995, avec un numéro (double, lui aussi) dédié aux « Traitements probabilistes et corpus », qui donnent chacun un panorama assez varié des travaux de l'époque.

Ces progrès matériels s'accompagnent bien sûr de l'apparition du Web (inventé en 1989) qui donne facilement accès à un nombre de plus en plus important de textes. Les logiciels documentaires (Salton 83), jusque-là confinés aux bibliothèques, vont inspirer la conception des premiers moteurs de recherche, qui vont s'avérer de plus en plus indispensables. Google, rappelons-le, naît en 1998.

Conceptuellement, l'informatique en tant que discipline scientifique a aussi beaucoup évolué durant ces années. Nous avons vu dans les sections précédentes que certains aspects de la lexicométrie (on pourrait aussi citer la théorie de l'inférence inductive de [Solomonoff 1964]) étaient précurseurs en matière d'approche fondée sur les données. Les réseaux de neurones, très à la mode dans les années 1980, sont aussi à ranger dans le camp des méthodes qui font confiance à la « force brute » du calcul plus qu'aux raisonnements formels. Un moment passés de mode, ils reviennent en grâce de nos jours, avec des modèles de plus en plus complexes. Le traitement de la parole, qui est un des domaines qui les utilisent dans les années 1980, exploite aussi des méthodes comme les chaînes de Markov (et des extensions comme les chaînes de Markov cachées), et se développe beaucoup dans ces années-là. Et les tout premiers systèmes de traduction automatique statistique sont proposés par des chercheurs d'IBM à la même époque [Brown et *al*. 1993].

D'autres branches émergent aussi, qui n'ont pas forcément à voir avec le TAL : le *data mining* ou fouille de données, est à la fin du XXème siècle la pointe avancée des systèmes de bases de données. L'objectif de ce domaine est en effet d'extraire le maximum d'informations possibles de grandes quantités de faits accumulés au fil des années et stockés de manière structurée dans des tableaux, afin de prévoir l'évolution de certains autres faits. Cela intéresse les professionnels de la banque (qui veulent par exemple prévoir à l'avance la capacité de leurs clients à rembourser un prêt), du marketing (qui cherchent à anticiper les achats pour

faire de la publicité ciblée), les médecins (qui espèrent décider ainsi des médicaments ayant le meilleur effet possible sur leurs patients), etc. Jusque dans les années 1980 ces problèmes étaient abordés dans le cadre de l'intelligence artificielle par des *systèmes experts*, lourds dispositifs qui visent à formaliser la connaissance d'experts humains dans des règles formelles écrites à la main. Le *data mining* opère de façon radicalement différente : il se propose d'accumuler des données pour lesquelles la réponse recherchée est déjà connue (historiques des prêts déjà accordés et remboursés –ou non-, effets mesurés de traitements déjà prescrits, etc.) et cherche à concevoir des programmes capables d'exploiter ces connaissances initiales pour construire lui-même des règles (ou au moins des critères) lui permettant de prévoir l'avenir. C'est le fondement de l'apprentissage automatique supervisé, dont les bases théoriques ont été posées dans les années 1980 [Valiant 84] mais qui explose en proposant des solutions pratiques aux problèmes évoqués précédemment dans les années 1990 (les « arbres de décision » sont par exemples inventés en 1986 [Quinlan 86]).

2.1.2. Les effets sur le TAL

Comment ces nouvelles approches influencent-elles le TAL ? Le partage des données via les réseaux (principalement le Web, bien sûr), et la prédominance des visées applicatives du *data mining* incite le domaine à se restructurer en *tâches*. Jusque-là, en effet, c'était plutôt les niveaux d'analyse de la langue (morphologie, syntaxe, sémantique…) qui caractérisaient les différentes branches du TAL. Dans les années 1990, des conférences commencent à proposer des « *challenges* » (défis applicatifs) pour mettre en concurrence les programmes des participants sur des données communes. Certains d'entre eux sont encore calqués sur les niveaux d'analyse traditionnels (il existe toujours des *challenges* portant sur l'analyse syntaxique par exemple), mais d'autres se focalisent sur des traitements plus originaux : la conférence MUC (Message Understanding Conference) met ainsi à disposition lors de sa première édition, en 1987, des dépêches d'agences de presse évoquant des attentats terroristes (on ne s'étonnera pas que le DARPA, l'agence de recherche du département de la défense américaine, soit impliquée dans cette proposition), et demande aux participants d'extraire automatiquement de ces textes, par programme, des informations factuelles précises (date, lieu, nombre de victimes, revendications éventuelles, etc.), de nature à remplir les champs d'une base de données [Poibeau 2003]. C'est un changement de point de vue fondamental : plutôt que de viser une analyse linguistique pertinente, on cherche juste dans ce type de défi à réaliser une tâche applicative le plus efficacement possible, par quelque moyen que ce soit (« *whatever works !* »). Ce n'est plus la méthode qui est évaluée en tant que telle, mais

