Introduction[1]

L'automatisation des sciences du langage[2] commence avec les premières expériences de traduction automatique (désormais TA) entreprises à la fin de la seconde guerre mondiale, en 1948-1949 aux États-Unis et en Grande-Bretagne. On parlera de « tournant de l'automatisation »[3] ; même s'il caractérise la façon dont les linguistes ont adopté ou intégré les concepts ou les méthodes de l'informatique et/ou des mathématiques, ce tournant comporte des traits qui lui confèrent un statut tout à fait spécifique. Il se caractérise d'abord par l'apparition brutale d'un ensemble de concepts, méthodes et pratiques totalement nouveaux, qui n'appartenait pas à l'« horizon de rétrospection » (Auroux 1987, 2007) des linguistes et des sciences du langage.

> Parce que c'est un phénomène social (et non biologique), l'activité scientifique est impossible sans transmission du savoir, sans le fonctionnement institutionnel des formations, du cumul des connaissances (par exemple sans bibliothèque), et sans la mémoire individuelle. Le texte scientifique comprend essentiellement deux types d'expressions. Les unes font référence au seul domaine des phénomènes, manipulés

1 Je tiens à remercier les relecteurs d'ENS Éditions pour leurs très précieuses remarques qui ont beaucoup contribué à l'amélioration de cet ouvrage.

2 Cet ouvrage est une synthèse de travaux menés entre 1997 et 2010 sur l'automatisation du langage. Il est issu d'un mémoire d'habilitation à diriger des recherches soutenue en décembre 2010. Certains chapitres sont originaux, d'autres reprennent partiellement des articles déjà parus et figurant dans la bibliographie.

3 Voir Rorty (1967) *The Linguistic Turn* sur la méthode philosophique et l'attention portée à l'importance du langage dans la formulation des questions philosophiques.

> à l'aide de concepts appartenant à la connaissance commune et/ou produits par l'énonciateur. Les autres font référence à d'autres travaux, par définition antérieurs. Nommons *horizon de rétrospection* HR, cet ensemble de références […] la structure de l'horizon de rétrospection est une cause dans la production momentanée de la recherche […] Mais à l'inverse, la structure du système scientifique détermine celle des horizons de rétrospection. (Auroux 1987, p. 29)

Ce nouvel horizon de rétrospection est instauré par une technologie, la traduction automatique, issue des sciences de la guerre (Dahan et Pestre 2004). Celles-ci, caractérisées par l'interaction entre sciences de l'ingénieur et sciences fondamentales, comprennent notamment les mathématiques, la logique, la physique, les neurosciences, l'acoustique, et les sciences nouvellement apparues que sont la cybernétique et la théorie de l'information. La linguistique, notons-le, ne fait pas partie des sciences de la guerre. Développées essentiellement au MIT, les sciences de la guerre ont permis l'élaboration de technologies de pointe comme les radars, les systèmes de défense antiaérienne et les ordinateurs, puis, après-guerre, la traduction automatique.

Le tournant de l'automatisation du langage se divise en deux temps. La TA, tout en instaurant un nouvel horizon de rétrospection, projette un avenir, un *horizon de projection*, pour les sciences du langage.

> Parce qu'il est limité, l'acte de savoir possède par définition une épaisseur temporelle, un horizon de rétrospection, aussi bien qu'un horizon de projection. (Auroux 1995, p. 49)

C'est la linguistique computationnelle et le programme chomskyen qui vont constituer cet horizon de projection, anticipant l'avenir de l'automatisation-mathématisation du langage ainsi instituée.

Cette période d'une quinzaine d'années (1948-1966), entre le début des premières expériences de TA et la mise en place de la linguistique computationnelle, peut être considérée comme un véritable *événement*[4], constitutif du tournant de l'automatisation.

