Version de travail
Paru dans
2000, Léon Jacqueline, "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.

TRADUCTION AUTOMATIQUE ET FORMALISATION DU LANGAGE. LES TENTATIVES DU CAMBRIDGE LANGUAGE RESEARCH UNIT (1955-1960)

Jacqueline Léon
UMR7597, CNRS , Paris 7

INTRODUCTION

Le groupe britannique CLRU (Cambridge Language Research Unit) commence ses recherches en traduction automatique (infra TA) en 1954-55, dans le sillage de la première démonstration de TA sur ordinateur effectuée par les américains en janvier 1954 à New York. Certains membres du groupe font déjà partie des pionniers dans le domaine. Ainsi Booth[1] et Richens ont entrepris des expérimentations de traduction mot à mot dès 1948, en même temps que les Américains. Le CLRU développe des modèles théoriques de TA en privilégiant une approche sémantique et une méthode par thesaurus, formalisé selon la méthode des treillis mise au point par Birkhoff. Les travaux sont foisonnants et semblent parfois manquer de cohérence, c'est ce qu'a dénoncé du moins Bar-Hillel dans son rapport de 1960. Pourtant, nous faisons l'hypothèse qu'à travers cette diversité des recherches émergent différentes tentatives de formalisation du langage et que celles-ci doivent beaucoup au fait qu'il s'agit d'élaborer des dispositifs de traduction automatique. Le projet défendu par le CLRU permet en effet de s'interroger sur ce qu'il en est de la formalisation quand elle se donne pour objectif l'élaboration d'une nouvelle technique. Par ailleurs, l'approche sémantique du CLRU introduit la question de savoir s'il est possible de formaliser les langues quand on écarte précisément ce qui leur est propre, à savoir la grammaire ou la morphologie. Enfin il est intéressant de voir comment, à cette époque, se met en place une conception de la formalisation du langage fondée sur une réflexion sur le contexte et le sens, indépendamment et concurremment aux hypothèses logico-mathématiques de Chomsky ou Bar-Hillel.

---

[1] A.D. Booth dirigeait le second groupe britannique de TA à Londres et est l'un des premiers utilisateurs des calculateurs électroniques

1. PRÉSENTATION DU GROUPE DE CAMBRIDGE

1.1. LE CLRU ET LE RAPPORT BAR-HILLEL

Le CRLU, dirigé par Margaret Masterman, philosophe et logicienne, est composé de linguistes comme M.K. Halliday, alors assistant en chinois, d'un statisticien A.F. Parker-Rhodes, de mathématiciens, physiciens et philosophes. A ce groupe se joindront ultérieurement, à partir de 1958, les linguistes Martin Kay et Paul Bratley puis Yorik Wilks. Il ne comprend pas d'ingénieurs mais bénéficiera à partir de 1957 des conseils de l'américain Gilbert King, auteur de la technologie de la mémoire photoscopique, qui permettait d'accroître de façon considérable la mémoire des calculateurs électroniques. Bien qu'appartenant à l'université de Cambridge, le CRLU est constitué hors du cursus universitaire.

Le CLRU fut l'objet d'attaques sévères de la part de Bar-Hillel dans son rapport de 1960, alors que le groupe était loin de partager l'ambition de certaines équipes américaines prétendant obtenir des résultats rapides de bonne qualité et produire en masse des traductions en série. Bar-Hillel reconnaît l'aspect hautement spéculatif de l'activité du CLRU mais lui reproche le caractère désordonné des productions en mutation perpétuelle. Le thesaurus notamment désigne des entités différentes selon les auteurs, voire même change de sens pour un auteurs précis au cours du temps. Bar-Hillel critique spécifiquement la faisabilité de la TA par une méthode de langue intermédiaire, avec l'argument qu'il est plus difficile de traduire une langue naturelle dans un langage artificiel que dans une autre langue naturelle. Il critique l'utilisation de la théorie des treillis qui, selon lui, ne fait que formuler dans un symbolisme différent de choses déjà faites sans rien apporter de nouveau. Bar-Hillel conclut sa critique par en affirmant qu'il lui semble peu probable que le groupe de Cambridge amène quoi que ce soit à la TA ou à l'étude du langage.