uniquement ce qu'elle produit. Les programmes en concurrence sont comparés via des mesures quantitatives précises (précision, rappel, F-mesure…) : ceux qui produisent les résultats les plus proches de solutions de référence (construites ou validées par des humains) sont annoncés vainqueurs. La culture de l'évaluation quantitative commence à occuper le terrain, en TAL comme ailleurs...

2.1.3. Les nouvelles tâches du TAL

Quelles sont donc ces nouvelles *tâches* qui apparaissent, qui héritent de la fouille de données pour constituer ce qu'on appelle désormais la *fouille de textes* [Ibekwe-SanJuan 2007]. On peut lister les principales :

- la recherche d'information (RI, voir l'article de B. Grau et P. Bellot dans le numéro 1, ou [Amini et Gaussier 2013]) est la tâche consistant à sélectionner à l'intérieur d'un corpus les documents qui répondent le mieux à une requête : c'est évidemment ce que font les moteurs de recherche généralistes du Web mais aussi, à une échelle plus limitée, ceux qui opèrent à l'intérieur d'un site particulier. Une version avancée, celle des systèmes « question-réponse », accepte les questions formulées en « langue naturelle » et est censée y répondre de manière factuelle (et non en se contentant de fournir un document pertinent)

- la classification est la tâche consistant à associer automatiquement à un texte une étiquette parmi une liste d'étiquettes possibles : c'est ce que font les logiciels de courriers électroniques qui distinguent tout seuls les « spams » (courriers indésirables) des autres, voire proposent un dossier de rangement parmi ceux créés dans la boîte aux lettres. Mais c'est aussi une tâche qui peut se décliner en de multiples autres variantes : identification de l'auteur d'un texte anonyme (parmi une liste de suspects), de la rubrique dont il relève (pour un article de journal, par exemple), de sa couleur politique, de sa date de rédaction… La variante la plus contemporaine de la classification est celle qui vise à étiqueter comme « positif », « négatif » ou « neutre » un texte porteur d'opinion (critique d'un produit culturel ou matériel, d'un événement, d'une marque…) pour évaluer la « e-réputation » de ce dont il parle : c'est un marché valorisable financièrement, actuellement en plein développement ! Cette tâche de classification peut être envisagée comme le prolongement direct dans le domaine des textes du *data mining*.

- l'extraction d'information (EI) est l'héritière directe de la tâche du challenge MUC, évoqué précédemment : elle vise à remplir les champs d'un formulaire à partir du contenu d'un texte. Cette tâche est emblématique de la mutation subie par le TAL à l'occasion du « data turn » : à défaut de « comprendre » un texte, cherchons juste à en extraire (à en « distiller » dit

McCallum, un des chercheurs de référence du domaine [McCallum 2005]) le contenu pour le rendre exploitable par les machines, c'est-à-dire en le rangeant dans les cases prédéfinies d'une base de données. Le traitement automatique des petites annonces ou des CV, l'analyse bibliométrique des articles scientifiques… sont autant d'applications possibles pour des systèmes capables d'aborder cette tâche [Tellier, Tommasi 2011].

Ces tâches n'ont de potentiel applicatif que si les systèmes qui les traitent sont capables de gérer rapidement de grandes quantités de données. Elles font basculer la recherche du côté de l'innovation technologique, donnant lieu à des applications commerciales capables de gérer des données réelles. On parle d'ailleurs plus volontiers d'ingénierie linguistique que de TAL dans les sociétés qui commercialisent des produits les mettant en œuvre.

2.1.4 Les habits neufs du TAL

Comment réaliser un programme capable de traiter ces nouvelles tâches ?

Remarquons tout d'abord que leur unité de traitement est le *texte,* considéré comme un ensemble de phrases. Cette notion se passe d'ailleurs très bien dans ce cas d'une vraie définition linguistique : c'est l'application qui définit la granularité des textes qu'elle traite. Un texte, pour une tâche de recherche d'information, c'est « ce que fournit un moteur de recherche » (l'identifiant URL d'une page Web, par exemple, même si elle contient plusieurs « textes » au sens linguistique), ou bien c'est la donnée d'entrée d'un système de classification ou d'extraction d'information. Les tweets ou SMS actuels peuvent jouer ce rôle, au même titre que les articles de journaux ou les livres, si le seul objectif visé est de leur attribuer à chacun une étiquette (tâche de classification).

