La constitution du TAL

Marcel Cori, Jacqueline Léon

La constitution du TAL
Étude historique des dénominations et des concepts

Marcel Cori* — Jacqueline Léon**

* *Laboratoire Modèles, Dynamiques, Corpus*
CNRS/Université Paris X
200 avenue de la République, F-92001 Nanterre cedex
mcori@u-paris10.fr
** *Laboratoire d'Histoire des Théories Linguistiques*
UMR 7597 CNRS/ Université Paris 7/ ENS Lettres et Sciences Humaines
Case 7034, 2 place Jussieu, F-75251 Paris cedex 05
jacqueline.leon@linguist.jussieu.fr

RÉSUMÉ. *Pour désigner le champ d'investigations et d'applications à l'intersection de la linguistique, des mathématiques, de l'informatique et des sciences cognitives hérité des expériences pionnières en traduction automatique, plusieurs termes sont ou ont été en concurrence,* Computational Linguistics *ou* Natural Language Processing *dans le domaine anglo-américain, Traitement automatique des langues, Informatique linguistique ou Linguistique informatique en France. Cet article se propose, en retraçant le parcours historique de ces dénominations, de montrer que le flottement sur les termes est symptomatique des tensions à l'œuvre dans le domaine, sur le plan des enjeux institutionnels, économiques, théoriques et conceptuels.*

ABSTRACT. Several terms have been in competition as names for the theoretical and applied discipline that lies in the intersection of linguistics, mathematics, computer sciences and cognitive sciences and which developed out of early experiments in Machine Translation. They include Computational Linguistics and Natural Language Processing in English, and Traitement automatique des langues, Informatique linguistique *and* Linguistique informatique *in French. This paper traces the history of these terms and considers whether the terminological variation may be a symptom of the conflicts at work in the field, concerning the institutional, economical, theoretical and conceptual issues.*

MOTS-CLÉS : histoire, épistémologie, traduction automatique, TAL, linguistique informatique, linguistique appliquée, linguistique mathématique

KEYWORDS History, Epistemology, Machine Translation, Computational Linguistics, Natural Language Processing

TAL. Volume 43 – n° 3/2002, pages 21 à 55

1. Introduction

Le Traitement automatique des langues (ou du langage), connu sous l'acronyme TAL, a une réalité indiscutable. En témoignent le titre de la revue dans laquelle est publié cet article, ainsi que l'usage répété du sigle dans les intitulés de cursus et de colloques[1]. Mais la question de ce que recouvre ce terme, et même de savoir s'il fait référence à une réalité bien délimitée, se pose si on déplace les frontières géographiques et temporelles de notre réflexion. Ainsi, on ne sait pas très bien quelle est la signification de ce terme quand on cherche à le traduire en anglais : on a alors le choix entre *Computational Linguistics* (infra *CL*) et *Natural Language Processing* (infra *NLP*). Par ailleurs, il y a une dizaine d'années en France, on aurait parlé plutôt de Linguistique informatique ou d'Informatique linguistique que de TAL.

Ce flottement dans la dénomination est un symptôme de la difficulté de déterminer si le TAL désigne un domaine scientifique, une technologie ou une communauté de chercheurs et d'ingénieurs. C'est également un symptôme de la difficulté du TAL, sous l'emprise simultanée de contraintes technologiques, pratiques et sociales, de se développer en tant que nouveau champ spécifique et de se situer par rapport aux quatre principaux pôles disciplinaires autour duquel il gravite :

- la linguistique ;

- l'informatique ;

- les mathématiques (sous la forme de l'algèbre, de la logique ainsi que des statistiques) ;

- l'intelligence artificielle, la psychologie expérimentale ou, plus récemment, les sciences cognitives qui se préoccupent de définir des modèles de l'esprit et du langage.

Afin de donner des éléments permettant d'aller vers une caractérisation précise de ce qu'est le TAL, il nous faut revenir aux origines. Nous appréhenderons en premier lieu la constitution du domaine au travers des termes qui l'ont désigné. Les changements de termes ont en effet la capacité de dévoiler les perspectives théoriques et pratiques mises en jeu[2]. Nous allons ensuite étudier comment les acteurs du TAL ont défini le domaine dans lequel ils travaillaient, en nous appuyant sur des textes qui ont tenté de mettre en avant une synthèse théorique en se donnant pour tâche d'expliciter les enjeux, textes que l'on peut qualifier d'épistémologiques, mais qui ont bien souvent une visée politique.

---

[1] On trouve aussi Traitement automatique du langage naturel/des langues naturelles (TALN).

[2] La terminologie et la traduction des termes étant un des éléments principaux de notre argumentation, nous conserverons dans notre texte les termes anglais en anglais.

Nous abordons la description chronologique de l'apparition des différents termes qui se sont succédé en traitant de manière séparée les domaines américain et français qui, bien qu'étroitement liés, présentent des différences qui font apparaître les points de fracture[3]. L'origine du TAL peut être située aux États-Unis, où sont nées les premières idées de traduction automatique (TA infra) associées à l'apparition des machines électroniques. Ces idées ont été exposées dans le *Memorandum* de Weaver de 1949 intitulé « Translation » et diffusé auprès des quelque deux cents universitaires ou commanditaires susceptibles d'être intéressés. Les Français ont commencé une dizaine d'années après les Américains, au moment où l'idée d'une traduction entièrement automatisée est déjà publiquement mise en doute, et dans un contexte culturel et scientifique très différent : importance moindre accordée à l'informatique, attitude spécifique des linguistes et des institutions par rapport à la TA. Aux États-Unis, les impasses de la traduction automatique ont fait le lit de la *Computational Linguistics*, puis, dans un second temps, du *Natural Language Processing*. En France, c'est le terme *Traitement automatique des langues* qui s'est imposé, mais cela a pris beaucoup de temps.

Nous nous interrogeons ensuite sur le contenu sous-jacent aux termes mis en avant, en essayant tout d'abord de voir comment le domaine a pu se situer par rapport aux quatre pôles autour desquels il gravite, puis en examinant quel contenu effectif a été donné au TAL par ses acteurs. Nous montrons enfin en quoi le tournant déclaratif a entraîné une redéfinition du domaine.

2. Aux États-Unis, la traduction automatique et ses suites

2.1 La naissance d'un champ d'investigation

Aux États-Unis, les termes *Mechanical Translation* et *Machine Translation* sont utilisés de façon indifférente et semblent s'être fixés assez tôt. Warren Weaver dans son *Memorandum* de 1949 n'utilise pas le terme de *Machine Translation* à proprement parler. Il parle de *computer translation*, et de *mechanization of the translation problem*. Le terme *Mechanical translation* apparaît, quant à lui, sous la plume de Norbert Wiener[4], comme problème, « the problem of mechanical translation », en relation avec « the mechanization of speech and the mechanization of language ». Il s'agit donc à l'origine, pour ces deux cybernéticiens, d'un projet dont il faut étudier la faisabilité.

[3] Les Soviétiques ont une approche tout à fait particulière qu'il faudrait traiter à part (cf. sur le domaine Archaimbault, Léon, 1997). Quant aux Britanniques, leur histoire se confond au début, notamment avec les premières expériences de Booth et Richens, avec l'histoire américaine ; ce qui n'exclut pas des approches tout à fait originales comme celles du *Cambridge Language Research Unit* (cf. Léon 2000).

[4] Lettre du 30 avril 1947 à Weaver, citée dans (Weaver, 1949).

Le texte de Weaver aborde déjà la plupart des questions théoriques et pratiques de l'automatisation du traitement du langage, dont celles de l'automatisation de la traduction : structure logique du langage et des langues, grammaires universelles et universaux. Les aspects pratiques ne sont pas négligés. Ce cybernéticien-cryptographe propose de résoudre les ambiguïtés syntaxiques et sémantiques en utilisant la redondance du langage écrit dans le cadre de la théorie de l'information (Shannon et Weaver, 1948) ; par ailleurs, Weaver soutient une position modeste en écartant d'emblée toute possibilité de traduction littéraire par la machine et en faisant l'hypothèse que la traduction parfaite est impossible mais qu'il peut y avoir des degrés inférieurs de traduction utilisables par les scientifiques. On notera que cette acceptation d'une fiabilité relative des résultats nous renvoie d'ores et déjà à l'une des difficultés propres au TAL.

La *Mechanical Translation* ne devient que plus tard un thème de recherche, lorsqu'une partie des deux cents scientifiques auxquels Weaver avait envoyé son *Memorandum* s'attaquent au problème. Dès janvier 1950, Erwin Reifler, directeur du groupe de TA de l'Université de Washington à Seattle, intitule une série de rapports de recherche *Studies in Mechanical Translation.* C'est d'ailleurs au sein du groupe de Washington que le terme *Mechanical translation* va être associé à l'idée de machine à traduire, c'est-à-dire à la production en série de traductions automatisées, rentables sur le plan industriel. L'idée d'un dispositif autonome de traduction suppose la réalisation d'une machine entièrement consacrée à la traduction qui n'aurait pas les mêmes caractéristiques physiques que les calculatrices électroniques spécialisées dans le calcul numérique. C'est ainsi que la machine Mark I (*the US Air Force Automatic Language Translator Mark I*) de l'Université de Washington, fondée sur la mémoire photoscopique de Gilbert King ne correspond pas à l'architecture Von Neumann.

Pour désigner les premières recherches, on utilise aussi le terme *Machine Translation*. Weaver utilise les deux termes dans sa préface au recueil de 1955 *Machine Translation of Languages*, intitulée *The new tower* en référence à la tour de Babel. Ce recueil rassemble les contributions du premier colloque de TA, intitulé *Conference on Mechanical Translation*, qui a lieu en juin 1952 au MIT sous la direction de Yehoshua Bar-Hillel. Enfin la revue fondée par Victor Yngve en 1954 s'intitule *Mechanical translation – devoted to the translation of languages by the aid of machines*.

Bar-Hillel utilise aussi les deux termes alternativement dans quatre articles parus sur le sujet en 1953. A noter que dans son premier article sur la TA (1953a), rédigé en 1951, il pose les questions fondamentales de la TA, mais de façon un peu différente de Weaver. Bien que convaincu comme lui de l'impossibilité d'une traduction parfaite, qu'il nomme FAHQT (*Full Automatic High Quality Translation*), il ne préconise pas de traductions de niveaux inférieurs[5], mais plutôt une traduction assistée par ordinateur ; d'où la nécessité d'expérimenter avec une

[5] On trouve dans ce texte une des premières discussions sur l'évaluation des systèmes de TA.

assistance humaine de pré-édition et de post-édition. Il s'interroge sur le rôle des statistiques, la faisabilité d'une analyse automatique morphologique et l'utilité des domaines restreints (vocabulaires de spécialité). Comme Weaver, il s'interroge sur les fondations logiques du langage et préconise la construction de diverses grammaires universelles et leur évaluation. Son apport le plus important est la nécessité de construire une analyse syntaxique opérationnelle qu'il développera dans son article « A Quasi-Arithmetic Notation for Syntactic Description » (1953b). Par ailleurs, l'idée que, pour traduire, la machine doit appréhender les questions de sens, incompatibles avec les options structuralistes, apparaît dans un autre article de la même année (1953c) où il déclare p. 217 : « the task of instructing a machine how to translate from one language it does not and will not understand into another language it does not and will not understand presents a real challenge for structural linguists, in that their thesis that language can be exhaustively described in non-referential terms undergoes here an experimentum crucis. If, in a translation program, some step has to be taken which directly or indirectly depends upon the machine's ability to understand the text on which it operates, then the machine will simply be unable to make this step, and the whole operation will come to a full stop. I have in mind present day machines that do not possess a semantic organ. The situation will change in the not too distant future. » Cette question sera développée notamment dans son rapport de 1960 au travers de son fameux exemple *the box was in the pen* :

« The linguistic context from which this sentence is taken is, say, the following: Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

Assume, for simplicity's sake, that pen in English has only the following two meanings: (1) a certain writing ustensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word pen in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this "automatically". » (Bar-Hillel (1960), Appendix III A demonstration of the Nonfeasibility of Fully Automatic High Quality Translation, p.158-159).

Est ainsi abordée dès cette époque la question de la compréhension du langage et du rapport du TAL à l'Intelligence artificielle.