Ceux-ci répondent en juin 1959[2] par une série de huit essais dédicacés à Yehoshua Bar-Hillel et intitulés *Essays on and in Machine Translation by the Cambridge Language Research Unit*. Ce document, associé à leurs autres publications constituent un ensemble éclairant sur les tentatives de formalisation du langage à des fins pratiques de traduction automatique.

1.2. LA TA: NAISSANCE D'UNE NOUVELLE TECHNIQUE

Si le CLRU a surtout développé des modèles théoriques, cela ne veut pas dire qu'il néglige l'expérimentation. Au contraire, le physicien E.W. Bastin, dans le chapitre intitulé*Considérations générales* de la réponse à Bar-Hillel (1959), revendique une approche empirique. Une étape décisive du développement d'une science est franchie,

[2] Le CLRU répond à une version du rapport Bar-Hillel datée de février 1959.

lorsque la combinaison d'une discussion théorique et l'examen attentif du matériel expérimental ont débouché sur la découverte d'une technique. C'est la possession d'une technique qui rend une théorie publique: elle peut être transférée d'une équipe à une autre sans changer de sens et de façon à ce que les tests expérimentaux puissent être reçus consensuellement. Bastin reproche à Bar-Hillel de sous-évaluer l'importance de la technique et des tests expérimentaux et de ne considérer la TA que comme un problème de codage complexe pour lequel le savoir nécessaire sur le langage est déjà, pour l'essentiel, disponible.

La mécanisation des tests est jugée indispensable. Sur ce point, le choix du CRLU qui peut, à première vue, paraître paradoxal, d'utiliser des cartes perforées (le procédé mécanographique Hollerith) plutôt que l'ordinateur se justifie par les capacités réduites des calculateurs électroniques de l'époque et leur complexité d'utilisation. Même si la mécanographie est plus lente que les calculateurs électroniques, les membres du CLRU considèrent sa programmation plus aisée, plus rapide et plus facile à apprendre. Cet aspect est important puisque les expérimentations en TA exigent des modifications fréquentes. Selon Needham, toujours dans le même rapport, "l'autocode" Comit mis au point par Yngve au MIT à des fins de programmation en TA, est plus adapté aux opérations de réarrangement et de substitution que la machinerie Hollerith plus adaptée à la logique et aux opérations booléennes d'intersection et union propre à la formalisation en treillis[3].

1.3. FILIATION LINGUISTIQUE

Les conceptions linguistiques du groupe de Cambridge se sont développées à partir de celles de Firth, fondateur de l'Ecole de Londres, qui a participé aux premières réunions du groupe[4]. C'est d'ailleurs au sein du groupe de Cambridge que Halliday développera sa théorie systémique, inspirée des conceptions de Firth.

Firth élabore une théorie fondée sur le concept de fonction comme principe organisateur dans la langue et sur une conception contextuelle du sens. Son approche est polysystémique: les langues sont structurées selon plusieurs systèmes, organisés en réseaux. Chaque langue a ses propres systèmes. Ce qui implique que les catégories, même si elles ont le même nom, n'ont pas le même sens selon les langues. Ainsi le nominatif dans un système à quatre cas n'a pas le même sens que le nominatif dans un système à deux cas ou encore à quatorze cas. Pour Firth comme pour Halliday, il n'y a pas de catégories universelles. Cela a des implications importantes pour la traduction.

---

[3] Il est clair que des raisons économiques ont aussi présidé à ce choix. Le CLRU a bénéficié de l'aide de la compagnie de mécanographie *ICT International Computers and Tabulators Ltd* qui leur a fournit des cartes perforées et mis à leur disposition l'Unit machinery.