Pour des raisons d'efficacité opératoire, la quasi-totalité des systèmes vont renoncer complètement à procéder à une analyse syntaxique (et encore moins sémantique) des phrases qui y figurent. Les analyseurs syntaxiques sont pourtant de plus en plus performants, mais ils sont soit trop coûteux en temps de calcul (dans le cas de textes longs en particulier), soit trop ambigus (ils produisent plusieurs analyses possibles pour une même phrase), soit encore pas assez fiables (dans le cas de textes ne respectant pas les normes usuelles, comme la plupart des tweets ou des SMS). L'intuition qui prévaut est qu'un traitement superficiel des données suffit, dans la plupart des cas, à réaliser la tâche requise. Puisque le pari est d'utiliser les méthodes les plus simples possibles, pour aboutir aux meilleurs résultats possibles, le *mot* (en tant que suite de caractères comprise entre deux séparateurs dans un texte) retrouve également une nouvelle actualité. C'est l'unité de base sur lequel opèrent en effet la plupart des

programmes du TAL contemporain. Suivant que l'ordre dans lequel les mots apparaissent dans un texte est pris en compte ou non, on distinguera toutefois deux familles d'approches :
- la première est l'approche dite « sacs de mots ». Comme son nom l'indique, elle consiste à ramener un texte à l'ensemble des « mots-formes » qu'il contient, en négligeant leur ordre d'apparition dans le texte. Les mots sont ici de simples unités de découpage du texte. Représenter un ensemble de textes en sac de mots revient ainsi à créer un tableau dont chaque colonne est un « mot » (ou toute autre unité de segmentation préalablement définie : n-grammes de lettres ou de mots, racine si on a appliqué un « raciniseur », lemme si on dispose d'un lemmatiseur, etc.) présent au moins une fois dans ce corpus, et chaque ligne correspond à un texte : la case à l'intersection d'un texte t et d'un mot m est le nombre d'occurrences du mot m dans le texte t (ou toute autre valeur obtenue par une pondération de cette première). Les lignes et les colonnes sont dans un ordre arbitraire, toute notion de séquentialité a été perdue. Procéder ainsi permet de transformer un corpus en un *tableau de nombres*, et de lui appliquer directement les procédures qui ont fait leur preuve pour les tâches de *data mining*. Les mots jouent ainsi pour les textes le rôle que les descripteurs jouaient pour les individus qui cherchent à obtenir un prêt, par exemple : ce sont des attributs dont les valeurs (nombres d'occurrences) sont censées caractériser la donnée. Les systèmes de fouille de textes font en fait la plupart du temps implicitement l'hypothèse encore plus forte que chaque ligne du tableau est un *vecteur* dans un espace normalisé, qui comprend autant de dimensions qu'il y a d'attributs (de colonnes) dans le tableau. C'est le même principe que les coordonnées (x,y) servant à repérer des points ou à dessiner des vecteurs dans le plan à deux dimensions, sauf que ces objets ont maintenant autant de coordonnées qu'il y a de colonnes. Tout l'attirail mathématique des espaces vectoriels devient alors disponible, ce qui permet par exemple de définir simplement des notions de distances entre données (donc entre textes). Chaque mot caractérise donc dans ce cas une dimension de l'espace de représentation des textes (comme les vecteurs de base (x,y) dans l'espace à deux dimensions) : chacun est « orthogonal » à chaque autre, autrement dit complètement indépendant. Ces hypothèses sont bien sûr linguistiquement aberrantes, mais elles présentent l'avantage de simplifier les calculs. Le mot a ainsi perdu son caractère linguistique pour devenir une unité mathématique : c'est une base vectorielle à partir de laquelle on peut définir ce texte par combinaison linéaire (somme pondérée par les nombres d'occurrences). C'est ainsi que fonctionnent à l'heure actuelle les meilleurs systèmes de recherche d'information, de même que ceux qui réalisent des tâches de classification.