Le tournant de l'automatisation est associé à la seconde mathématisation du langage. La première mathématisation du langage, qui a eu lieu dans les années 1930, avec la formalisation proposée par l'École de Vienne, et en particulier Carnap, comme horizon commun à toutes les sciences, instituait les mathématiques comme un langage parmi d'autres. La première mathématisation du lan-

4 « Événement » est utilisé ici au sens d'événement historique, qui, lorsqu'il arrive, a une importance sur le cours des choses. Il apporte quelque chose de nouveau qui servira de référence pour un groupe social ou une communauté scientifique donnée. Il est donc susceptible d'une mise en récit.

gage se caractérise par la mise en interaction d'algorithmes et de langages formels issus de la logique mathématique. La seconde mathématisation mise en place grâce à la TA institue un domaine faisant l'interface entre l'analyse syntaxique, les langages formels et la programmation. Les algorithmes abstraits de la première mathématisation s'inscrivent dans la seconde mathématisation, dans l'espace et le temps de la programmation sur ordinateur. C'est pourquoi on appellera cette dernière automatisation-mathématisation.

Ce second tournant, automatisé, de la mathématisation du langage a commencé par la mise en œuvre de méthodes d'analyse syntaxique pour la TA, avant de s'imposer comme domaine de recherche autonome et institutionnalisé. On peut avancer que c'est grâce à la TA, c'est-à-dire grâce à la nécessité stratégique de produire des traductions rentables en série, que les langages formels, ancrés dans le développement de la logique mathématique des années 1930-1940, se sont investis dans des algorithmes d'analyse syntaxique qui ont déterminé l'essor des grammaires formelles, notamment celles de Chomsky.

Ce tournant de l'automatisation est marqué au départ par un paradoxe, à savoir que, bien que la traduction automatique implique le traitement (automatique) des langues, la linguistique ne fait pas partie des sciences de la guerre[5]. Ainsi, pour les sciences du langage, le nouvel horizon ne serait pas le produit de l'annulation d'un horizon antérieur (Auroux 1987). Il est entièrement nouveau et constitué de façon externe. Mais, parce que la traduction automatique, et à sa suite la linguistique computationnelle, sont avant tout aussi affaire de traitement des langues, ce champ nouveau s'impose aux sciences du langage et interrompt le (ou les) processus cumulatifs en cours, qui vont devoir l'intégrer ou dans lesquels elles vont devoir s'intégrer. Au premier moment de l'événement constitutif du nouvel horizon de rétrospection va succéder un mouvement d'intégration.

Un deuxième moment-clé de l'automatisation du langage peut être identifié dans les années 1990, lorsque la puissance des ordinateurs va permettre de traiter des données textuelles en nombre et que la mise à disposition des micro-ordinateurs va conduire les linguistes à utiliser des données informatisées et de nouveaux outils linguistiques. Ce second tournant, qu'on pourrait qualifier de « *corpus turn* », a cependant des caractéristiques bien différentes du premier tournant constitué par la TA et la linguistique computationnelle. Contrairement à celui-ci, l'utilisation des corpus s'inscrit dans la continuité. Elle permet de mettre en œuvre des hypothèses appartenant à des courants des sciences du

5 Martin Joos, ingénieur acousticien et phonéticien, fait exception. C'est probablement le seul linguiste ayant eu une activité dans les sciences de la guerre (voir chapitre 3, § 2.1).

langage antérieurs au premier tournant de l'automatisation, ou de renouer avec des méthodes apparues au moment de l'événement fondateur puis abandonnées ensuite, comme les méthodes probabilistes issues de la théorie de l'information.

Dans cet ouvrage, on s'intéressera moins aux conséquences sociales de la mécanisation du langage (Auroux 1996) qu'aux divers modes d'intégration par les sciences du langage du nouvel horizon de rétrospection institué par l'automatisation. Nous l'appréhenderons à travers un certain nombre de questions :

(i) est-ce que, comme le laisserait supposer le développement de la linguistique computationnelle, l'automatisation des sciences du langage est associée à une seule forme de mathématisation, logico-mathématique, ou d'autres formes d'automatisation-mathématisation sont-elles possibles ?