[4] Les membres du groupe discutent aussi les positions de Hjemslev dont ils prétendent se démarquer (cf. *Mechanical Translation* vol.3, n°1, 1956).

1.4. CONCEPTIONS DE LA TRADUCTION

Selon la conception de Firth et du groupe de Cambridge, la traduction est un transfert de sens et non un transfert de structures. Ce qu'on appelle transfert de structures entre les langues n'est que le résultat de parallèles accidentels entre les structures de deux langues et n'est pas systématisable dans le processus de traduction. Toutefois ce parallélisme de structures augmente quand on compare des langues de plus en plus apparentées; ce qui conduit Firth à envisage la possibilité d'une interlangue conçue comme un pont entre deux langues, chacune d'elles ayant été soumise à une description adéquate. Ce projet de langue intermédiaire sera développé par Halliday (cf.§6).

Pour Masterman, ce qui va être transféré ce sont les 'bits of information' portés par des 'chunks of text' sémantiquement significatifs. Parker-Rhodes (1956) précise cette position en définissant la traduction comme une procédure qui exprime, par l'intermédiaire de la langue cible, l'information qui est véhiculée par le texte en langue source. Il désigne par information les informations grammaticales (par exemple le verbe ou l'aspect) et signale que toute information n'est pas pertinente pour toutes les langues. Ainsi en anglais, on ne peut transférer un concept verbal sans à la fois ajouter les informations repérant le moment d'énonciation et l'action. En chinois, au contraire, toute cette information est considérée comme non-pertinente. Une bonne traduction, c'est celle qui transfère toute l'information pertinente et aucune information non-pertinente.

Cette conception a conduit le CLRU à adopter une approche sémantique de la traduction[5]. Le processus de traduction est conçu en deux étapes. La première étape consiste à obtenir une traduction mot à mot par l'application d'un dictionnaire bilingue à un texte source. Lors d'une seconde étape, l'application d'un thesaurus de la langue cible sur cette sortie mot à mot est chargé de réduire les ambiguïtés sémantiques et de donner une traduction la plus idiomatique possible[6].

Non soumis à la demande sociale de traductions en série, les membres du CLRU restent à tout moment lucides et conscients de l'ampleur de la tâche. Masterman (1957) rappelle qu'en TA, la recherche en informatique est plus avancée que la recherche linguistique et que les problèmes linguistiques sont loin d'être résolus. En effet, s'ils étaient résolus, dit-elle, il n'y aurait pas grande difficulté à les programmer. Spärck-Jones (1959) dénonce la division effectuée par Bar-Hillel dans son rapport de 1960 entre FAHQMT et MMPT (man-machine partnership translation), entre traduction

---

[5]Ceci contrairement aux américains, qui, pour dépasser la traduction mot à mot dépendante de la seule analyse morphologique, ont développé des modèles fondés sur l'analyse syntaxique (cf. Léon, 1998).

[6] Bien que les auteurs signale l'existence d'un module syntaxique, celui-ci est peu explicité dans leurs premiers travaux.

complètement automatisée de grande qualité et traduction assistée, en notant que la mise en oeuvre de la traduction assistée exige aussi un travail théorique approfondi qui est loin d'être accompli.

1.5. IMPORTANCE DU CLRU PARMI LES PREMIERS EXPÉRIMENTATEURS

Seuls parmi les premiers travaux de TA, le CLRU et Igor Mel'cuk, à l'Institut de linguistique de l'Académie des Sciences de Moscou, se sont préoccupés de la construction linguistique d'une langue intermédiaire comme méthode de TA[7]. Les deux recherches ont en commun, outre l'élaboration d'une langue intermédiaire sémantique, une certaine réflexion sur la diversité des langues et sur l'irréductibilité des traits morphologiques singuliers à la TA, et le fait de développer des modèles théoriques sans pression sociale (ou de l'état) exigeant de leur part des résultats rapides. Cependant les traditions linguistiques et le contexte économico-politique. dans lesquels s'inscrivent ces deux courants de recherche sont radicalement différents et, bien qu'ayant débuté à peu près à la même époque, en 1954-55, elles se sont développées de façon totalement indépendante[8].