- la seconde approche possible préserve la linéarité de la langue , à savoir l'ordre des mots dans le texte, et procède à des simplifications moins radicales : on pourrait la qualifier d' « approche annotative » car elle revient à annoter des portions de textes découpées en unités (ce peut être des phrases segmentées en mots, mais tout aussi bien des textes longs segmentés en phrases, ou en paragraphes) par des étiquettes (autant d'étiquettes qu'il y a d'unités) : l'ordre dans lequel les unités apparaissent se retrouve dans l'ordre des étiquettes. On retrouve là l'intuition initiale des chaînes de Markov (cf . section 1.2.1), appliquée aux séquences de mots. La tâche d'annotation morpho-syntaxique, qui consiste à étiqueter les mots d'une phrase comme « le petit chat est mort » en « DET ADJ NC V ADJ » (DET pour déterminant, ADJ pour adjectif, NC pour nom commun et V pour verbe) est l'instance la plus simple d'une telle tâche. Les systèmes d'extraction d'information ne procèdent pas autrement : pour repérer dans une phrase les noms propres qui y figurent (qui constituent en général la cible de l'extraction), ils cherchent à annoter chacune de ses unités en fonction de leur appartenance ou non à un tel nom propre. Les systèmes de traduction automatique fondés sur des modèles statistiques, quant à eux, exploitent des corpus bilingues alignés, c'est-à-dire des ensembles de phrases qui sont les traductions les unes des autres, exactement comme si chacune servait à « annoter » l'autre. C'est cet alignement qui est la cible principale de la phase d'apprentissage automatique mise en œuvre dans ces systèmes[14]. La figure suivante montre par exemple comment la traduction d'une phrase entre l'anglais au français, présentée dans un tableau de correspondance, se traduit par un couple de phrases annotées : chaque mot de chaque phrase est annoté par l'indice (la position) du mot qui le traduit dans la phrase de l'autre langue.

		J '	aime	le	Chocolat
		1	2	3	4
I	1	X			
like	2		X		
chocolate	3				X

J'aime le chocolat | I like chocolate

1 2 - 3 1 2 4

[14] Voir l'article d'Holger Schwenk dans *Information Grammaticale* n°141

Les traductions obtenues sont des « mot à mot » ou, au mieux, des « groupe de mots » à « groupe de mots » réordonnés. L'approche « annotative » est un peu moins frustrante pour le linguiste que l'approche « sac de mots » car les unités y sont considérées « en contexte » : pour choisir quelle étiquette associer à l'une d'elles, on a le droit de tenir compte des unités environnantes, voire des étiquettes des unités environnantes, quand elles sont déjà connues. Mais c'est toujours un contexte limité, borné (ce que recouvre exactement le terme « markovien » en mathématiques).

2.2. Un nouveau rapport à la linguistique

Les nouvelles recherches en TAL que nous venons d'évoquer, qu'elles se rattachent à une approche « sac de mots » ou à une approche « annotative », ont massivement recours à l'apprentissage automatique [Cornuejols et Miclet 2002]. La réalisation des tâches n'est ainsi pas directement programmée par un informaticien ou un linguiste : les chercheurs en TAL se contentent désormais de recueillir des exemples de données, si possible associées au résultat souhaité (exemples de mails transformés en sacs de mots pour lesquels on sait s'ils sont ou non des spams, par exemple, ou exemples de textes où les noms propres qui remplissent les champs d'un formulaire d'extraction d'information sont annotés) et de les confier à un programme d'apprentissage qui recherchera lui-même les paramètres pertinents qui relient les données aux résultats. La linguistique a-t-elle encore quelque chose à apporter à ces techniques ? C'est ce que nous allons envisager dans cette dernière partie (voir aussi [Tellier 2009]).

2.2.1. La linguistique comme pourvoyeuse de ressources

L'apprentissage automatique n'a rien de magique : pour que les programmes « apprennent » quelque chose de pertinent, c'est-à-dire qui soit applicable sur de nouvelles données non encore observées, il faut leur fournir des exemples de bonne qualité et en grand nombre. La plupart, comme on l'a vu, requièrent des *données associées au résultat attendu* : c'est ce que l'on appelle de *l'apprentissage supervisé*. De même que les banquiers se fient, pour accorder un nouveau prêt à un nouveau client, sur l'historique des remboursements passés d'autres clients, un détecteur de spams repose sur la mise à disposition d'exemples de courriers indésirables et d'exemples de textes acceptables. Les systèmes de gestion de courrier électronique effectuent donc d'autant mieux cette tâche que leurs utilisateurs leur ont signalé au fur et à mesure de leur arrivée les courriers indésirables non encore repérés. Ce type d'information est relativement facile à obtenir, mais collecter des textes intégraux où les noms