(ii) les modes d'intégration du nouvel horizon de rétrospection ne peuvent s'envisager que de façon comparative ; sont ainsi examinées les traditions américaine, britannique et française, et, dans une moindre mesure, la tradition russe dont les sources nous sont moins accessibles. Le choix de ces traditions n'est pas fortuit, il nous est imposé par la TA comme technologie de guerre. Les pays considérés sont les « vainqueurs » de la seconde guerre mondiale ; ils sont engagés dans le conflit de la guerre froide où la TA occupe une place stratégique. Beaucoup plus que d'autres qui suivront, et de façon beaucoup plus massive, ces États ont investi des moyens considérables dans la TA. On peut alors se demander si les traditions linguistiques et intellectuelles, encore bien distinctes en cette période de fin de guerre, ont déterminé des modes d'intégration différents, et de quelle façon ;

(iii) on examinera comment l'espace ouvert par le nouvel horizon et par son instanciation dans la linguistique computationnelle va susciter l'émergence du traitement automatique des langues (TAL) et de l'intelligence artificielle ;

(iv) on se demandera aussi dans quelle mesure la possibilité même de l'automatisation peut faire émerger de nouveaux objets, de nouvelles représentations ou de nouvelles méthodes dans les sciences du langage. On verra que, grâce à l'automatisation, la sémantique lexicale va se trouver renouvelée à partir d'anciennes questions sur le « mot » en tant qu'unité linguistique, selon des perspectives diverses ;

(v) on se demandera si les concepts et les méthodes sont intégrés globalement ou bien si des choix sont effectués, si certaines méthodes sont privilégiées par rapport à d'autres et comment. On pense notamment à la théorie de l'information, théorie centrale, unificatrice et universalisante, qui va connaître des destins variés au moment de l'intégration, distincts de celui de la linguistique computationnelle ;

(vi) une autre série d'interrogations va porter sur la périodisation. On se demandera si, à partir de cet événement que constitue le tournant de l'automati-

sation, on peut délimiter une périodisation linéaire, avec un commencement, un début et une fin d'intégration ; ou bien, au contraire, si les divers modes d'intégration vont déterminer des périodisations diverses, parfois ancrées dans les siècles antérieurs, et toujours en cours aujourd'hui ;

(vii) enfin, on peut se demander si cette troisième révolution technologique constitue une révolution des sciences du langage comparable aux deux premières, déterminées par l'écriture et la grammatisation des vernaculaires (voir Auroux 1994).

Cet ouvrage a pour objectif de rendre compte de trois mouvements, la traduction automatique comme événement fondateur de l'« *automatic turn* », l'intégration par les sciences du langage du nouvel horizon de rétrospection, et le second tournant constitué par les corpus. Ces trois mouvements seront développés sous forme de neuf chapitres. Les quatre premiers chapitres sont consacrés aux États-Unis, où tout a commencé. Le premier chapitre « La traduction automatique comme technologie de guerre » permet de rendre compte de l'événement constitutif du tournant de l'automatisation. Dans le second chapitre « De la TA à la linguistique computationnelle et au TAL » est examinée la façon dont le nouvel horizon de rétrospection des sciences du langage s'est transformé en linguistique computationnelle, grâce à l'analyse syntaxique condensant les résultats de la linguistique structurale, de la première mathématisation et de l'algorithmisation rendue possible par la TA, puis comment s'est constitué le domaine appelé actuellement traitement automatique des langues. Le chapitre 3 « Effort de guerre, technologisation de la linguistique et naissance de la linguistique appliquée » est orienté vers la technologisation des sciences du langage. Il est consacré à l'effort de guerre entrepris par les Américains en matière d'enseignement des langues, dans lequel la plupart des linguistes américains étaient engagés. Beaucoup étaient également impliqués dans la cryptographie, la plupart comme simples traducteurs de messages en langues « rares », mais certains ont aussi participé aux travaux de décodage proprement dits. Cet effort de guerre a conduit à l'émergence de la linguistique appliquée aux États-Unis, qui se caractérise par une importante technologisation des méthodes. Automatisation et technologisation des sciences du langage sont ici étroitement associées. Le chapitre 4 « La théorie de l'information : transfert de termes, concepts et méthodes » est moins concerné par l'automatisation que par la mathématisation du langage. Il s'agit d'examiner le processus par lequel certains concepts et certaines méthodes de la théorie de l'information, faisant interagir ingénierie des télécommunications et théories mathématiques, ont pu être intégrés dans les sciences du langage ; la théorie des traits distinctifs de Roman Jakobson présente un cas exemplaire de ce processus, associant linguistique et ingénierie européennes et environnement des sciences de la guerre