Pour conclure sur ce point, il faut signaler que le CLRU, bien que produisant plus de travaux théoriques que de résultats spectaculaires, n'en est pas moins reconnu par les expérimentateurs américains en TA. Il participe dès 1956 aux colloques de TA organisés par le MIT. En 1956, il reçoit un financement de la NSF. Un des premiers numéros de la revue *Mechanical Translation*, créée en 1954 par Victor Yngve, lui est entièrement consacré (vol.3, n°1, 1956); M.Masterman et A.D Booth font partie du comité de rédaction à partir de 1958. Au milieu des années 60, le CLRU expérimenta un des premiers dispositifs de TA interactifs et assistés, en collaboration avec les canadiens. A partir de 1967, les activités en TA du groupe ont diminué. Certains de ses membres, comme Martin Kay ou Yorik Wilks, ont gardé un intérêt à la TA, tout en s'orientant plus particulièrement vers l'intelligence artificielle (cf. Hutchins, 1986).

2. LE THESAURUS: UNE CERTAINE CONCEPTION PHILOSOPHIQUE DE LA FORMALISATION DU LANGAGE

Dans *Fictitious sentences in language,* Margaret Masterman (1959a) réfute l'argumentation de Bar-Hillel visant à montrer la non faisabilité de la FAHQMT (Fully-Automatised Reasonably Idiomatic Machine Translation). Pour ce faire, elle oppose sa conception philosophique de la formalisation du langage, inspirée des *Investigations*

---

[7] On peut aussi citer les travaux d'Andreev à l'Université de Léningrad, mais ceux-ci n'ont pas véritablement donné de résultats.

[8] Mel'cuk (1961) fait état des travaux du groupe de Cambridge dans un rapport de l'Académie des sciences sur les recherches en TA en-dehors de l'URSS; mais il ne les cite pas en référence de ses propres travaux· Le CLRU, quant à lui, ne mentionne pas les travaux de Mel'cuk. (Sur les travaux soviétiques sur la langue intermédiaire en TA, cf. Archaimbault, Léon, 1997).

philosophiques de Wittgenstein (1953) et sous-jacente à l'idée de thesaurus, à celle de Bar-Hillel, issue de la logique mathématique de Carnap[9]. Masterman, et à sa suite l'ensemble du groupe de Cambridge, préconise la faisabilité de la FARIMT (Fully-Automatised Reasonably Idiomatic MT). La TA doit fournir un dispositif, programmable sur machine, destiné à traduire des usages idiomatiques ou métaphoriques d'un mot (cf. Masterman, 1957).

2.1. *THE BOX WAS IN THE PEN*: CONTRE-EXEMPLE OU PHRASE FICTIVE?

Afin de réfuter l'argument de Bar-Hillel (1960) contre la faisabilité de la traduction par machine, elle réexamine les phrases données par celui-ci comme contre-exemples:

(i) the box was in the pen
(ii) the inkstand was in the pen

Dans les deux cas, dit Bar-Hillel, il faut des connaissances de sens commun pour identifier *pen* comme *parc à bébé* et non comme *plume*, connaissances dont ne dispose pas la machine. Même une méthode par thesaurus comme celle du groupe de Cambridge, utilisant les contextes, ne sera pas capable de résoudre la difficulté. Elle traduira toujours *pen* par *plume* que ce soit dans la phrase *the pen was in the inkstand* ou dans *the inkstand was in the pen*.