propres sont identifiés, ou des corpus bilingues alignés, est nettement plus problématique. Les linguistes sont ainsi très souvent mis à contribution pour produire (ou au minimum corriger et valider à la main des versions préliminaires imparfaites) ces corpus de référence qui serviront à alimenter les systèmes fondés sur l'apprentissage automatique. Une partie de ces précieuses données de référence (on ne les appelle sans doute pas pour rien des « gold standard » : elles valent de l'or !) sert aussi à valider les programmes appris : on les fait fonctionner sans, bien sûr, leur donner accès au résultat attendu, et on compare ce qu'ils produisent avec ce résultat, afin de calculer les fameuses mesures d'évaluation qui servent à quantifier leur performance. La « linguistique », dans ce cas, intervient en amont (comme fournisseuse de données d'apprentissage) ou en aval (pour l'évaluation du résultat final) du processus de construction du programme.

Mais elle peut aussi être mise à contribution pendant la phase d'apprentissage elle-même. Les programmes d'apprentissage automatique, en effet, sont capables de tirer parti de tous les types d'informations qu'on leur fournit. De même qu'un banquier se trompera d'autant moins dans l'appréciation de son client qu'il dispose sur lui d'informations pertinentes, de même un programme chargé de classer des textes ou de les annoter bénéficiera en général d'informations de nature linguistique intégrées à ces textes : les « mots vides » (mots grammaticaux ou très fréquents) peuvent ainsi être exclus de la liste de ceux qui définissent l'espace de représentation des « sacs de mots » sans perte de performance, tandis que la présence d'un mot dans une liste de noms propres prédéfinie est une indication précieuse qui peut aider un programme à décider si on l'étiquette en vue de l'extraire ou non. Des « ressources linguistiques » sont ainsi intégrées parmi les indices dont dispose le système d'apprentissage automatique pour prendre sa décision. On remarque, là encore, que ce sont principalement des ressources de nature lexicale (présence dans des listes, des dictionnaires...) qui sont exploitées. La linguistique qui regarde au-delà du niveau des mots est très rarement utilisée dans ce cadre.

2.2.2. La linguistique comme effet secondaire à commenter

Un autre type d'interaction entre TAL et linguistique peut également émerger de ces nouvelles approches, en se focalisant sur la compréhension des résultats des programmes et l'interprétation de leurs erreurs. Les programmes d'ingénierie linguistique, on l'a vu, visent surtout à la réalisation d'une tâche : c'est là-dessus seul qu'ils seront évalués. Néanmoins, certains d'entre eux produisent également, pendant leur fonctionnement (soit pendant la phase d'apprentissage sur des exemples étiquetés, soit pendant leur application sur de nouvelles

données), des traces exploitables. Les « arbres de décision », par exemple, sont des modèles qui non seulement classent les objets (en l'occurrence les sacs de mots, quand ils sont employés en fouille de textes) sur lesquels ils opèrent, mais qui représentent dans une arborescence la succession des critères ayant permis d'aboutir à ce classement. Pour des textes, chaque noeud de l'arbre sera ainsi un critère du type : « le mot m est présent au moins n fois dans le texte » : l'une des branches qui suit ce noeud correspond à une réponse positive, l'autre à une réponse négative. Chaque classement global se « lit » alors en parcourant les critères depuis la racine jusqu'à une feuille, qui contient le diagnostic final : la classe choisie. Plus un critère se trouve proche de la racine, plus il est discriminant, c'est-à-dire important pour la décision finale. On obtient ainsi indirectement un classement de l'ordre d'importance des mots dans un texte en vue de le classer, qui peut être confronté à l'intuition d'un linguiste. Mais tous les programmes d'apprentissage automatique (et il existe de nombreux, très différents les uns des autres !) ne proposent pas des « sorties » aussi lisibles pour un humain que les arbres de décision : il est souvent difficile d'interpréter les raisons qui leur ont permis d'aboutir à un résultat plutôt qu'à un autre, et ce d'autant plus quand cette décision résulte de la combinaison pondérée d'une multitude de critères. A défaut d'en comprendre vraiment les motifs, les linguistes en sont alors réduits à commenter et analyser les erreurs commises par les programmes, à chercher les sources des confusions qu'ils ont faites, et à essayer d'y remédier en fournissant des exemples qui lèvent les ambiguïtés, ou une ressource qui contient des indices qui manquaient. C'est un travail souvent difficile et ingrat.