En 1960, bien que le terme *Machine translation* semble unanimement adopté[6], il recouvre en fait des approches et des méthodes extrêmement différentes. Même si l'objectif avoué – subventions obligent – reste la traduction automatisée de textes en série à visée industrielle, les projets n'ont parfois en commun que l'automatisation du traitement du langage ou des langues à l'aide de calculatrices électroniques :

[6] Seule la revue *Mechanical Translation* conservera le terme *Mechanical* jusqu'au bout, c'est-à-dire jusqu'en 1974 quand Yngve en quittera la direction. Yngve tient d'ailleurs au terme *mechanical* et refuse de le troquer pour *machine* (communication personnelle).

dictionnaires électroniques fondés sur différentes approches de l'analyse morphologique (Oettinger, 1960), analyseurs syntaxiques, mise au point de langues intermédiaires sémantiques (Masterman, 1957, Mel'cuk, 1960), ébauches de mémoires de traduction (Koutsoudas et Humecky, 1957), résolution d'ambiguïtés à l'aide de méthodes statistiques (Kaplan, 1950), traitement des unités lexicales composées (Reifler, 1955), etc. Les approches, théoriques, formelles ou délibérément empiriques, sont diverses, et les conceptions linguistiques multiples. Toutefois, stimulées par ce brassage parfois désordonné, on peut avancer que la plupart des idées et des méthodes fondamentales du TAL apparaissent pendant ces dix premières années (1948-1958). Certaines d'entre elles disparaîtront au milieu des années 1960, au profit exclusif de l'analyse syntaxique, pour réapparaître quelques décennies plus tard.

On notera que les acteurs de la TA aux États-Unis sont des ingénieurs, des mathématiciens appliqués, des philosophes, des spécialistes de langues naturelles. Les linguistes structuralistes sont peu présents. Même si Harris et Hockett se sont associés à des anthropologues linguistes comme Voegelin ou Garvin pour publier en 1954 un numéro spécial sur la traduction dans l'*International Journal of American Linguistics*, aucun des auteurs du numéro, à l'exception de Paul Garvin, ne poursuivra de travaux en TA[7]. Quant à Chomsky, bien qu'embauché en1955 au LRE par Yngve sur un projet de TA, avec une subvention de la National Science Foundation, en compagnie de trois autres linguistes Joseph R. Applegate, Fred Lukoff et Betty Shefts (cf. *Mechanical Translation* vol. 2, n°1, 1955), il n'a jamais travaillé sur un projet de TA.

Si l'on situe le TAL à la croisée d'objectifs parfois difficilement compatibles, tels que fournir des outils directement dépendants de la demande sociale et dont la rentabilité industrielle constitue un critère essentiel d'évaluation d'une part, proposer des dispositifs dynamiques de représentation des connaissances linguistiques d'autre part, et enfin constituer des bancs d'essai pour la validation de théories linguistiques[8], force est de constater que, dès les premiers projets de TA, ces trois perspectives sont souvent intriquées. Cela tient au fait que, dans la conception des dispositifs de TA, interagissent différentes conceptions de la traduction et de l'objet « texte à traduire », les théories linguistiques en cours, la demande sociale et le développement technologique.

On connaît l'issue fatale des projets de TA suite au rapport de Bar-Hillel (1960) et au rapport de l'ALPAC (*Automatic Language Processing Advisory Committee*), publié en 1966. Les raisons de la suppression des subventions et de la mise à l'index de la TA pour plus de deux décennies aux États-Unis, ont été abondamment commentées (cf. Hutchins 1996, entre autres)[9]. Ce qui est clairement posé par

[7] Voir sur ce point Léon (2001).

[8] Ainsi Kay (1982) conçoit les systèmes de TA et les analyseurs syntaxiques comme des méthodes d'évaluation des modèles linguistiques par la technologie.

[9] Il ne faut pas oublier qu'en optant pour une démarche délibérément empirique et en rejetant toute réflexion théorique, beaucoup de recherches avaient été menées isolément, chacun

l'ALPAC, c'est que la *Machine Translation* au sens de « going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing » n'est pas possible, et qu'il faut encourager la *machine-aided translation* et un recentrage des travaux de recherche sur l'analyse syntaxique automatique, c'est-à-dire sur l'interaction entre langages formels logico-mathématiques, analyse grammaticale et programmation. C'est ce qui constituera la *Computational Linguistics*.

2.2 La Computational Linguistics

La *Computational Linguistics*, que ce soit sur le plan théorique ou institutionnel, ne date pas du rapport de l'ALPAC ; elle lui est bien antérieure. C'est le rapport Bar-Hillel publié en 1960 qui a donné l'alarme et a conduit certains des acteurs de la TA à déplacer les enjeux de l'automatisation du langage. En 1961, conscients du danger que la TA courait, ceux-ci ont tenté de se recentrer sur les aspects théoriques et de se rassembler autour d'une association professionnelle.

Quant aux premiers analyseurs syntaxiques, qui sont au cœur de la « nouvelle linguistique » que constitue la *Computational Linguistics*, ils datent du début des années 1950. Bar-Hillel (1953a) fait l'hypothèse qu'on peut construire une machine capable de déterminer la structure de toute phrase d'une langue source pourvu que la syntaxe de cette langue soit présentée sous forme opérationnelle à la machine. Dans sa grammaire catégorielle, mettant en correspondance la méthode de Harris et la notation de Ajdukiewicz, il envisage une méthode automatique pour tester la connexité syntaxique d'une séquence et trouver les constituants immédiats de toute séquence syntaxiquement connexe.

En 1955, Yngve (1955) élabore une procédure traitant les phrases réduites à une suite de classes de mots qui comprennent l'information grammaticale et syntaxique de la phrase. Il s'agit d'une procédure de reconnaissance ascendante, gauche-droite, chargée de construire la structure syntaxique niveau par niveau. Par la suite, Yngve (1960) élabore un modèle prédictif d'analyse syntaxique fondé sur une grammaire syntagmatique, et sur des hypothèses psycholinguistiques sur la mémoire à court terme pour fixer la profondeur de l'arbre de représentation des phrases. Alors qu'en 1955 il utilise les automates à états finis, il se réfère en 1960 au modèle de Chomsky (1957) mais en rejetant les transformations. Il met ainsi au point la méthode dite de transfert pour la TA. Parallèlement, en 1957-58, Yngve met au point un langage

---

travaillant pour soi et refaisant le même chemin. Les objectifs, essentiellement industriels pour certains, avaient exacerbé une concurrence entre équipes peu enclines à la communication de résultats. Face à une demande militaire de traductions du russe dans un contexte de guerre froide, et à une demande croissante en traductions technologiques et scientifiques, l'absence de publications et les annonces de presse triomphalistes avaient même pu faire soupçonner certaines équipes de véritables escroqueries (Bar-Hillel, 1960), (Mounin, 1964).

spécialisé, COMIT, premier langage de traitement de symboles non numériques (chaînes de caractères) fondé sur le *pattern matching*.

Au début des années 1960, il existe déjà de nombreux analyseurs syntaxiques développés à partir de différents modèles de grammaires, comme par exemple la grammaire de dépendance de Tesnière (Hays, 1962) ou la grammaire stratificationnelle de Lamb (1962).

Sur le plan institutionnel, c'est en février 1961 que la *computational linguistics* voit le jour, grâce aux conférences hebdomadaires organisées par David G. Hays à la Rand Corporation de Los Angeles. Ces conférences seront reprises comme communications lors de la *First International Conference on Machine Translation of Languages and Applied Language Analysis* de Teddington en septembre 1961. Y participent des linguistes ou informaticiens, acteurs de la TA, comme Paul Garvin, Sydney M. Lamb, Kenneth E. Harper, Charles Hockett, Martin Kay (appartenant alors au groupe britannique *Cambridge Language Research Unit)*. Bernard Vauquois y propose un exposé sur la « linguistique appliquée en France ».

C'est lors de cette conférence, où étaient présents tous les groupes de recherche en TA subventionnés par le gouvernement américain, qu'est décidée la création d'une société américaine pour la traduction automatique, appelée à défendre les intérêts scientifiques et professionnels des chercheurs dans ce domaine. Cette société sera créée en juin 1962 sous la présidence de V. Yngve et la vice-présidence de Hays. Elle prendra le nom d'*Association for Machine Translation and Computational Linguistics (AMTCL)* et, en 1973, ne s'appellera plus que *Association for Computational Linguistics* (ACL). L'association se dote d'un bulletin, *The Finite String*, et organise un colloque tous les deux ans. Le premier a lieu en août 1963 à Denver, Colorado. La revue *Mechanical Translation* s'appelle, à partir de 1965 (et jusqu'à 1974 où Yngve quitte sa direction) *Mechanical Translation and Computational Linguistic*s. De 1974 à 1983, elle s'appellera *American Journal of Computational Linguistics*, puis à partir de 1984 *Computational Linguistics*[10].

Institutionnellement, la *Computational Linguistics* est née. Le premier colloque intitulé *International Conference on Computational Linguistics* a lieu à New York en mai 1965, regroupant 150 participants. Il est organisé par plusieurs associations (l'AMCTL, l'ATALA, des associations scandinave, japonaise et sud-américaine) qui seront à l'origine de la création d'une Fédération Internationale, l'*International Committee on Computational Linguistics* (ICCL)[11].

En 1966, l'ALPAC légitime définitivement la nouvelle discipline, « the new linguistics ». Le rapport recommande de subventionner les recherches sur les

[10] D'après Martin Kay (2000), c'est David Hays, membre de l'ALPAC dès 1964, et au courant des probables coupures de subventions qui menaçaient la TA, qui a convaincu Yngve de changer le nom de la revue.

[11] Ce colloque inaugure la série des célèbres *Coling*, qui seront désormais organisés par l'ICCL.

méthodes informatisées du traitement du langage, sur les outils d'aide à la découverte, à la généralisation et à la validation pour le linguiste, enfin sur les méthodes permettant aux linguistes de tester les théories grammaticales et sémantiques considérant que « The tools of computational linguistics are considerably less costly than the multibillion-volt accelerators of particle physics. »

Parallèlement à ce recentrage théorique et institutionnel, rendu nécessaire par l'arrêt brutal des subventions, on constate cependant que, sous couvert précisément de la *Computational Linguistics*, sont menés toutes sortes de travaux associant informatique et traitement du langage, et faisant une large place aux applications. De fait, dans sa profession de foi, publiée dans le premier numéro du Bulletin de l'AMTCL, *The Finite String*, qui paraît en 1964, David Hays, président de l'association depuis 1963, revendique l'existence d'un domaine de savoir spécifique défini par l'interaction entre linguistique et informatique, et une intégrité disciplinaire pour toute application linguistique de l'informatique, sans exclusion.

Cette revendication d'un champ disciplinaire unique à travers la diversité des applications est manifeste dans les thèmes des premiers colloques organisés par l'AMTCL, la revue *Machine Translation and Computational Linguistics* et le bulletin de l'association *The Finite String*.

Ainsi, l'appel à communications de la première conférence, qui a lieu à New-York en 1965, stipule que la *Computational Linguistics* inclut toutes les applications de l'ordinateur à des fins de traitement des langues naturelles ou artificielles. On y encourage tout de même les aspects plus théoriques tels que les travaux sur les fondements mathématiques ou sur les recherches en linguistique effectuées avec l'aide de l'ordinateur. Le compte-rendu, publié dans le numéro double de *TA informations* de 1965, met l'accent sur la diversité du domaine, tout en s'interrogeant sur cette nouvelle science que constituerait la *computational linguistics :*

« Deux courants principaux paraissent se dégager :

1) les recherches purement pragmatiques avec acceptation d'une marge d'erreur devant aboutir à des résultats concrets dans des délais limités, mais dont l'exploitation effective est essentiellement liée à l'introduction rapide et peu onéreuse des données en machine.

2) Des travaux relevant de la recherche fondamentale, soit à tendance linguistique, soit à tendance mathématisante.