Pour répondre à cette argumentation, Margaret Masterman s'interroge sur le statut de ces phrases. Tout d'abord, dit-elle, elles ne posent pas le même problème que les phrases 'truquées', du type *the whiskey was good but the meat had gone bad*[10] qui, traduite en russe et retraduite en anglais, donne *the spirit is willing but the flesh is weak*. Ces phrases peuvent être facilement traitées par la méthode du thesaurus qui comprend des références croisées aux proverbes, des phraséologies, des citations bibliques etc.

Elle se propose quatre critères définissant un contre-exemple à la faisabilité de la FAHQT:

(1) la phrase doit provenir d'un texte effectif et non construit par un logicien
(2) elle doit être accompagnée par un contexte d'au moins une page
(3) elle doit passer par un test de traduisibilité par un humain
(4) elle doit démontrer une inconsistance de la méthode, et pas seulement une déficience.

Or les phrases données par Bar-Hillel ne répondent pas à ces critères. Par exemple pour le critère (2), Bar-Hillel n'a donné un contexte que pour (i), et pas pour (ii). De plus Masterman observe que ces phrases sont des transpositions:

S1 the pen is in the box	->	S1' the box is in the pen
S2 the pen is in the inkstand	->	S2' the inkstand is in the pen

[9] M.Masterman raconte qu'elle suivait les cours de Wittgenstein à Cambridge en 1933, mais que celui-ci l'en a chassée, l'accusant de n'y rien comprendre.

[10] Selon Hutchins (1995) l'origine de cette phrase date probablement de 1956. Ironie du sort, cette phrase, donnée comme exemple d'erreur de traduction humaine, est devenue une des phrases mythiques destinées à montrer les limites de la traduction automatique.

Cette transposition peut s'analyser de la façon suivante: *phrases normales en anglais -> phrases en anglais normal.* Le problème est de savoir si cette propriété de transposition est une caractéristique essentielle que devraient posséder tous les ensembles de phrases mettant en défaut tout type de TA par thesaurus, ou si c'est seulement un cas particulier.

2.2. DEUX CONCEPTIONS PHILOSOPHIQUES DE LA FORMALISATION DU LANGAGE

Sur le plan philosophique, c'est entre les deux sens de *normal* que réside le gouffre qui sépare les deux conceptions du langage. L'une, favorisant l'interprétation *normal sentences in English*, s'appuie sur une conception logico-mathématique des transformations et utilise le sens mathématique de *normal*. C'est le cas des phrases S' qui sont bien formées, ont un sens, bien que ce soit des permutations, des parenthésages ou des répétitions de phrases en ordre *normal*. Toutes ces opérations, permutation, parenthésage, répétition, supposent une conception du langage selon laquelle l'ordre de l'ensemble total des phrases peut être défini en utilisant un calcul (par exemple la logique combinatoire). C'est la conception du langage que préconisent Bar-Hillel ou Chomsky.

Si au contraire *normal* signifie *ce qui est effectivement dit,* comme dans *sentences in normal English*, cette normalité est socio-scientifique et non plus mathématique. Le langage est pris comme une totalité et ce n'est que dans un second temps que l'on y reconnaît une éventuelle structure mathématique. Alors que pour Bar-Hillel, l'ordre est inverse: il considère d'abord la formalisation mathématique comme une totalité et tente ensuite de voir combien de langues naturelles peuvent s'y adapter.

Dans une note, Masterman soutient que la position de Bar-Hillel est incompatible avec l'automatisation du langage. Elle prend l'exemple de la classification intuitive des concepts avec références croisées telle qu'elle peut être effectuée par un bon bibliothécaire. Celui-ci vise à organiser les concepts sémantiques et à les encoder de façon à fournir une échelle de pertinence pour chaque recherche donnée; or s'il écoute les conseils logico-mathématiques de Bar-Hillel, il inférera, de façon tout à fait correcte, qu'une telle classification est impossible. Enfin, Masterman considère que la seconde école de pensée devrait davantage intéresser la linguistique structurale et descriptive dans la mesure où les linguistes s'intéressent au langage comme il est et non en fonction de sa capacité d'adaptation à un système logique.