Conclusion

Qu'est-ce qui caractérise le *data turn* ? Alors que dans les années 1960, la linguistique apparaît comme le moteur des traitements statistiques, elle perd ce rôle pilote dans les années 1990, où elle n'est plus souvent qu'une ressource, qu'un recours éventuel en cas de nécessité. La tâche est au premier plan et abolit toute réflexion théorique sur le statut des unités traitées, qui est désormais fixé par les mathématiques employées.

Dans les années 1960, les lois statistiques sont l'objet de débats afin de déterminer si elles peuvent être considérées comme des universaux linguistiques ou des propriétés intrinsèques des unités de langue ou de discours. En tout cas les statistiques ne sont jamais de simples outils ; leur utilisation en linguistique est un problème de linguistique. Ce type de position est manifeste dans les conclusions du premier colloque *Statistique et analyse linguistique* qui a eu

lieu à Strasbourg en 1964 : « (1) il est essentiel de distinguer entre la quantification des éléments du discours et la description quantitative de la langue ; (2) l'utilité des méthodes quantitatives semble évidente dans le domaine de la philologie et particulièrement dans l'histoire des textes, l'histoire externe des langues, les recherches sur la datation et l'attribution des textes ».

Les unités « texte » et « mot » sont au premier plan. Bien qu'admis unanimement comme unité de traitement statistique, le mot-code - chaîne de caractères alphanumériques entre deux séparateurs - reste l'enjeu d'un débat linguistique. Un mot est un vocable, une unité de discours ou de traduction. Les collocations sont des unités sémantiques. Le nombre de mots dans une phrase est un indice du style d'un auteur. Grâce au mot-code, émerge une nouvelle entité linguistique, les unités lexicales complexes. Autrement dit, le mot-code issu de la cryptographie et des télécommunications est habilité comme unité linguistique. Quant au texte, échantillon de langue pour Zipf, il est un objet pour l'analyse stylistique (travaux de Guiraud et de Muller) ou discursive dans les travaux lexicométriques de Moreau ou Tournier, où il devient une unité de corpus.

A partir des années 1990, les tâches qui structurent désormais le domaine sont plus applicatives, plus neutres au sens où elles ne requièrent pas une analyse linguistique intermédiaire. Avec la montée en puissance des systèmes d'apprentissage automatique issus des laboratoires d'informatique, le pouvoir d'initiative a changé de camp. Désormais, la linguistique est la plupart du temps assujettie à l'exécution d'un programme d'apprentissage automatique, aux données dont ce programme a besoin pour fonctionner, ou aux traces qu'il produit pendant son exécution. Certes, à performance équivalente sur une tâche donnée, un programme qui produit en outre une trace linguistiquement valide sera préféré à un autre. On lui fait en quelque sorte crédit d'un « supplément d'âme » linguistique. Mais celui-ci ne suffit jamais à compenser un défaut de performance, qui prime avant tout.

Une question subsiste : le fait que « ça marche », que la méthode par alignement produise des traductions meilleures que les méthodes à base de règles (linguistiques), n'apporte certes rien à la linguistique, mais est-ce que ça nous apprend quelque chose sur les usages d'une langue? On peut tenter de répondre à cette question en disant que le TAL actuel est « fondé sur les données » au lieu d'être « fondé sur des modèles abstraits (types grammaires formelles…) ». Il fait l'hypothèse que l'usage prime sur l'intuition, la performance sur la compétence, pour reprendre la distinction chomskyenne. A défaut de donner des clés sur comment fonctionne le langage, ce parti pris rend les systèmes plus opérationnels, la quantité compensant la qualité.

Dans les traitements apparus à partir des années 1990, les aspects linguistiques des unités texte et mot ont été en grande partie abandonnés. Dans la fouille de textes, le mot devient une « base » au sens vectoriel du terme : chacun est considéré comme orthogonal (indépendant) des autres. Et le texte, transformé en tableau, est alors réduit à une combinaison linéaire de mots (un vecteur) dans l'espace ainsi défini. Les deux se ramènent à des entités mathématiques. Dans la tâche d'annotation, certains traits linguistiques de l'objet traité sont préservés, en particulier l'ordre des mots, que l'on peut notamment traiter par les chaînes de Markov cachées. Mais, hormis lorsqu'il s'agit de corpus qui suscitent des débats d'homogénéité et de représentativité d'ordre linguistique, le texte désigne désormais avant tout le donné sur lequel s'applique la tâche.

Autrement dit d'unités linguistiques dans les années 1950-60, « mot » et « texte » sont devenus, à partir des années 1990, des entités mathématiques (d'ailleurs une formation en mathématiques devient un prérequis indispensable pour comprendre les systèmes de TAL actuels).