Ces deux courants se trouveraient réunis dans une science nouvelle que les Américains appellent *computational linguistics* et qui paraît se situer à l'intersection de la linguistique proprement dite, l'appareil formel offert par les sciences logico-mathématiques, et certaines contraintes, disons 'certaines règles de jeu' suggérées par les machines destinées au traitement automatique de l'information, de sorte que l'étage ultime d'une recherche aboutirait à un modèle algorithmique. Cette jeune science ne doit être confondue ni avec la linguistique mathématique susceptible de

se développer indépendamment de l'existence même des machines concrètes, ni avec la linguistique quantitative, principalement axée sur les méthodes statistiques. » (*TA Informations* 1965, p. 62).

Les deux colloques suivants présentent la même diversité. Ainsi le second, intitulé « conférence internationale du traitement automatique des langues », et organisé par le CETA (Centre d'Étude pour la Traduction Automatique) en août 1967 à Grenoble, propose trois grands thèmes : analyse automatique des langues naturelles, analyse statistique et sémantique des données linguistiques, théorie algébrique des langages.

On retrouve la même hétérogénéité dans le *Finite String*, bulletin de l'AMTCL, publié de 1964 à 1973. Sous la rubrique générale *Recent publications in Computational Linguistics*, sont recensés de façon régulière les travaux dans les domaines étiquetés de la façon suivante : *Linguistics and Computation, Mechanical translation*[12], *Quantitative Linguistics, Languages, Computer Science, Information Science, Related fields and applications.* On constate donc que la rubrique *Computational Linguistics* regroupe tous les domaines de la linguistique, y compris des domaines non concernés par le traitement par ordinateur comme la sous-rubrique *Languages*, où figurent des descriptions de langues naturelles. Ce qu'on peut appeler le « noyau dur » de la CL, c'est-à-dire les travaux utilisant à la fois les langages formels et l'ordinateur, est plus ou moins représenté par la sous-rubrique *Linguistics and Computation :* grammaires formelles, analyse syntaxique, génération, avec également des travaux sur les aspects psycholinguistiques de la syntaxe et de la sémantique, et les travaux en linguistique algébrique des chercheurs des pays de l'Est.

2.3 Le Natural Language Processing

Alors que la *Computational Linguistics* tente de se constituer comme domaine autonome, avec ses ancrages théoriques (cf. 4.2) et institutionnels, les travaux associant linguistique, langages formels et informatique se reconnaissent aussi, dans la même période, sous le terme de *automatic* ou *automated language processing*. *Automatic*, comme en témoignent les titres d'articles recensés par le *Finite String*, est essentiellement associé à *linguistic* ou *syntactic analysis*. On trouve ainsi *automated/automatic linguistic/language analysis, automated/automatic syntactic analysis, automated/automatic sentence analysis.* Il est clair que l'essor de ces termes marque la prédominance absolue de l'analyse syntaxique automatique comme seule méthode de TA pendant les années 1960.

Automatic permet aussi de marquer la distance avec la TA sur laquelle pèsent nombre de soupçons, inaugurés par le rapport de Bar-Hillel, dont le titre « The

[12] Cette sous-rubrique *Mechanical Translation* ne contient, dans cette période, que quelques articles canadiens, allemands, français ou japonais, et très peu d'articles américains.

Present Status of Automatic Translation of Languages » comporte le terme *automatic translation*[13] et non celui de *machine translation. Automatic* est plus abstrait et plus théorique que *mechanical* et réfère davantage à l'automatisation de modèles ou bien à une implication partielle de l'ordinateur dans le traitement du langage. Le terme *automatic* est d'ailleurs repris par les promoteurs de la CL. Ainsi, Hays (1962), se référant explicitement à Bar-Hillel (1960), forge le terme *Automatic Language-Data Processing* (ALDP)[14] où l'on peut voir une des premières occurrences du terme *Language Processing* qui gagnera du terrain dans les années 1970.

L'ALPAC, en se constituant en 1964 comme *Automatic Language Processing Advisory Committee*, contribue à promouvoir le terme *Automatic Language Processing*. C'est en 1964 également que paraît l'ouvrage de Paul Garvin « Natural Language and the Computer », dont une des trois parties s'intitule *Language Data Processing*, et que trois cours de *Language Data Processing* sont proposés, l'un à Harvard dans une école d'été organisée par Oettinger, le second à l'Université d'Indiana, sous la direction de Garvin, et le troisième à la Rand avec Hays[15].

Le début des années 1970, c'est aussi la rencontre entre automatisation du langage et Intelligence artificielle[16]. Celle-ci, apparue avec la première cybernétique des *Macy Conferences* en 1946, est dominée à partir de 1956 par le modèle cognitiviste et la résolution de problèmes. Elle s'essouffle à la fin des années 1960 dans les « solutions ad-hoc de problèmes habilement choisis pour donner l'illusion d'une activité intellectuelle complexe et complète » (Dreyfus, 1972) et doit reconnaître la nécessité de construire des systèmes de représentation des connaissances, ce que préconisait, pour la TA, Bar-Hillel dans son rapport de 1960.

---

[13] A noter qu'Anthony G. Oettinger, qui fera partie de l'ALPAC, publie, également en 1960, un ouvrage intitulé *Automatic Language Translation*. Cet ouvrage est pionnier, puisqu'il s'agit du premier dictionnaire électronique russe-anglais et du premier PhD dans le domaine, entrepris en 1954.

[14] Dans ce texte de 1962, l'opinion de Hays sur la TA est, à vrai dire, tout à fait surprenante. Il considère que la TA constitue une des parties les plus simples de l'*Automatic Language-Data Processing*, dans la mesure où elle ne requiert pas de traitement du sens, contrairement par exemple à la documentation automatique : « MT is not the whole, of ALDP. It is, however, the earliest species to develop and probably the simplest. Reasonably good translations seem to require little or nothing more than knowledge of grammar, whereas indexing and abstracting seem to require much more, and the increment must be drawn from the terra incognita of semantics. » (article cité p. 396). En 1972, Hays ajoutera que ce n'est pas en trouvant le bon modèle formel en linguistique qu'on va résoudre les problèmes de TA, mais en fournissant du travail solide fondé sur une linguistique simplifiée (Melby, 1992).

[15] D'autres termes apparaissent, qui n'auront pas de lendemain : de1962 à 1964, Sidney Lamb organise, à Berkeley, des colloques de *Mechanolinguistics.* D'après Martin Kay (communication personnelle), le terme *Computational Linguistics* a été préféré à celui de *Mechanolinguistics* lors d'une réunion regroupant Hays, Kay, Lamb et Pendergraft, au lendemain de la publication du rapport de l'ALPAC, pour désigner les recherches théoriques et les distinguer clairement de l'ALDP consacré aux applications industrielles.

[16] Le premier numéro de la revue *Artificial Intelligence* paraît au printemps 1970.

C'est dans le sillage de l'IA qu'on voit apparaître le *Natural Language Understanding*.

Le terme *Natural Language Processing* (NLP) s'installe dans les années 1980[17] et semble bien établi dans les années 1990. En 2001, l'ACL définit ses objectifs comme devant promouvoir la recherche en NLP.

Pour observer l'installation du terme NLP, il est intéressant d'examiner l'intitulé des premiers cursus universitaires. On a vu que les premiers cours ou séminaires de *Computational Linguistics* ont été créés en 1964 et avaient pour nom *(Automatic) Language (Data) Processing*. En 1967-68, on voit apparaître des cours nommés *Computational Linguistics* organisés par Garvin, Hays et Kay. Au début des années 1980, c'est le terme *Computational Linguistics* qui reste le terme général pour désigner la discipline dans les universités. Ainsi le premier annuaire général des cursus, publié dans la revue *Computational Linguistics* en 1984, s'appelle le *Directory of graduate programs in Computational Linguistics.* Pourtant le terme *Natural Language Processing* est aussi utilisé, et ce sont les départements pluridisciplinaires, informatique et linguistique, qui utilisent conjointement les deux termes. La *Computational Linguistics* y est alors davantage orientée vers l'analyse syntaxique, ou l'analyse sémantique (voire discursive ou pragmatique) et/ou le *Natural Language Understanding*, c'est-à-dire vers certains éléments de l'Intelligence artificielle (IA) comme les *frames*. Les départements qui utilisent le terme NLP seul sont massivement les départements d'informatique, et ce qu'ils appellent NLP est fortement orienté vers l'IA.

Dans les années 1990, la *Computational Linguistics* reste la discipline de référence : c'est le nom de la principale revue, de l'association et des cursus universitaires consacrés au domaine. Pourtant, l'activité elle-même est de plus en plus désignée comme NLP. Ainsi, dans l'*Encyclopedia of Language and Linguistics* dirigée par Asher (1994) il n'y a pas d'entrée *Computational Linguistics*. L'index renvoie à l'entrée *Natural Language Processing*, dont les auteurs Mellish, Hirst et Hayes (entre autres) insistent sur la séparation de plus en plus nette, datant du début des années 1990, entre la *Computational Linguistics*, plus théorique, et le NLP, à visée plus pratique. Pour Hayes, en effet, une application en NLP requiert certaines conditions : un problème réel à résoudre, répondant à une demande sociale ; et une solution viable, en terme de fiabilité, robustesse, rapidité et coûts[18]. Ce qui a conduit, dit-il, a une dichotomie de plus en plus marquée entre application et recherche en *Computational Linguistics*, de telle sorte que celle-ci, orientée exclusivement vers des recherches théoriques, ne se donne plus du tout comme objectif la réalisation

[17] On peut trouver des occurrences du terme plus tôt. C'est en effet en 1966 que l'AMTCL finance une session spéciale intitulée « Natural Language Processing » à la Spring Joint Computer Conference à Boston.

[18] Toutefois, l'auteur précise que ces conditions sont de plus en plus difficiles à satisfaire. Même constat pour la TA : Kay (1992) déplore le peu de progrès accomplis depuis les années 1960, estimant que les systèmes se sont beaucoup améliorés sur le plan de la vitesse mais peu sur le plan de la qualité.

d'applications industrielles, cette tâche restant dévolue au NLP. On peut tout de même se demander si la disparition d'une entrée spécifique *Computational Linguistics* n'est pas la marque d'un désintérêt croissant des acteurs du domaine pour ses aspects théoriques.

On retrouve une démarche analogue chez les promoteurs du *Natural Language Engineering* (Cunningham, 1999), terme forgé au début des années 1990 dans le cadre de programmes européens. En définissant la CL comme partie de la linguistique théorique, partageant les mêmes objectifs et utilisant l'informatique comme outil d'investigation et de modélisation, on l'écarte du domaine. Et c'est le NLP qui se voit octroyer les objectifs de la CL, tels que la recherche sur les algorithmes d'analyse syntaxique ou de génération. Ce jeu de taquin a pour but de doter les applications industrielles d'une véritable légitimité scientifique et ceci grâce à une nouvelle dénomination, le *Natural Language Engineering*.

Pour un historien du TAL, cette tension entre théorie et pratique liée à une demande sociale paraît familière. N'est-ce pas la discordance entre demande sociale, moyens investis et médiocrité des résultats qui a conduit l'ALPAC à promouvoir la *Computational Linguistics* en dégageant, au détriment des objectifs empiriques de la TA, les aspects théoriques de la mécanisation et de la formalisation du traitement des langues ? N'assisterait-on pas actuellement à la démarche inverse qui tend à privilégier, par l'utilisation dominante du terme NLP, les applications industrielles au détriment des investigations théoriques ? Ces mouvements de balancier ne témoignent-ils pas d'une difficulté intrinsèque de la *Computational Linguistics* à se constituer en véritable discipline scientifique ?

3. La constitution du TAL en France

En France, l'automatisation du langage a d'abord concerné la mécanisation du vocabulaire, avant la traduction automatique qui n'a démarré qu'en 1959, soit plus d'une dizaine d'années après les États-Unis et la Grande-Bretagne[19]. Cette prééminence des études mécanisées du vocabulaire aura des conséquences importantes sur l'histoire du TAL en France.

L'essor de la lexicologie en France a commencé avec Mario Roques et son *Inventaire Général de la Langue Française* créé en 1936 (Chevalier et Encrevé, 1984). Sa mécanisation a été encouragée lors de différents colloques à la fin des années 1950, et a été marquée par la création, en 1959 à Besançon, du Laboratoire d'analyse lexicologique, des *Cahiers de Lexicologie* et des *Études de Linguistique Appliquée*, tous trois sous la direction de Bernard Quemada.