américain. Le chapitre 5 traite du mode d'intégration de l'automatisation dans la linguistique américaine. Intitulé « Tournant de l'automatisation et formalisation chez les linguistes distributionnalistes néo-bloomfieldiens », il examine comment la possibilité d'automatisation a suscité de nouveaux enjeux pour les linguistes structuralistes américains autour des questions de traduction et de formalisation. À partir des chapitres suivants, on quitte le domaine américain proprement dit. Dans les chapitres 6, 7 et 8, ce sont d'autres traditions que la linguistique américaine qui sont examinées. Dans le chapitre 6 « Automatisation de la traduction, sémantique et lexique : l'inscription de nouvelles questions et de nouveaux objets dans le temps long », on s'attache à montrer que la possibilité même de l'automatisation a déterminé la mise au jour d'objets qui, bien qu'inscrits dans des traditions linguistiques et intellectuelles différentes (britannique, russe ou française), ont renouvelé certaines questions concernant notamment le lexique. Où l'on voit également qu'un changement de focale fait apparaître un changement de périodisation, et qu'au temps très court de l'événement TA et du tournant de l'automatisation peut être opposé un temps long remontant parfois à plusieurs siècles (voir Chiss et Puech 1999).

On examinera plus particulièrement la situation en France dans le chapitre 7 « Tradition linguistique française et réception externe de la mathématisation-automatisation du langage » et le chapitre 8 « Documentation automatique et analyse automatique de discours. Spécificité des réceptions de Harris en France ». Contrairement à ce qui se passe aux États-Unis, le nouvel horizon de rétrospection est totalement étranger à la tradition linguistique en France, d'où une réception complètement externe de la TA et de la linguistique computationnelle, et la nécessité de passeurs, lieux et personnalités. L'automatisation connaît en France une configuration singulière où sont associées documentation automatique, analyse automatique du discours et réception de Harris. Enfin, le dernier chapitre (chapitre 9) est consacré au « tournant empiriste de l'automatisation-mathématisation. Grands corpus, langages restreints, sous-langages ». Inscrit dans la continuité, ce tournant prend ses sources dans la tradition britannique, et fait émerger de nouveaux objets pour le TAL. Il a permis de renouveler un débat entre empirisme et chomskysme entrepris dans les années 1960.

Un point de méthode : le corpus des textes concernant les expérimentations de la TA (1948-1966) est sinon fini, du moins aisément répertoriable. Le nombre relativement restreint des textes publiés rend leur recensement possible : premiers ouvrages collectifs, revues (*Machine Translation*, *La traduction automatique* et leurs successeurs). Par ailleurs une de nos tâches est de recueillir la « littérature grise » et les archives personnelles auprès des institutions et des pionniers du domaine. Leur classement et archivage constituent une composante essen-

tielle d'une telle recherche. Nous tenons à la mener à bien et espérons que de futurs jeunes chercheurs pourront utiliser ce fonds d'archives et poursuivre ainsi les recherches dans ce domaine[6].

Pour cette recherche, j'ai largement utilisé les notices et les textes du *Corpus de textes linguistiques fondamentaux* (*CTLF*), et je tiens à exprimer mes remerciements à Bernard Colombat et Arnaud Pelfrêne qui m'y ont donné accès.

6 La constitution d'un fonds d'archives et de documentation sur l'histoire de la traduction automatique et du traitement automatique du langage (1954-1975) fait l'objet d'une convention signée en 2006 entre l'ATALA (Association pour le traitement automatique des langues), le CNRS, l'université Paris Diderot et l'ENS Lyon. Ce projet est mené dans le cadre de l'UMR7597 (Histoire des théories linguistiques) avec la collaboration d'Elisabeth Lazcano (documentaliste). Dans la bibliographie les documents faisant partie de ce fonds d'archives sont référencés comme [archives *Histoire du traitement automatique des langues* HTAL].