2.3. LE THESAURUS: UN PROJET WITTGENSTEINIEN

La méthode par thesaurus est, selon Masterman, directement inspirée des intuitions de Wittgenstein (1953), en ce qu'elle justifie la définition du sens d'un mot à partir de ses contextes:

L'unité de la logique pour l'étude du langage n'est pas le mot (le terme d'Aristote) et encore moins la phrase (la proposition de Russell) mais le contexte d'un mot. Un concept du langage peut être considéré comme une figure, au sens de la Gestalttheorie, à savoir qu'on le verra différemment selon l'aspect à travers lequel on le regarde. Chaque aspect est représentable par un contexte. Les contextes d'un mot ne peuvent, par définition, être distingués par les méthodes normales. Ils ne peuvent l'être que si on leur fournit une analogie. Chaque analogie montre le concept original, c'est-à-dire l'ensemble imaginable des contextes distincts d'un mot, sous un nouvel aspect.

Ainsi les phrases S1', S2' ne sont pas de simples transpositions. Elles concrétisent de nouveaux contextes. Dans le langage ordinaire, ces phrases sont fictives jusqu'à ce qu'elles soient acceptées comme 'nouvelles phrases', c'est-à-dire des phrases qui créent de nouveaux contextes dans le langage. De ce point de vue les phrases sont ritualisées: elles doivent s'entourer des 'ornements contextuels' pour pouvoir accéder la première fois à 'l'édifice sacré' du langage (cf. Masterman, 1959a: 26).

2.4. INFINITÉ DES CONTEXTES D'UN MOT, LIMITE DES SITUATIONS EXTRA-LINGUISTIQUES

Masterman (1959b) définit un thesaurus comme un système de mots (têtes) classées selon un ensemble de contextes. Elle rappelle que, comme l'usage des mots change continuellement, l'ensemble total des contextes est infini. L'hypothèse fondamentale qui étaye la faisabilité d'un thesaurus est que, bien que l'ensemble des usages possibles des mots dans une langue soit infini, le nombre de situations extra-linguistiques primaires nécessaires pour communiquer est fini. Contrairement à ce qu'on pourrait penser, nous ne nous référons pas à une nouvelle situation extra-linguistique à chaque fois que nous utilisons un nouvel usage d'un mot. Nous empilons les synonymes, élargissons les références en intégrant de nouveaux aspects à partir du stock de situations extra-linguistiques de base que nous avons déjà.

Cette hypothèse à des conséquences importantes pour la TA La traduction, comme la communication, n'est possible, que si les deux populations et les deux cultures correspondant aux langues envisagées, même si elles sont très différentes, partagent un stock commun de contextes extra-linguistiques. Pour élaborer un thesaurus, il est important de savoir que la totalité des contextes d'une langue forment un continuum alors même que l'ensemble des contextes d'un mot forment un ensemble discret.

3. UN FORMALISME DE REPRÉSENTATION DES CONNAISSANCES

Revendiquée par l'ensemble des membres du groupe de Cambridge, la notion de thesaurus revêt des sens différents selon les divers concepteurs. C'est d'ailleurs, on l'a vu, un des reproches formulés par Bar-Hillel à l'égard du CLRU. Pourtant, on peut soutenir que, dictionnaire ou langue intermédiaire, le thesaurus constitue une véritable

tentative de représentation sémantique formalisée, destinée à résoudre les problèmes de polysémie à l'aide du contexte.

3.1. UN OU DES THESAURUS?

Masterman elle-même modifie au cours du temps sa définition du thesaurus. Dans un article de 1957, thesaurus et langue intermédiaire sont conçus comme deux projets différents et complémentaires[11]. Pour l'unification des deux méthodes, consistant soit à transformer le thesaurus en langue intermédiaire, soit à inclure la syntaxe dans le thesaurus, il faut avoir recours aux linguistes qui auront pour tâche de faire de la syntaxe dans la sémantique ou de la sémantique dans la syntaxe. Pour le moment, dit-elle, cette unification est prématurée.