Concernant la TA, c'est en 1959 qu'a été créée l'ATALA à la suite d'un petit groupe international d'études sur la traduction automatique animé par Émile

[19] Sur les raisons du retard français, le début de la mécanisation du vocabulaire et les débuts de la TA en France, voir Léon (1998).

Delavenay à l'UNESCO[20]. A la fin de la même année, le CNRS, par une convention avec les services concernés du ministère de la Défense, crée, à l'Institut Blaise Pascal[21], le CETA (Centre d'Étude pour la Traduction Automatique) comprenant deux sections, l'une parisienne dirigée par Aimé Sestier, l'autre grenobloise dirigée par Bernard Vauquois.

Avec ce décalage de dix ans et cette différence d'ancrage culturel, les différents termes utilisés par les Français pour désigner le domaine illustrent en les amplifiant les lignes de tension qui ont divisé les Américains et le milieu international en général. En France, le consensus sur le terme *TAL* date seulement du début des années 1990. Auparavant, on assiste à l'apparition de divers termes concurrents qui se disputent le domaine. Certains, c'est le cas du terme *TAL* lui-même, disparaissent pour réapparaître plus de deux décennies plus tard.

En France, les opinions ont souvent divergé quant aux thèmes qui devaient faire partie du TAL. On peut avancer que ces divergences, tout en reflétant les lignes de tension et les partages de territoires dans l'Université et au CNRS, ont rendu perceptible et incontournable la question de la légitimité du domaine d'investigation.

3.1 Machine à traduire et traduction automatique

Le choix des termes chargés de désigner l'automatisation de la traduction des langues, machine à traduire ou traduction automatique, montre les traces laissées par les dix ans d'histoire qui ont précédé son introduction en France.

Le terme *Machine* (ou *Mechanical) Translation* embarrasse considérablement les Français qui s'intéressent à la TA. Il est souvent assimilé à l'idée d'une machine à traduire, bien qu'un tel projet, aux États-Unis, ait connu un succès très limité. Le premier ouvrage français sur la TA, publié en 1959 par Émile Delavenay, fondateur de l'ATALA, s'intitule *la Machine à traduire*. L'auteur justifie ce terme en alléguant qu'il est déjà bien établi aux États-Unis, tout en se défendant de promouvoir par le terme *machine à traduire* (il parle également de *traductrice automatique* ou *électronique*) un outil simulant complètement une activité humaine. Delavenay déclare que l'intérêt réside davantage dans l'automatisation du processus de traduction, recherche essentiellement linguistique, que dans la création d'une machine clé en main, comme le font certains groupes américains. Ce sont les mêmes arguments que reprend Mounin en publiant en 1964 son ouvrage intitulé lui aussi *La machine à traduire* avec comme sous-titre *Histoire des problèmes linguistiques*.

[20] Sur l'histoire de l'ATALA, voir le numéro spécial de *TAL* 1992/1-2.

[21] L'Institut Blaise Pascal, berceau du calcul électronique en France, est créé en 1946, sous la supervision de Joseph Perès, directeur adjoint du CNRS pour les sciences. C'est à l'IBP que sera implantée, en 1955, une des premières calculatrices électroniques acquises par la France.

Il en va de même pour les sections de linguistique du CNRS. Dans le rapport de conjoncture du CNRS de 1959-60, la section « Linguistique et philologie non classique », dont fait partie Delavenay, salue la naissance des « machines à traduire », c'est-à-dire « les traductions automatiques effectuées par des machines de type logique », faisant écho aux vœux des mathématiciens appliqués dont une des priorités de recherche est l'étude logique de circuits et d'organes de machine adaptés à des fins de traduction ; l'emploi des calculatrices existantes est considéré comme un pis-aller.

Ces mêmes mathématiciens, qui parrainent la TA au CNRS, adoptent cependant sans atermoiements le terme *traduction automatique* (cf. le texte de création du CETA et les rapports de conjoncture du CNRS, section mathématiques appliquées, 1959-1960). Il est vrai que les mathématiciens sont familiers de l'automatique « traitant des procédés et des moyens réalisant l'exécution d'opérations données par des dispositifs mécaniques sans interventions humaines » qui constitue un des domaines couverts par les mathématiques appliquées au CNRS. Les machines à programme, destinées à effectuer le traitement automatique de l'information, sont appelées « automatiques »[22].

La revue de l'ATALA s'appellera d'ailleurs, à partir de 1965, *TA Informations. Revue Internationale des Applications de l'Automatique au Langage*.

Le terme *traduction automatique* s'impose donc rapidement, et c'est le rapport rédigé par Sestier en novembre 1959, un rapport circonstancié sur sa conception de la TA, qui va servir de base à la définition des orientations de la recherche en TA en France, notamment au CETA. Celui-ci doit avant tout poursuivre un objectif pratique et être capable de produire, dans un délai de cinq ans, des traductions en série d'articles scientifiques russes (objectif que reprend le rapport de conjoncture du CNRS 1959-60). Pour ce faire, il faut élaborer une « technologie linguistique » spécifique consistant à classer automatiquement les faits linguistiques, syntaxiques, morphologiques et lexicaux, à partir de critères définis par des linguistes et d'un corpus de textes à traduire. Mais la lenteur de ce travail de classement préalable, les difficultés théoriques quant à la validité des modèles utilisés – Sestier discute essentiellement les modèles de Chomsky et de Hays –, l'impossibilité d'obtenir un financement pour une machine plus puissante, et enfin la parution du rapport Bar-Hillel, rendent le délai de cinq ans complètement illusoire et conduisent Sestier à démissionner en 1962.

Le groupe de Vauquois partage avec celui de Sestier une conception de la traduction des langues naturelles, assimilée à celle des langages formels et de

[22] Le CNRS choisit de nommer *Section d'automatique documentaire* le centre créé en décembre 1960 sous la direction de Jean-Claude Gardin et destiné à la recherche en automatique non numérique - recherche automatique des informations bibliographiques, apprentissage automatique (*learning machines*) et programmes heuristiques (*intelligent machines*). C'est aussi un laboratoire d'automatique, le LADL (*Laboratoire d'Automatique Documentaire et Linguistique*), qui lui succédera en 1970, sous la direction de Maurice Gross.

programmation. Vauquois s'attachera à développer, dans la mise au point de systèmes de TA, l'analogie entre traduction et compilation, commentée par Boitet (2000) de la façon suivante : comme il n'y a pas de système de représentation explicite disponible dans les machines de première et deuxième génération, la tâche de traduction se réduit à transformer la forme en préservant le contenu. De la même façon, un compilateur transforme un programme en un programme équivalent, à savoir un programme qui calcule la même fonction, sans reconnaître cette fonction en tant que telle.

3.2. Confusion des termes et instabilité du champ

3.2.1 L'impossibilité de la linguistique computationnelle : un simple problème de traduction ?

Le terme *Computational Linguistics* qui s'impose aux États-Unis au début des années 1960 est source de difficultés innombrables pour les Français. En particulier, il semble impossible de le traduire par *linguistique computationnelle*. L'*Association for Machine Translation and Computational Linguistics*, créée en 1962, est annoncée dans *La Traduction Automatique*, revue de l'ATALA, comme *association pour la traduction mécanique et le calcul linguistique*. Quant à la rubrique *Recent publications in Computational Linguistics* du bulletin *The Finite String*, elle devient la rubrique *Linguistique mathématique et traduction automatique.*

En défaveur de l'usage du terme « computation » en français, il faut reconnaître qu'existe le terme « calcul ». Dans la section de Mathématiques appliquées du CNRS, on parle de la TA comme de la première application de calcul non numérique. Les centres d'informatique sont appelés « centres de calcul », etc.[23]

3.2.2 Linguistique quantitative et linguistique mathématique

En fait, la difficulté à nommer le domaine relève de causes plus profondes. La division entre disciplines ainsi que leur ancrage ne sont pas les mêmes qu'aux Etats-Unis, et les termes qui les désignent ne peuvent coïncider. Les Français se lancent dans la TA au moment où celle-ci commence à décliner et où l'analyse syntaxique

[23] Le débat est encore vif dans les années 1980. Le rapport du CNRS sur les Applications des Mathématiques aux Sciences de la Société et à la Linguistique, publié en 1984 sous la responsabilité d'A. Lentin, soutient le terme *Linguistique computationnelle* : « L'usage français hésite entre *linguistique algorithmique, linguistique calculatoire, informatique linguistique*. Les vieux mots français [souligné dans le texte] de *comput*, computation, très anciennement spécialisés pour désigner les calculs nécessaires à l'établissement du calendrier liturgique romain (date de Pâques) paraissent aujourd'hui disponibles pour de nouveaux usages. Alors pourquoi pas linguistique computationnelle ? » (p. 32).

automatique, fondée sur des grammaires formelles, s'impose comme méthode incontournable. Or on peut se demander si les Français, et en particulier les linguistes, sont vraiment prêts à adopter simultanément automatisation du traitement des langues et formalisation.

Les premiers acteurs de la TA en France sont avant tout des ingénieurs, Sestier ou Gross à Paris, Veyrunes ou Veillon à Grenoble. Vauquois a une formation de physicien. Il y a très peu de linguistes directement acteurs de la TA. Si l'on excepte Yves Gentilhomme pour le CETAP et Bernard Pottier et Guy Bourquin pour le groupe de Nancy, les linguistes intéressés étaient présents dans les instances d'évaluation du CETA et/ou membres de l'ATALA, mais pas engagés dans des projets concrets de TA. C'est le cas, pour ne citer que les plus prestigieux, d'Émile Benveniste, Georges Gougenheim, Michel Lejeune, André Martinet et Jean Fourquet. Il sera d'ailleurs beaucoup reproché au CETA de manquer de chercheurs et surtout de linguistes (cf. notamment le rapport Gross de 1967).

Des tentatives de formation réciproques entre linguistes et mathématiciens-informaticiens ont pourtant existé, parfois couronnées de succès. Jean Favard crée en 1959, en collaboration avec Daniel Hérault et René Moreau, une association chargée de « jeter un pont » entre linguistes et mathématiciens qu'ils nomment *Centre de Linguistique Quantitative*. Ce centre constituera un des hauts lieux, avec l'ATALA, de découverte des langages formels pour les linguistes français. Y enseigneront A. Martinet, J. Dubois, B. Pottier, M. Gross, A. Lentin, etc. Le Centre publiera pendant plusieurs années les *Documents de linguistique quantitative* consacrés à des questions de formalisation.

Le choix du terme *Linguistique quantitative* est explicitement commenté par les créateurs du centre : aucune traduction française ne semble convenir pour le terme *Computational Linguistics*, et à défaut des termes *Linguistique algorithmique* ou *mathématique* qui auraient fait peur aux linguistes, c'est celui de *Linguistique quantitative* qui est choisi. En choisissant ce terme, les objectifs des créateurs du centre sont clairs : ils ne veulent pas dissocier les aspects formels de la linguistique et les méthodes statistiques[24]. « L'enseignement du centre associe l'enseignement de la syntaxe formelle et celui de l'aspect aléatoire de la langue, syntaxe et aléatoire étant liés de façon indissociable dans la réalité du fonctionnement de la langue » (Hérault et Moreau, 1967, p. 114). Ainsi, alors que la *Quantitative Linguistics* (cf. *The Finite String*) reste aux États-Unis réservée aux strictes études statistiques, elle désigne en France à la fois les méthodes statistiques appliquées à l'écrit et l'étude des grammaires formelles.

[24] A noter que la théorie de l'information de Shannon et Weaver, et l'étude de la redondance lexicale des textes, a toujours droit de cité en France au début des années 1960, grâce notamment aux efforts de René Moreau qui avait fait partie du chiffre pendant la guerre, alors qu'elle était fortement critiquée et même mise sur la touche aux États-Unis (cf. Church et Mercer, 1993).