Enfin, les contextes sociologiques sont différents. Dans les années 1960, l'analyse statistique était discutée par des ingénieurs, des mathématiciens, des informaticiens et des linguistes. Même si la primauté de la tâche a toujours suscité des conflits entre linguistes et talistes et même si les objectifs économiques de rentabilité ont toujours eu une place importante dans le TAL – c'est vrai dès les premières expériences de traduction automatique – celui-ci a longtemps été un domaine pour la recherche et développement. A l'heure actuelle, la fouille de textes a largement dépassé le cadre des informaticiens, des mathématiciens et des linguistes. Le débat a échappé au cadre universitaire et aux objectifs de production du savoir, pour se focaliser sur des comparatifs de performances.

Bibliographie

Amini M.-R. et Gaussier E. (1993). *Recherche d'information*, Eyrolles.

Antié J., Kelih E., Grzybek P. (2006). Zero-Syllable Words In Determining Word Length » In Grzybek Peter (éd.), 2006, *Contributions to the Science of Text and Language. Word Length Studies and Related Issues* Dordrecht, Springer, 117-156.

Baratin M., Cassin B., Rosier-Catach I., Ildefonse F., Lallot J., Léon J. (2004). Le mot, In Cassin B. (éd.),*Vocabulaire Européen des Philosophies, Dictionnaire des intraduisibles,* Seuil- Le Robert, 830-844.

Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L. (1993) The Mathematics of Statistical Machine Translation : Parameter Estimation, *Computational Linguistics* 19-2, 263-311.

Chevalier J.-C. avec Encrevé P. (2006). *Combats pour la linguistique, de Martinet à Kristeva*, Lyon, ENS Editions.

Computational Linguistics, 1993, *Computational Linguistics*, Special Issue on Using Large Corpora (I et II), volumes 19-1 et 19-2

Cori M. et Léon J. 2002. La constitution du TAL. Etude historique des dénominations et des concepts, *Traitement Automatique des Langues*, 43-3, 21-55.

Cornuéjols A., Miclet L. 2002 *Apprentissage artificiel ; concepts et algorithmes*, Paris, Eyrolles.

Cotteret J.-M. et Moreau R. 1969. Le vocabulaire du général de Gaulle 1958-59. *Cahier de l'Herne* 21, 217-245.

Estoup J.-B. 1916. *Les gammes sténographiques* Paris

Firth J.R. 1957 [1951]. Modes of Meaning, In *Papers in Linguistics (1934-1951),* Oxford University Press, 190-215.

Firth J.R.,1968 [1957]. A synopsis of linguistic theory 1930-55, In Palmer F.R. (éd.), *Selected papers of J.R. Firth (1952-59),*168-205.

Gougenheim G., R.Michea, P.Rivenc et A.Sauvageot. 1956. *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

Grzybek P. (éd.). 2006. *Contributions to the Science of Text and Language. Word Length Studies and Related Issues,* Dordrecht, Springer.

Guiraud P. 1954. *Les Caractères statistiques du vocabulaire*, Paris, PUF.

Hockett Ch. F. 1953. Review of Shannon Cl. E. et Weaver W. *The Mathematical Theory of Communication* 1948, Urbana Ill. *Language* 29-1, 69-93.

Kyheng R. 2005. Hjelmslev et le concept de *texte* en linguistique. *Texto* [en ligne], 10-3. <http://www.revue-texto.net/Inedits/Kyheng/Kyheng_Hjelmslev.html.

Ibekwe-SanJuan F. 2007. *Fouille de textes*, Lavoisier, Hermès.

Lafon P., 1981, Analyse lexicométrique et recherche des cooccurrences, *Mots* 3, 95-147.

Lafon P. et Salem A. 1983. L'inventaire des segments répétés d'un texte, *Mots* 6, 161-177.

Léon J. 2001. Conceptions du mot et débuts de la traduction automatique, *Histoire Epistémologie Langage*, 23-1, 81-106.

Léon J. 2003. Proposition, Phrase, Enoncé : parcours historique. *L'information grammaticale* 98, 5-16.

Léon J. 2007. Meaning by collocation. The Firthian filiation of Corpus Linguistics, In (D. Kibbee (éd.), *Proceedings of ICHoLS X, 10th International Conference on the History of Language Sciences*, John Benjamins Publishing Company, 404-415.

Léon J., 2008. Automatisation du traitement des unités lexicales composées. Perspective historique », B. Kaltz (éed.) *Regards croisés sur les mots non simples*, Lyon, ENS-LSH Editions, 177-191.