Deux ans plus tard, Masterman (1959b) admet que le CLRU développe une langue intermédiaire à base de thesaurus. En outre, elle répond aux reproches formulés par Bar-Hillel en reconnaissant que le terme de thesaurus recouvre plusieurs entités: les thesaurus naturels, les thesaurus terminologiques à usage bibliographique[12] et les thesaurus à vocation de langue intermédiaire, comme Nude. Les thesaurus naturels, comme le Roget's thesaurus, utilisé par le groupe dans ses premières expériences de TA, ou comme certains regroupements lexicaux en ancien chinois, en sanscrit ou en sumérien, sont incomplets et peu rigoureux. Leur objectif qui consiste à accroître le savoir des lecteurs, n'est pas celui de la TA. Les têtes sont classées selon une hiérarchie arborescente simple alors qu'une hiérarchie multiple d'archi-têtes (archeheads) est nécessaire, d'où l'utilisation de structures en treillis.

3.2. VERS UN LANGAGE DE REPRÉSENTATION SÉMANTIQUE UNIVERSEL

Masterman (1957, 1959b) se pose la question de savoir quels sont les critères qui définissent les archi-têtes, catégories sémantiques de base du thesaurus. De façon plus générale, Masterman se pose la question de la définition du mot, auquel, dit-elle, personne ne s'est encore attaqué de façon intellectuellement satisfaisante.

Le problème, c'est de trouver des concepts sémantiques qui aient du sens sans pour autant être des mots appartenant à une langue donnée. Masterman caractérise ces archi-têtes comme devant être 'juste en-dessous de la ligne de signification'. Ce sont des mots qui n'ont pas besoin d'exister dans toutes les langues. Mais ils doivent ressembler à des mots existant dans n'importe quelle langue. Ainsi l'archi-tête TRUE! doit ressembler au mot anglais *true*; ou du moins TRUE! doit plus ressembler à *true* qu'à *please*. Le choix

---

[11] Elle oppose le thesaurus, "tentative de mathématisation de Platon", à la langue universelle de Richens, "tentative de mathématisation d'Aristote". "Vieille opposition, dit-elle entre nominalistes et réalistes qui prend un tour nouveau, étrange, fascinant et ésotérique." (1957:38).

[12] Le CLRU a mis au point un thesaurus terminologique dans le cadre de son programme Library Retrieval Scheme pour retrouver des références bibliographiques à partir de mots clefs.

des archi-têtes peut être arbitraire puisque ce ne sont pas des entités empiriques mais une structure mathématique codée, une structure en treillis, utilisée comme véhicule intermédiaire de traduction dans l'algorithme. Mais il ne doit pas y avoir de recouvrement entre les différentes archi-têtes. Enfin il ne faut pas que la complexité des entrées augmente avec le nombre de langues traitées.

A travers la définition des archi-têtes, ce qui est souligné c'est la difficulté d'abstraction du sens et de son attribution à des catégories dont les noms doivent être apparentées au langage ordinaire. Cette construction d'un métalangage sémantique est jugée d'autant plus ardue qu'elle est confiée à des humains. D'où les risques de choix arbitraire ou de recouvrement. La justification d'une telle opération est d'ordre pratique: c'est l'utilisation en machine à des fins de traduction[13].

3.3. UN EXEMPLE DE TRADUCTION À L'AIDE DU THESAURUS: UN EXTRAIT DES *GÉORGIQUES* DE VIRGILE

Dans leur rapport de 1959, les membres du groupe de Cambridge présentent la traduction de la phrase *agricola incurvo terram dimovit aratro* (le laboureur a fendu la terre avec une charrue incurvée), extraite des *Géorgiques* de Virgile.