Notons également que *Computational Linguistics* est parfois traduit par *Linguistique mathématique*. Ce terme désigne soit les études statistiques, comme c'est le cas dans les pays d'Europe de l'Est, soit les travaux sur les grammaires formelles[25]. C'est ce dernier usage qui tend à se généraliser. C'est d'ailleurs celui que choisit le CETA de Grenoble (cf. *La Traduction automatique* de mars 1962). Dans un article intitulé « Linguistique mathématique et traduction automatique », les auteurs postulent que la TA s'est appuyée sur la syntaxe structurale et la formalisation des grammaires, et a apporté en retour une augmentation de la rigueur et un développement de la formalisation pour donner naissance à la « Linguistique mathématique ». Sont cités en exemple les travaux de Bar-Hillel, Lamb, Yngve et Lecerf.

C'est aussi ce dernier usage que choisit Tabory, chroniqueur des recherches américaines dans la revue de l'ATALA et auteur lui-même d'une méthode de TA : « la linguistique mathématique permet d'évaluer et de comparer les grammaires... c'est dans cette voie que désormais toute contribution à la linguistique théorique devrait se formaliser facilement… dans le cadre de la théorie chomskienne » (*La Traduction Automatique*, juin 1963).

Quelques années plus tard (*TA Informations*, 1969-1), Desclés et Fuchs tentent de mettre de l'ordre dans la confusion terminologique régnante, suivant des propositions formulées par Solomon Markus lors d'un séminaire international de linguistique formelle organisé en septembre 1968, à l'initiative du séminaire animé par Antoine Culioli à l'ENS et du Centre de Linguistique Quantitative. Parmi la dizaine de termes définis, la linguistique automatique ou linguistique cybernétique, qui « s'appuie souvent sur des travaux mathématiques mais travaille toujours en vue d'une application de la linguistique aux ordinateurs : d'une part études de langages formels et de programmation, d'autre part application des théories linguistiques, par exemple à l'élaboration des langages documentaires ». On notera que n'est nulle part définie la linguistique formelle, dont se réclament les auteurs, et qu'il y a une difficulté à admettre les travaux statistiques dans le champ, alors même que le Centre de Linguistique Quantitative, refusant de rejeter les méthodes statistiques au profit des seules grammaires formelles, est co-organisateur de la rencontre.

Institutionnellement pourtant, TA et linguistique formelle d'une part, études statistiques du vocabulaire d'autre part, sont nettement séparées. Au CNRS, les premières font partie de la section 'Linguistique générale, langues modernes et littérature comparée' (devenue en 1969 'Linguistique générale, langues et littératures étrangères') alors que les secondes font partie de la section « linguistique française et études littéraires » (devenue en 1969 « Études linguistiques et littérature française »). Il est vrai que, selon certains témoins de l'époque, les études statistiques, ancrées, on l'a dit, dans une solide tradition d'étude du vocabulaire français, étaient particulièrement florissantes et richement pourvues de moyens, surtout depuis la création du Trésor de la Langue Française en 1960.

[25] Ce qui en Europe de l'Est est désigné par le terme de *Linguistique algébrique.*

3.2.3 La Linguistique appliquée

Il y a un terme que les Français tentent d'adopter de façon consensuelle, c'est celui de *Linguistique appliquée*. La linguistique appliquée est d'ailleurs comprise dans le nom d'origine de l'ATALA (Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée). Le domaine de la linguistique appliquée comprend en France, outre l'enseignement des langues, toutes les applications de l'ordinateur au traitement des langues. Dans le rapport de conjoncture de 1963, les deux sections de linguistique du CNRS se félicitent du développement de cette discipline renouvelée qui recouvre deux ordres d'applications, l'enseignement des langues et la traduction automatique, en soulignant à la fois ses liens avec la linguistique fondamentale et son caractère d'application de nouvelles techniques à des fins linguistiques : ordinateur électronique pour la traduction, magnétophone à double piste en pédagogie.

La notion d'*Applied Linguistics* aux États-Unis est plus limitée et le *Center for Applied Linguistics*, créé en 1959, se consacre essentiellement à l'enseignement des langues étrangères et à l'enseignement de l'anglais à l'étranger. Mais *Linguistique appliquée* traduit aussi *Applied Language Analysis*, terme contenu dans le titre du colloque de Teddington de 1961 : *First International Conference on Machine Translation of Languages and Applied Language Analysis*.

Le premier Colloque international de Linguistique Appliquée, réuni en octobre 1964 à Nancy, propose comme sous-thème l'automatisation de la linguistique, en traduction automatique, en documentation automatique, et l'apport de la linguistique quantitative. En 1984 (cf. le rapport du CNRS consacré aux applications des mathématiques aux Sciences de la société et à la linguistique), la linguistique appliquée comporte toujours les deux types d'applications. Si elle a le mérite d'unifier l'ensemble des applications des technologies nouvelles à l'étude ou l'enseignement des langues, favorisant ainsi leur visibilité et leur subventionnement, la linguistique appliquée est loin de pouvoir constituer une discipline. Comme le signalent Desclés et Fuchs (1969, article cité), c'est un terme général ne faisant référence à aucune méthode particulière et la linguistique appliquée est loin de pouvoir revendiquer, comme tente de le faire la *Computational Linguistics*, le statut de discipline à part entière.

En 1965, au moment de la crise de l'ATALA, qui faillit disparaître atteinte de plein fouet par le rapport Bar-Hillel (1960) et par la constitution de l'ALPAC (1964), la revue change de nom et Delavenay cède son poste de président. Il salue alors, dans le premier numéro de *TA Informations* (1965, 1-2), l'apparition d'une nouvelle technologie linguistique, d'une linguistique appliquée, adaptée aux besoins de la machine. Ce terme assez ambigu chez Delavenay désigne d'une part la *MT linguistics* de Reifler, c'est-à-dire la linguistique pour la machine qui doit ignorer les résultats de la linguistique théorique, d'autre part la *Computational Linguistics*, cette *« new linguistics »* encouragée par l'ALPAC, et enfin les diverses applications du calcul automatique au langage. C'est ainsi qu'il promeut le terme *Traitement*

automatique du langage, dont on peut constater une des premières occurrences. Première occurrence ne signifie pas constitution du domaine. Celui-ci, on va le voir, peine encore à se constituer.

3.3 L'entrée en scène difficile du TAL

3.3.1. Vauquois et le Traitement automatique des langues

C'est Vauquois qui tente d'imposer le terme *Traitement automatique des langues*. La seconde conférence internationale de *Computational Linguistics*, organisée entre autres par l'AMTCL et l'ATALA, se tient à Grenoble en août 1967 sous le nom de « deuxième conférence en traitement automatique des langues ». Toujours en 1967, la revue *TA informations* change de sous-titre qui devient *Revue internationale du traitement automatique du langage*[26]. Sur la seconde page de couverture, figure cependant toujours l'ancien sous-titre : *Revue Internationale des Applications de l'Automatique au Langage*.

En 1969, Vauquois, président de l'ATALA depuis 1966, signe un article dans *TA Informations* intitulé « Dix ans d'ATALA : de la traduction automatique au traitement automatique des langues ». Vauquois y insiste sur le changement de nom, considérant que ce que les Américains ont appelé *Computational Linguistics* comme activité de remplacement de la traduction automatique n'a jamais été clairement défini. Il souhaite distinguer deux types de recherche (p. 60) :

1) Les travaux de linguistes conduits par des linguistes et qui utilisent des ordinateurs à titre d'outil : analyse de concordances, calculs de fréquence de mots, recherches lexicographiques, confection de dictionnaire à consultation automatique, et surtout grammaires génératives [sic].

2) Les équipes pluridisciplinaires formées de linguistes et de mathématiciens dont l'objectif n'appartient exclusivement à aucune de ces disciplines, mais au traitement automatique des langues proprement dit. Un tel projet recouvre une masse d'applications, comme la communication homme-machine, l'enseignement programmé, la documentation automatique, les études de formalisation sémantique, enfin la traduction automatique.

Cette prise de position de Vauquois, dans un contexte où les études statistiques restent dominantes, constitue une tentative de s'en démarquer pour former un champ autonome. Curieusement, toutefois, Vauquois exclut les travaux en grammaire générative du traitement automatique des langues, alors même que l'ALPAC les considère au cœur de la nouvelle linguistique. Est-ce une façon de se démarquer du centre de linguistique quantitative qui allie linguistique formelle et études statistiques ?

[26] Il semble qu'à cette époque on trouve alternativement *Traitement automatique du langage* ou *des langues*, sans que la différence soit explicitée.

Dans son rapport de conjoncture (1969-1975) et son rapport d'activité de 1970, la section 28 du CNRS « Linguistique générale, langues et littératures étrangères » consacre plusieurs pages au TAL (sous sa forme d'acronyme), où l'on retrouve les idées de Vauquois : « Le TAL se situe au carrefour des diverses disciplines : principalement linguistique, mathématiques, informatique. Ce caractère rend obligatoire la constitution d'équipes mixtes permettant l'élaboration d'un but commun atteint au moyen d'une méthodologie qui met en jeu les trois disciplines sources… les trois principaux résultats sont des modèles logico-linguistiques et algorithmiques permettant le traitement sur ordinateurs des problèmes d'analyse et de synthèse des langues naturelles, des programmes permettant des extractions et analyses de données sur fichiers linguistiques, et des débuts d'applications en traduction et documentation automatiques. Les grands axes de recherche à venir s'orientent vers des modèles sémantiques évolués et la communication homme-machine. »

Alors que le rapport de 1970 consacre plusieurs pages au langage pivot mis au point par le CETA, le Traitement automatique des langues disparaît des rapports d'activité du CNRS en 1971. Cela coïncide avec le déclin du CETA, amorcé en 1968, qui est alors divisé en deux équipes, le GETA, sous la direction de Vauquois, qui continue à se consacrer à la TA et à la formalisation, et l'équipe de Traitement automatique des langues et applications, dirigée par Jacques Rouault, qui développe un programme de reconnaissance automatique du français[27].

3.3.2. L'informatique linguistique

Un nouveau terme apparaît : l'*Informatique linguistique*. Dès 1972, les rapports de conjoncture tentent de promouvoir ce domaine « nouveau », « privilégié de la linguistique », et de lui donner autant d'importance que « la linguistique générale, les systèmes de communication, les études textuelles, et les études sur le terrain ». Toutefois l'informatique linguistique, telle que définie par le CNRS, fait reculer les modèles logico-mathématiques et la linguistique formelle, au profit des traitements statistiques dont on assiste au retour en force. Ainsi peut-on lire p. 53 du rapport de conjoncture de 1974 : « L'ordinateur intervient dans tous les cas où des décomptes, des statistiques attendent de l'ordinateur non seulement des résultats chiffrés, mais aussi des éditions : textes, listes de mots, tableaux morphologiques ou syntaxiques, lexiques, concordances, etc. »

Le rapport du CNRS sur les Sciences humaines pour 1976-77 témoigne du désarroi et de la dispersion du domaine : cinq contributions, toutes intitulées

[27] A la même époque, Michel Pêcheux élabore son Analyse automatique du discours (Pêcheux, 1969). AAD69 se présente comme un dispositif automatisé, chargé de généraliser les procédures mises au point par Harris (1952) afin de découvrir les structures invariantes de corpus repérées par leurs conditions socio-historiques de production. L'informatisation devait garantir « l'objectivité » de la lecture visant à découvrir les structures sous-jacentes. Le programme, écrit en Fortran, constitue sans doute une des premières tentatives de modélisation informatisée en linguistique structurale et sciences humaines.

« Informatique linguistique », se succèdent, avec pour auteurs respectifs : Bernard Vauquois, Bernard Quemada et Robert Martin, Françoise Bader et Christian Amphoux, Jean Leclant, Mario Borillo et Jacques Virbel, chacun proposant une répartition du champ différente selon ses objectifs, sans qu'il soit défini par des concepts ou des propositions théoriques ou méthodologiques claires. Les enjeux sont importants : il s'agit notamment de distribuer les heures de calcul du nouveau Service de Calcul Sciences Humaines créé en 1970 par le CNRS. De plus, la division de la linguistique en deux sections différentes ne contribue pas à unifier le domaine, et ce n'est qu'en 1983 que le CNRS créera une section unique de Sciences du langage.