Léon J. 2010. Automatisation-mathématisation de la linguistique en France dans les années 1960. Un cas de réception externe. In F.Neveu, V.Muni-Toke, J.Durand, T.Kingler, L.Mondada, S.Prévost (éds.) *Actes du 2e Congrès Mondial de Linguistique Française*, Paris:EDP Sciences, 825-838. (www.linguistiquefrancaise.org)[DOI10.1051/cmlf 2010158]

Léon J. (sous presse), *Histoire de l'automatisation du langage*, Lyon, ENS-Editions.

Mandelbrot B. 1957. Linguistique macroscopique » In Apostel, Mandelbrot et Morf (éds.) *Logique, langage et théorie de l'information,* Paris, PUF, 1-78.

Mandelbrot B. 1961. On the theory of word frequencies and on related Markovian Models of Discourse. In Jakobson R. (éd.) *Structure of Language and its Mathematical Aspects* Proceedings of Symposia in Applied Mathematics, vol XII, Providence, Rhode Island, American Mathematical Society, 190-219.

Mandelbrot B. 1968. Les constantes chiffrées du discours In Martinet A. (éd.), *Encyclopédie de la Pléiade Le Langage*, 46-57.

Markov A.A. 1913. Exemple d'une étude statistique d'un texte extrait de 'Eugene Oneguine' illustrant les probabilités liées, *Bulletin de l'Académie Impériale des Sciences de St Pétersbourg*, 153-162. Traduction Française Comité d'action scientifique de défense nationale T/R/-427-561 [Archives HTAL].

McCallum A. 2005. Information Extraction : Distilling Structured Data from Un- structured Text *Bulletin of the Association for the Computing Machinery*, 3-9.

Michéa R. 1967. La relation rang-fréquence et la structure statistique de la langue parlée, *BSL* 62, 9-14

Moreau R. 1962. Au sujet de l'utilisation de la notion de fréquence en linguistique, *Cahiers de lexicologie, vol 3*, 140-158.

Moreau R. 1963. Sur la distribution des formes verbales dans le français écrit, *Etudes de Linguistique Appliquée* 2, 65-88.

Moreau R. 1963b. Linguistique statistique et calcul automatique *IBM Point* 6, 6-12.

Muller Ch. 1968 *Initiation à la linguistique statistique*, Paris, Larousse.

Petruszewycz M. 1973. L'histoire de la loi d'Estoup-Zipf : documents, *Mathématiques et Sciences Humaines* 44, 41-56.

Petruszewycz M. 1981. *Les Chaînes de Markov dans le domaine linguistique*, Genève, Paris, Editions Slatkine.

Poibeau T. 2003. *Extraction automatique d'information*, Hermès, Paris.

Quinlan J.R. 1986. Induction of Decision Trees, *Machine Learning* 1, 81-106.

Salton G. 1983. Introduction to Modern Information Retrieval, M.J. McGill.

Solomonoff R. J. 1964. A Formal Theory of Inductive Inference, *Information and Control* 7, 1-22 et 224-254.

Statistique et analyse linguistique. Colloque de Strasbourg (20-22 avril 1964) COLLECTIF. 1966. Paris : PUF.

Stubbs M. 1993. British Traditions in Text Analysis – From Firth to Sinclair In Baker M., Francis G., Tognini-Bonelli E. (éds.) *Text and Technology. In Honour of John Sinclair,* Amsterdam, John Benjamins, 1-36.

TAL ,1995, *TAL*, Traitements probabilistes et corpus, 36 1-2.

Tellier I., 2010. Préface au numéro 50-3 "Apprentissage automatique pour le TAL", revue *TAL*, 7-21.

Tellier I., Tommasi M. 2011. Champs Markoviens Conditionnels pour l'extraction d'information, In Gaussier E. et Yvon F. (éds), *Modèles probabilistes pour l'accès à l'information textuelle*, Paris, Hermès, 223-267.

Tournier M. 1980. D'où viennent les fréquences de vocabulaire ? *Mots* 1, 189-209.

Tournier M. 1985. Sur quoi pouvons-nous compter ; réponse à Charles Müller? *Verbum* 8, 481-492.

Valiant L. G. 1984. A Theory of the Learnable, *Communications of the ACM*, 27-11, 1134-1142.

Zipf G. K. 1935. *Psycho-Biology of Language*, Boston, Houghton Mifflin.

Zipf G. K. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology,* Cambridge MA, Addison-Wesley Press.