Dans les années 1980, certains tentent de redonner à l'Informatique linguistique des perspectives théoriques, en dénonçant par là même l'inconsistance du domaine. Ainsi Desclés présente dans *TA informations* 1984-1 les articles issus d'une journée d'études de l'ATALA ayant eu lieu en 1981 sur le thème « La notion d'arbre en linguistique et en informatique » chargés « d'organiser le débat en termes conceptuels clairs et de ne pas réduire le domaine *informatique linguistique* (ou *linguistique informatique* – en italique dans le texte) à n'être seulement qu'un ensemble de techniques informatiques de traitement du langage naturel sans problèmes et sans fondements théoriques » (p. 3). En 1985, le premier cursus de Linguistique et informatique est fondé à l'Université Paris 7, à l'initiative de François Charpin.

3.4. L'officialisation du TAL

Au début des années 1990, deux événements en France vont marquer la prééminence du terme *TAL* et son installation comme dénominateur commun. D'une part, la revue de l'ATALA change une troisième fois de nom en 1992 pour s'appeler *TAL*. D'autre part, le ministère codifie un nouveau diplôme national en 1993 : il s'agit de la licence de Sciences du langage, mention « Traitement automatique des langues », assortie de la maîtrise de Sciences du langage, mention « Industries de la langue » qui vient à sa suite – on notera, au travers de cette double dénomination, une certaine incohérence de la part du ministère.

Le prix à payer est d'accepter de renoncer à dissocier les aspects théoriques et formels des applications industrielles, les statistiques ayant retrouvé droit de cité avec le traitement des grands corpus et les mémoires de traduction. Cette évolution a été favorisée par l'apparition, dans la seconde moitié des années 1980[28], de ce qu'on a appelé les industries de la langue et l'ingénierie linguistique, liée à certains développements technologiques, comme la micro-informatique, rendant accessibles

[28] En 1984, on parlait encore « d'applications des logiciels linguistiques pour l'industrie », comme en témoigne la journée d'études organisée par l'ATALA en avril 1984 par Bernard Normier (directeur d'ERLI fondée en 1977) et Alexandre Andreewsky. Le premier colloque sur les industries de la langue a eu lieu en 1987.

certaines applications du traitement automatique des langues aux particuliers et permettant l'émergence de petites entreprises spécialisées.

D'innombrables termes fleurissent pour désigner le domaine au début des années 1990 : Industries de la langue, Ingénierie linguistique, *Natural Language Engineering*, Technologies de la langue, etc. Il n'est toutefois pas certain que cette inflation de termes et cette frénésie de la dénomination parviennent à masquer l'inanité d'une impossible quête, celle de définir un champ unifié qui, tout en englobant les applications industrielles, soit scientifiquement fondé.

4. Quels contenus pour le TAL et la CL ?

Quand on étudie l'histoire du TAL et de sa constitution, on s'aperçoit que le problème majeur qui se pose est celui de l'unité d'une discipline, de l'existence d'un contenu unifié qui justifierait que le TAL soit une discipline. Or, en plus de la tension à laquelle il est soumis, comme on l'a vu, entre point de vue théorique et applications pratiques, le TAL est soumis à une tension qui provient des disciplines qui ont contribué à sa fondation et qui continuent d'exercer des attractions centrifuges.

4.1 Les quatre pôles autour desquels gravite le TAL

Afin de donner une caractérisation précise des disciplines en relation permanente avec le TAL, on peut se référer à Miller et Torris (1990, p. 15)[29]. Ils remarquent que le TAL fait intervenir « des domaines d'investigation très variés », qu'ils regroupent autour de cinq axes. Nous laissons de côté le premier de ces axes, la « *linguistique informatique* au sens plus étroit du terme[30] ». Les quatre autres axes sont la « *linguistique théorique* » qui doit fournir des « descriptions entièrement explicites », l'« *informatique théorique* » qui « permet d'optimiser les algorithmes et programmes de traitement » et qui développe « des techniques formelles qui influencent la linguistique informatique », l'« étude *mathématique* des propriétés formelles des outils de traitement et des théories linguistiques » et enfin les « recherches en *sciences cognitives* et en *intelligence artificielle* sur la représentation du savoir ».

Ces domaines sont de nature différente. Comme le notait déjà Oettinger (1960), à propos des deux seuls domaines qu'il considérait : « Research in automatic language translation draws from and contributes to two major disciplines – linguistics and

[29] Dont l'ouvrage, intitulé « Formalismes syntaxiques pour le traitement automatique du langage naturel », a certainement contribué à l'installation du terme TAL.

[30] Il est assez étrange de distinguer la linguistique informatique comme une sous-composante du TAL. Ceci est sans doute lié à l'époque où écrivent Miller et Torris, époque charnière où l'ancienne dénomination ne peut être oubliée, et où la nouvelle ne s'est pas encore imposée.

automatic information processing – one with firm roots in antiquity and tradition, the other newborn and not yet fully conscious of its identity » (p. xvii). En fait, parmi les quatre domaines, deux sont très anciens, les mathématiques et la linguistique, ils forment des disciplines bien établies. Mais ces disciplines anciennes ont avec le TAL des interactions qui ne se ressemblent pas : l'existence du TAL n'a qu'un effet négligeable sur les mathématiques, alors que la linguistique n'a pu ignorer le TAL. Qui, d'ailleurs, a eu l'ambition de renouveler profondément la linguistique et par là même a mis en cause la façon de travailler des linguistes traditionnels.

Deux domaines sont beaucoup plus récents, et leur définition ne va pas nécessairement de soi : l'informatique, et le pôle constitué par les sciences cognitives, l'intelligence artificielle, auquel on peut rattacher certains pans de la psychologie. Ce dernier pôle a sans doute, comme le TAL, des difficultés à se former et à se définir, d'où des conflits avec le TAL sur la délimitation des frontières et des tentatives d'absorption. C'est ainsi que Pereira et Grosz (1993) présentent le NLP comme un cas particulier des problèmes généraux de l'IA. « The papers exhibit the strong connections between NLP and such other areas of AI research : knowledge representation, reasoning, planning, and integration of multiple knowledge sources. » (p. 1). L'informatique a moins de mal à affirmer son existence. Elle introduit toutefois son ambiguïté inhérente[31] dans le TAL : est-ce une discipline scientifique, ou ne regroupe-t-elle que des applications pratiques ?

Le premier moment de la TA est sans doute un moment plus original, puisque c'est un moment de convergence : des acteurs, venus d'horizons divers et quelquefois très éloignés, convergent dans la réalisation d'une tâche commune. Mais très vite, les quatre disciplines ou regroupements de disciplines ont constitué des points de référence, des pôles d'attraction parfois antagonistes pour le TAL, exerçant constamment une influence sur le TAL, tant sur le plan institutionnel que sur le plan conceptuel.

En premier lieu, on s'aperçoit que les organismes de rattachement des acteurs du TAL dépendent presque toujours de l'un de ces quatre pôles, de même que les départements où sont proposés les enseignements de TAL. Il est intéressant également de noter que les articles qui ont trait au TAL sont souvent publiés dans des revues d'informatique. Ainsi, en 1962, Kuno et Oettinger publient un des tout premiers algorithmes d'analyse syntaxique basé sur les grammaires de type 2 dans *Information Processing* (IFIP Congress) (le domaine considéré est donc celui du traitement de l'information). Le congrès est en fait un très vaste congrès d'informatique, qui fait le point sur tous les travaux en cours, où l'on traite aussi bien des structures des machines que de l'informatisation des banques ou de traduction automatique. En 1964, Oettinger crée une section de *Computational*

[31] On notera le caractère imprécis en français du terme « informatique » qui renvoie aussi bien à ce qui se veut une discipline scientifique qu'à un rayon dans une grande surface, alors qu'en anglais on emploie le terme « computer science », sans doute plus précis.

Linguistics dans les *Communications of the ACM*. C'est dans cette revue que paraîtra en 1970 l'article de Woods sur les ATN.

Dans la période qui suivra, on trouvera des articles importants dans des revues d'intelligence artificielle, ainsi l'article de Pereira et Warren sur les DCG (*Definite Clause Grammars*, 1980) est publié dans *Artificial Intelligence*. Inversement, assez peu d'articles de TAL paraîtront dans des revues généralistes de linguistique ou de mathématiques.

On remarquera une confusion certaine dans les étiquettes, des croisements qui prouvent que les positions ne sont pas fermement établies. Pour ne citer que deux exemples, Colmerauer qui appartient au « Groupe d'intelligence artificielle » publie en 1979 « Un sous-ensemble intéressant du français » dans *RAIRO Informatique théorique*. D'un autre côté un article de Levy et Joshi (« Skeletal Structural Descriptions »), membres de départements de *Computer and Information Science(s)*, placé par ses auteurs dans le cadre de la « mathematical linguistics », est publié en 1978 dans une revue d'informatique, *Information and Control*[32].

On peut penser que si des articles qui, à l'évidence, relèvent du TAL ou de la CL sont publiés dans des revues rattachées à d'autres disciplines – alors qu'existent des revues de TAL –, c'est bien parce que le TAL cherche sans cesse à confirmer sa légitimité à travers une reconnaissance par d'autres disciplines.

En deuxième lieu, le TAL est constitué par des acteurs venus de ces quatre horizons, et il hérite des acquis théoriques accumulés par ces disciplines. Il en découle naturellement la question suivante : l'amalgame s'est-il vraiment fait ? A-t-on formé des « talistes » dont les bases de compétence seraient bien définies ? Ou bien compte-t-on toujours sur des équipes interdisciplinaires, faites de chercheurs aux compétences variées, comme le souhaitait par exemple Vauquois en 1969 ? S'appuie-t-on toujours essentiellement sur des travaux issus de disciplines externes au TAL, comme quand Winograd (1983, p. vii) indique que ses références bibliographiques proviennent de la linguistique *et* de l'informatique : « it is a reference source with many pointers into the literature of both linguistics and computer science » ?

Mais, en troisième lieu, se pose la question la plus importante : celle des objectifs et des problématiques du TAL. L'objectif, qui était unifié quand il s'agissait de faire de la TA, ne devient-il pas éclaté et divergent selon le pôle auquel les auteurs des différents travaux se rattachent prioritairement ? Le critère de qualité des travaux ne dépend-il pas de la problématique de la discipline interne/externe qui est privilégiée ?

Ainsi, si on se place dans la perspective de la linguistique qui est, dit grossièrement, de décrire les langues et l'activité langagière, pourquoi faire du traitement automatique ? Deux raisons possibles ont été proposées, et ce dès le

[32] Ce que l'on peut voir comme un indice de rattachement de la linguistique mathématique au TAL. Cf. ci-dessous.

rapport de l'ALPAC : d'une part la possibilité d'accéder à de grands corpus, et d'autre part la constitution de modèles formels, créés pour le traitement automatique, mais qui peuvent se révéler féconds dans l'activité de description linguistique.

L'objectif de l'informatique est ambigu : il s'agit d'une part, dans le cadre de ce qu'on appellera bien après leur apparition les Industries de la langue, de construire des logiciels efficaces et qui répondent à une demande sociale ou économique ; d'autre part, dans un cadre qu'on peut rattacher à l'informatique théorique, de définir des algorithmes généraux, des structures de données (formelles) ou même des langages de programmation qui s'inscrivent dans des recherches plus larges sur les algorithmes, les structures de données ou les langages de programmation. Et donc qui puissent être appliqués à d'autres domaines que le traitement des langues naturelles.

Les mathématiques n'exportent pas vraiment un objectif au sein du TAL, mais elles imposent plutôt des critères de rigueur aux objets utilisés, rejoignant dans une certaine mesure l'informatique théorique. Mais, si on se livre à des recherches mathématiques sur les objets en question considérés comme de pures abstractions, on s'éloigne du TAL.

L'Intelligence artificielle vise la « compréhension », la simulation des comportements humains, l'étude du raisonnement. On peut citer de nombreux travaux comme celui de Kaplan (1975) paru dans un volume intitulé *Explorations in cognition* et dont la première phrase est « How does the human listener comprehend sentences ? ». La thèse de l'auteur dont découle sa contribution est signalée comme ayant été préparée dans un « Department of Psychology and Social Relations ».

D'autres objectifs ont été donnés au TAL, ainsi la « communication homme-machine », que l'on peut rattacher à l'informatique, mais l'ergonomie et la psychologie ne sont pas loin. Par exemple, en 1978 paraît un numéro de *Lecture Notes in Computer Science* sur « Natural Language Communication with Computers », avec notamment l'article de Colmerauer sur les grammaires de métamorphose.

Il y a néanmoins eu la volonté de définir des objectifs propres au TAL (ou à la CL) et unifiés, comme nous allons le voir à présent.

4.2 Les contenus conceptuels du TAL et de la CL

La publication du rapport de l'ALPAC constitue un des actes fondateurs de la *Computational Linguistics*. Non seulement il porte un coup d'arrêt à la TA, mais il suggère de développer une nouvelle discipline, qu'il nomme, et dont il caractérise le contenu.

En fait, le contenu proposé est déjà porteur des contradictions internes à la discipline promue. En effet, un point important est la volonté de fédérer toutes les applications sous une même étiquette. En même temps, la recherche fondamentale est favorisée par rapport aux applications – ce qui constitue une rupture par rapport à la période précédente.

Et, ce qui paraît s'opposer à la perspective de fonder une discipline unifiée, les recherches en CL sont justifiées notamment par des effets bénéfiques envisagés sur deux des disciplines qui sont à son fondement : l'informatique et la linguistique.

C'est ainsi aux relations entre langages de programmation et linguistique qu'est consacrée l'annexe 18 du rapport. Par exemple, ce sont des règles équivalentes aux règles de réécriture des grammaires de type 2 qui ont servi à définir et à déterminer les spécifications d'Algol 60. L'existence de la linguistique mathématique permet d'envisager la conception des langages de programmation d'une manière plus rigoureuse, plus systématique. Il devient possible de définir des grammaires pour les langages de programmation, ce qui permet d'en donner des descriptions précises et transmissibles, et de bâtir des analyseurs syntaxiques généraux, qui s'appliquent à tous les langages de programmation d'un type donné.

Mais c'est sur la linguistique que les effets les plus importants sont escomptés : « The advent of computational linguistics promises to work a revolution in the study of natural languages. » (p. 29) L'analyse syntaxique par des ordinateurs apporte des connaissances nouvelles au linguiste. Ainsi, aucun linguiste ne peut être indifférent à l'accès aux grands corpus. L'ALPAC propose, dans les recherches à développer, de construire des outils de manipulation du langage, afin d'aider les linguistes à découvrir et énoncer leurs généralisations, puis à vérifier ces généralisations en regard des données. Les ordinateurs doivent permettre aussi aux « linguistic scientists » d'expliciter en détail les théories qu'ils produisent, par exemple les grammaires et les théories de la signification (p. 31).

L'ordinateur, premier objet susceptible de manipuler des symboles hors du cerveau des êtres humains, peut changer le niveau d'analyse des langues, comme le microscope a changé la biologie (annexe 19, p. 121). Comme on le voit, les attentes des auteurs du rapport sont considérables.

S'interrogeant sur l'influence du traitement automatique sur la linguistique, l'ALPAC s'intéresse aux avancées apportées à la linguistique par la formalisation, et spécialement aux travaux de Chomsky. Elle essaie de mesurer la contribution de l'informatique à ces avancées : « Though it seems clear that the computer was not at the center of most of this in a direct causal fashion, it has surely played a significant role, both of interplay and as a tool for validation. » (p. 121)

Mais, Chomsky[33], quant à lui, rejette le traitement automatique et refuse de se rattacher à la *Computational Linguistics*. Les travaux de Chomsky et de ses

[33] Bien qu'ayant été embauché par Yngve en 1955 sur un projet de TA, cf. § 2.1 ci-dessus.

successeurs, du moins les plus formalisés, seront étiquetés sous les termes de linguistique formelle, linguistique mathématique ou linguistique algébrique.

Le lien entre les travaux de linguistique formelle et la CL est pourtant patent. Sur le plan institutionnel, comme on l'a vu, les acteurs de la linguistique formelle sont très souvent soutenus par des départements d'informatique, les articles sont publiés dans des revues de CL. Ce n'est pas par hasard que la rigueur formelle est apparue exactement en même temps que débouchaient les premiers travaux de TA, et qu'elle a été le fait de chercheurs qui au moins connaissaient les travaux informatiques. Quant au plan du contenu, il faudrait être aveugle pour ne pas voir que la linguistique s'est enrichie de nombre de ses outils formels de description « au contact de l'informatique », selon l'expression de Cori et Marandin (2001). Il est ainsi intéressant de noter que le rejet des grammaires transformationnelles par Gazdar, à la source de la définition des GPSG, s'appuie notamment sur l'existence d'analyseurs syntaxiques efficaces fondés sur les grammaires syntagmatiques, alors que les GT ne peuvent donner lieu à de tels analyseurs. Par ailleurs, toute une série de formalismes de représentation – DCG, FUG, PATR II – sont explicitement des formalismes qui se donnent pour perspective l'informatisation. Et ces formalismes sont parmi les ancêtres des HPSG.

Il reste que, si la linguistique formelle doit beaucoup au traitement automatique, il peut être juste intellectuellement de séparer les travaux de linguistique formelle de leur éventuelle utilisation par des programmes informatiques, comme nous allons le voir à présent.

4.3 Le tournant déclaratif : des ATN aux DCG

Une tendance, dès les débuts de la TA, a été de vouloir produire des outils spécifiques au domaine. C'est ainsi que, dans les années 1950, circulait l'idée d'une machine à traduire, dont la structure physique même devait être adaptée à son utilisation. Dans les années 1960, cette idée a été remplacée par la recherche de langages de programmation spécialisés. On notera que l'ALPAC s'est félicité que le traitement automatique ait conduit à la définition de modèles linguistiques à la fois théoriques et spécifiquement construits pour être utilisés comme langages de programmation. Et il est vrai que certains langages ont été développés pour résoudre des problèmes linguistiques précis, c'est le cas de Snobol ou Comit, avec leurs structures de listes.

Oettinger (1960) va dans le même sens quand il déplore (p. 101) que la distinction entre problème linguistique, algorithme et programme ait besoin d'être faite et souhaite que ces trois étapes puissent être réduites à une seule opération.

Les ATN (Woods, 1970) constituent d'une certaine manière un aboutissement presque idéal de cette tendance à l'intégration des différents niveaux de traitement. Les ATN, en effet, peuvent être vus tout à la fois comme un outil informatique

d'analyse syntaxique, comme un langage de programmation défini à partir du langage LISP[34], ou comme un formalisme permettant de décrire des langues naturelles. Ceci est rendu possible par le fait que dans le langage LISP les programmes et les données ont la même structure. On notera que les ATN ont occupé une place considérable dans l'histoire du TAL, étant quasi-hégémoniques pendant une décennie (1970-1980).

La perspective est déjà différente quand des auteurs, dans un même travail, proposent tout à la fois la définition d'un modèle de grammaire, des algorithmes et la description de faits linguistiques, mais en séparant les différents niveaux. Tel est par exemple le cas de Yngve (1960) :

« In this paper, a simple and easily mechanized model for sentence production is set up. […]

The model arose out of research directed toward the mechanical translation of languages. […] » p. 444

« The model consists of a grammar and a mechanism. The grammar contains the rules of the particular language that is being produced. The mechanism, on the other hand, is quite general and will work with the grammar of any language. […]

The mechanism gives precise meaning to the set of rules by providing explicitly the conventions for their application. » p. 445

Yngve défend ses choix formels en s'appuyant sur la programmation, mais il ne mélange pas la définition de son formalisme et l'utilisation du formalisme. Mel'cuk (2000) rend hommage au caractère pionnier de Yngve, quand celui-ci conçoit le langage Comit : la représentation des règles de syntaxe sous forme de structures de listes, et l'utilisation de cette structure de liste par un programme indépendant. Mel'cuk ajoute que ceci permet de faire appel à des paquets de règles, comme si c'étaient des sous-programmes activés autant de fois que nécessaire dans le même programme. En ce sens, on peut dire que Yngve admet la séparation procédural/déclaratif, avant que celle-ci ne soit explicitement prônée, à la fin des années 1970.

Ce sont les systèmes experts, dont la première réalisation répertoriée date de 1976 (Shortliffe, 1976), qui ont conduit à privilégier la représentation déclarative des connaissances. Selon Laurière (1988, p. 10) « Déclaratif signifie :

a) qu'il s'agit de simples énoncés : quelque chose est affirmé et non pas ordonné (dans le sens "imposé") ;

b) que ces éléments de connaissances sont donnés indépendamment de leur mode d'emploi ;

c) que les énoncés en question sont fournis en vrac ; les affirmations successives ne sont pas ordonnées (dans le sens "classées") ; »

[34] Langage qui est lui-même dédié aux applications symboliques et à l'intelligence artificielle.

Les DCG (Pereira et Warren, 1980) sont sans doute le premier modèle qui se revendique clairement déclaratif.

« The greater clarity and modularity of DCGs is a vital aid in the actual development of systems of the size and complexity necessary for real natural language analysis. Because the DCG consists of small independent rules with a declarative reading, it is much easier to extend the system with new linguistic constructions, or to modify the kind of structures which are built. » (Pereira et Warren, 1980, p. 270).

Comme l'article fondateur des DCG démontre que leur puissance est équivalente à celle des ATN, la déclarativité devient un argument décisif en faveur de leur supériorité.

Quand un modèle est déclaratif, on doit distinguer très nettement les descriptions, qui sont fournies en vrac, dans un langage qui n'est en général pas un langage de programmation, et les procédures qui œuvrent sur ces descriptions, prises comme données. Cela induit une division du travail entre ceux qui élaborent les descriptions et ceux qui écrivent les procédures, c'est-à-dire en gros entre informaticiens et linguistes. Pereira et Warren insistent bien sur le fait que même les concepteurs de modèles ne sont pas habilités à résoudre les problèmes linguistiques. « It is NOT our intention to propose any definite solutions to the many unsolved linguistic problems of particular languages such as English; we describe only how DCGs can be used, not how they should be used. » (Pereira et Warren, 1980, p. 232).

Cette évolution vers des modèles déclaratifs a les plus grandes conséquences sur l'unité du TAL. Les acteurs, en effet, sont incités à se replier sur une spécialité donnée : les uns sur la description des données linguistiques, d'autres sur l'écriture de modèles, d'autres enfin sur la mise au point d'algorithmes. On peut dire que réapparaissent sous la forme de lignes de fracture les frontières entre les disciplines dont les apports variés ont permis que soit fondé le TAL en tant que domaine. Seuls peuvent se réclamer sans équivoque du domaine les ingénieurs qui réalisent des applications industrielles et qui, donc, ont simultanément besoin des algorithmes et de la description des données. Alors même qu'en adoptant des modèles déclaratifs le TAL cherche à s'imposer des critères de rigueur, il tend du même coup à rendre impossible sa constitution comme discipline scientifique.

5. Conclusion

Pour conclure, il faut souligner deux lignes de tension constantes dans l'histoire du TAL : la cohabitation paradoxale et nécessaire des recherches théoriques et des applications à visée industrielle d'une part, les antagonismes entre le TAL et les différentes disciplines qui le constituent, voire entre ces disciplines elles-mêmes quand elles rentrent en interaction dans un problème de TAL.

Si l'on considère le TAL aux débuts de la TA, ce sont les applications à visée industrielle qui sont privilégiées, avec la machine à traduire, dédiée à la traduction de bonne qualité entièrement automatisée. La machine à traduire, comme la machine à écrire ou à laver, en ces temps d'essor sans précédent de la consommation, doit fournir des traductions en série répondant à la demande sociale. Les questions théoriques comme les tensions entre les disciplines sont très réduites puisqu'on fait appel à une linguistique pour la machine, totalement *ad hoc* et déconnectée des résultats de la linguistique théorique. Les plus grands développements sont à attendre de la technologie et de cette nouvelle science que constitue l'informatique.

La première rupture au début des années 1960 a fait émerger la *Computational Linguistics*, comme tentative de séparer les applications à visée industrielle des recherches théoriques fondamentales. Elle a été rendue possible par la naissance des algorithmes d'analyse syntaxique : alors que, jusque dans les années 1960, toute programmation en TA constituait un tour de force, à cause de l'intrication entre programmation et grammaire qui empêchait toute évaluation linguistique des systèmes et tout progrès, le développement des langages formels va permettre de penser les problèmes de façon déclarative en distinguant la grammaire (la description linguistique), les langages formels (qui rendent les informations linguistiques traitables par la machine) et les stratégies d'analyse (Kay, 1982).

Toutefois, la *Computational Linguistics* n'a pas réussi à se centrer exclusivement sur des recherches théoriques. Institutionnellement, elle couvre en effet tous les domaines de l'application de l'informatique à des données langagières, incluant des études statistiques ou documentaires.

La seconde rupture, à la fin des années 1980, prend acte de l'irruption d'une quatrième discipline, l'Intelligence artificielle, en même temps que la nécessité, une fois de plus renouvelée, de séparer applications et recherche théorique. Devant la difficulté d'autonomiser les investigations théoriques en discipline à part entière, et face aux contraintes définissant une « bonne application », impossibles à satisfaire totalement, on tentera de séparer, parfois de façon tout à fait artificielle et au risque de la faire disparaître, la *Computational Linguistics*, traditionnellement consacrée à l'interaction entre linguistique, mathématique et informatique, et considérée comme plus théorique, au profit du *Natural Language Processing*, nouveau champ intégrant l'IA et délibérément orienté vers les applications.

En France, le TAL, qui a rencontré davantage de difficultés à s'implanter institutionnellement, semble avoir acquis un rôle fédérateur entre recherches théoriques, pour lesquelles on utilise encore parfois le terme de linguistique informatique, et applications. Quel statut alors donner à l'ingénierie linguistique ou aux industries de la langue, qui comme leur nom l'indique, désignent clairement les applications industrielles ? Ces termes sont-ils avant-coureurs d'une nouvelle rupture ? Alors qu'on pouvait penser qu'un point de stabilité du TAL était constitué par une communauté de chercheurs d'origines disciplinaires diverses, l'essor des applications au détriment des travaux théoriques ne va-t-il pas conduire les acteurs,

informaticiens et linguistes, à retourner dans leurs disciplines respectives en abandonnant le terrain aux seuls ingénieurs ? La pérennité du TAL comme champ autonome semble ne jamais pouvoir être définitivement acquise.

Nous tenons à remercier ici nos relecteurs pour leurs commentaires et suggestions dont nous espérons avoir tiré le meilleur profit.

Bibliographie

Rapports:

Language and Machines. Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, National Research Council. 1966.

Rapports de conjoncture et d'activité du CNRS (1959-1977), archives du CNRS.

Rapport du CNRS sur les Applications des Mathématiques aux Sciences de la Société et à la Linguistique, sous la responsabilité d'A. Lentin, *Mathématiques et Sciences Humaines* 1984 n°86.

Revues:

La Traduction Automatique (1960-1964), *T.A . Informations* (1965-1991), *TAL* (1992-).

Mechanical Translation (1954-1965), *Mechanical Translation and Computational Linguistics* (1965-1973), *American Journal of Computational Linguistics* (1974-1983), *Computational Linguistics* (1984 -).

The Finite String (1964-1973), bulletin de l'AMTCL (*Association for Machine Translation and Computational Linguistics*) devenue ACL (*Association for Computational Linguistics*) en 1973.

International Journal of American Linguistics, 1954, vol. 20 : 4, *Translation Issue*.

Archaimbault S., Léon J., 1997, « La langue intermédiaire dans la Traduction Automatique en URSS (1954-1960). Filiations et modèles », *Histoire Epistémologie Langage* 19-2 : 105-132.

Asher R.E. (ed.), 1994, *The Encyclopedia of Language and Linguistics*, Oxford, New York, Seoul, Tokyo, Pergamon Press.

Bar-Hillel Y., 1953a, « The present State of Research on Mechanical Translation [1951] », *American Documentation* 2 : 229-237.

Bar-Hillel Y., 1953b, « A Quasi-Arithmetic Notation for Syntactic Description », *Language* 29 : 47-58.

Bar-Hillel Y., 1953c, « Some linguistic problems connected with Machine Translation », *Philosophy of Science* n°20 : 217-225.

Bar-Hillel Y., 1960, « The Present Status of Automatic Translation of Languages », *Advances in Computers* vol.1, F.C. Alt ed. Academic Press, N.Y., London : 91-141.

Boitet C., 2000, « Bernard Vauquois' contribution to the theory and practice of building MT systems: a historical perspective » in Hutchins W.J., 2000, *Early Years in Machine Translation*, John Benjamins, Amsterdam, Philadelphia : 331-348.

Chevalier J.-C., Encreve P., 1984, « La création de revues dans les années 60. Matériaux pour l'histoire récente de la linguistique en France », *Langue Française* n°63 : 57-102.

Church K. W., Mercer R. L., 1993, « Introduction to the Special Issue on Computational Linguistics Using Large Corpora », *Computational Linguistics* vol 19 n°1 : 1-25.

Colmerauer A., 1978, « Metamorphosis grammars », in Bolc L. (ed.) *Natural Language Communication with Computers*, Lecture Notes in Computer Science 63, Springer-Verlag, Berlin.

Colmerauer A., 1979, « Un sous-ensemble intéressant du français », *RAIRO Informatique théorique*, vol. 13, n°4 : 309-336.

Cori M., Marandin J.-M., 2001, « La linguistique au contact de l'informatique : de la construction des grammaires aux grammaires de construction », *Histoire Épistémologie Langage*, vol.23-1 : 49-79.

Cunningham H., 1999, « A definition and short history of Language Engineering », *Natural Language Engineering* 5-1 : 1-16.

Delavenay E., 1959, *La machine à traduire*, Que-sais-je? Paris, PUF.

Descles J-P., Fuchs C., 1969, « Le séminaire international de linguistique formelle », *TA Informations*, 1969-1 : 1-5.

Dreyfus H., 1972, *What Computers can' t do*, Harper and Row.

Gross M., 1967, *Notes sur certains aspects des recherches en linguistique au CNRS*, rapport au directeur général du CNRS du 26 septembre 1967.

Harris Z.S., 1952, « Discourse Analysis », *Language* 28 : 1-30.

Hays D. G., 1962, « Automatic language-data processing », *Computer Applications in the Behavioral Sciences*, H. Borko ed. Prentice Hall, Englewood Cliffs, N.J. : 395-421.

Hays D. G., 1972, « The field and scope of Computational Linguistics », *The Finite String*, vol 9 nov-dec 1972 : 2 et sq.

Hérault D., Moreau R., 1967, « La linguistique quantitative », *Revue de l'enseignement supérieur*, n°1-2 : 113-127.

Hutchins W.J., 1996, « ALPAC: The (In)famous Report », *MT News International* 4 : 9-12.

Kaplan A., 1955 (1950), « An experimental study of ambiguity in context », *Mechanical Translation*, vol. 2, n° 2 : 39-46.

Kaplan R.M., 1975, « On Process Models for Sentence Analysis », in D.A. Norman, D.E. Rumelhart, *Explorations in cognition* W.H. Freeman and Company, San Francisco : 117-133.

Kay M., 1982, « Machine Translation », *American Journal of Computational Linguistics*, vol.8 n°2 : 74-78.

Kay M., 1992, Préface de *An Introduction to Machine Translation*, (W. John Hutchins et Harold L. Somers ed.), London, Academic Press Ltd.

Kay M., 2000, « David G. Hays », in Hutchins W.J., *Early Years in Machine Translation*, John Benjamins, Amsterdam, Philadelphia : 165-170.

Koutsoudas A., Humecky A., 1957, « Ambiguity of syntactic function resolved by linear context », *Word* 13 : 3 : 403-414.

Kuno S., Oettinger A.G., 1962, « Multiple-path Syntactic Analyzer », *Information Processing* VII, 1, North-Holland, Amsterdam : 306-311.

Lamb S.M., 1962, « On the mechanization of syntactic analysis », *Proceedings of the International Conference on Machine Translation and Applied Language Analysis*, Teddington 1961 : 673-686. London: HMSO [traduction française : 1964, « La mécanisation de l'analyse syntaxique », *Traduction automatique et linguistique appliquée*. Choix de communications présentées à la Conférence internationale sur la Traduction mécanique et l'Analyse linguistique appliquée, National Physical Laboratory Teddington 1961, ATALA (ed.), PUF : 169-183].

Laurière J.-L., 1988, *Intelligence artificielle, Tome 2 représentation des connaissances*, Eyrolles, Paris.

Léon J., 1998, « Les débuts de la traduction automatique en France (1959-1968) : à contretemps ? », *Modèles Linguistiques*, tome XIX, fascicule 2 : 55-86.

Léon J., 2000, « Traduction automatique et formalisation du langage. Les tentatives du Cambridge Language Research Unit (1955-1960) », *in The History of Linguistics and Grammatical Praxis* (eds. P.Desmet, L.Jooken, P.Schmitter, P.Swiggers) Louvain/Paris, Peeters : 369-394.

Léon J., 2001, « Conceptions du mot et débuts de la traduction automatique », *Histoire Épistémologie Langage*, vol.23-1 : 81-106.

Levy L.S., Joshi A.K., 1978, « Skeletal Structural Descriptions », *Information and Control* 39 : 192-211.

Masterman M., 1957, « The Thesaurus in Syntax and Semantics », *Mechanical Translation*, vol 4 : 1-2 : 35-44.

Melby A., 1992, « The translator workstation », *Computers in Translation, A practical Appraisal*, éd. par John Newton, London, Routledge : 147-165.

Mel'cuk I.A., 1960, « K voprosu o grammaticeskom v jazyke-posrednike », *Masinnyj Perevod i Prikladnaja Linguistika* 4 : 25-451 [Trad. angl. : « The problem concerning the 'grammatical' in an intermediate language », JPRS/8026].

Mel'cuk I.A., 2000, « Machine translation and formal linguistics in the USSR », in Hutchins W.J., 2000, *Early Years in Machine Translation*, John Benjamins, Amsterdam, Philadelphia : 205-226.

Miller P., Torris T., 1990, *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Hermès, Paris.

Mounin G., 1964, *La machine à traduire. Histoire des problèmes linguistiques*, La Haye Mouton.

Oettinger A.G., 1960, *Automatic Language Translation*, Harvard University Press.

Pêcheux M., 1969, *Analyse automatique du discours*, Dunod, Paris.

Pereira F., Grosz B.J., 1993, « Introduction », *Artificial Intelligence, Special volume Natural Language Processing*, 63 1-2 : 1-15.

Pereira F, Warren D, 1980, « Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks », *Artificial Intelligence*, 13 (3) : 231-278.

Reifler E., 1950, « Studies in Mechanical Translation, n°1, MT » Mimeographed, 51pp., Jan.10,.[cité dans W.N.Locke and A.D. Booth (eds.), 1955, *Machine Translation of Languages*, 14 essays, MIT et John Wiley].

Reifler E, 1955, « The Mechanical Determination of Meaning », Locke and Booth (eds.), *Machine Translation of Languages*, 14 essays, MIT et John Wiley : 136-164.

Sestier A., 1959, Comment doit être organisé à l'échelle française l'effort pour la TA (rapport du 23-11-59).

Shortliffe E., 1976, *Computer-based medical consultation, MYCIN*, Elsevier.

Vauquois B., 1962, « Langages artificiels, systèmes formels et traduction automatique », in *Automatic Translation of Languages*, Papers presented at NATO Summer School held in Venice, July 1962, Oxford, Pergamon Press : 211-236.

Vauquois B, 1969, « Dix ans d'ATALA: de la traduction automatique au traitement automatique des langues »,. *TA Informations* 1969-2 : 57-61.

Weaver W., [1949] 1955, « Translation » in *Machine Translation of Languages*, 14 essays, (W.N.Locke and A.D. Booth, eds.), MIT et John Wiley **:** 15-23.

Winograd T., 1983, *Language as a Cognitive Process, Volume I : Syntax*, Addison-Wesley, Reading, Massachusetts.

Woods W.A., 1970, « Transition Network Grammars for Natural Language Analysis », *Communications of the ACM*, vol. 13, n° 10 : 591-606.

Yngve V.H., 1960, « A Model and an hypothesis for language structure », *Proceedings of the American Philosophical Society*, vol.104, n°5 : 444-466.

Yngve V.H., 1962, « Random generation of English Sentences », *First International Conference on Machine Translation of Languages and Applied Language Analysis*, Teddington, sept 1961, HM Stationery Office, London: 65-82.