On ne retiendra ici que les deux premières étapes de la procédure de traduction, appliquées à un fragment de la phrase latine. Le thesaurus (il s'agit ici du Roget's thesaurus) est utilisé comme langue intermédiaire. Il est organisé selon un ensemble de têtes accompagnées d'une liste de synonymes. Il est formalisé en treillis, de façon à ce que chaque mot soit localisé selon son absence ou présence sous telle ou telle tête.

Une première phase de la procédure consiste à appliquer le dictionnaire latin-langue intermédiaire sur le texte d'entrée segmenté en 'chunks', c'est-à-dire en morceaux sémantiquement significatifs du texte. Ces 'chunks' sont alors remplacés par des têtes appartenant au thesaurus *agri-* et *-col-* pour *agricola*.

Soit les têtes de thesaurus suivantes et leur liste de synonymes:

agri-
181 region
189 abode
371 agriculture
780 property

-col-
188 inhabitant
186 presence
758 consignee
371 agriculture
342 land
876 commonalty

La seconde étape consiste à fournir une première traduction dite 'sémantique' en appliquant le thesaurus selon une procédure d'intersection. Les mots de la liste de

[13] Ces questions, précisons-le, semblent avoir disparu des préoccupations des chercheurs en intelligence artificielle des années 70-80 qui ne se posent guère cette question lorsqu'ils construisent des langage de représentation sémantique de type plans, cadres, scénarios etc. Souci davantage propre aux philosophes qu'aux informaticiens semble-t-il.

chaque tête retenue pour un 'chunk' sont comparés à chaque mot des listes des autres têtes retenues pour ce 'chunk'. Tout mot apparaissant plus de deux fois dans une phrase est retenu comme sortie. Dans notre exemple, le contexte *agriculture* appartenant aux têtes *agri-* et *-col* est retenu. Les têtes seront ensuite traduites dans la langue cible moyennant certains réarrangements syntaxiques[14].

4. DES LANGUES NATURELLES À LA LANGUE INTERMÉDIAIRE: LE MECHANICAL PIDGIN

Une des originalités du groupe de Cambridge consiste à considérer les langues naturelles du point de vue de leurs potentialités de formalisation. Ainsi ils s'intéressent à certaines langues parce qu'elles comportent des caractéristiques qui pourraient être les traits définissant une langue intermédiaire.

Le chinois, les pidgins et le latin attirent particulièrement leur attention. Ce choix de langues ne laisse pas de surprendre, surtout si l'on considère que les groupes américains de l'époque, et plus tard les Français, focalisaient leurs recherches sur la traduction du russe[15], et que les Soviétiques avaient pour arrière-plan la traduction des multiples langues de l'union. Ce n'est donc pas la demande sociale qui préside à ce choix, mais plutôt l'intérêt propre de ces langues au nom d'hypothèses philosophiques et linguistiques précises. De même, le choix d'une méthode par langue intermédiaire n'est pas lié seulement à des raisons économiques de traduction multilingue[16], il tient à une conception de la TA où la perspective d'automatisation ouvre sur la possibilité de traduire toutes les langues.

4.1. LES LANGUES DE LA TA: CHINOIS, LATIN, PIDGINS

Les logiciens du groupe de Cambridge considèrent que le chinois est une langue favorable à la TA puisqu'il est logiquement moins varié que les autres et construit sur une unité le 'tzu', concept plus fondamental que le mot .Certains, comme Bronowski (1956), établissent une analogie entre la sortie mot à mot et la poésie chinoise. Dans les deux cas, dit-il, c'est le lecteur qui fournit les connexions non explicitées dans le texte.

En ce qui concerne le latin, il permet de tester l'importance de l'ordre des mots sur la traduction mot à mot. Des essais comparatifs ont été faits entre un texte en latin classique, le premier paragraphe du premier livre "La Guerre des Gaules" de César et

---

[14] Dans certaines versions du thesaurus, des informations grammaticales sont associées aux "chunks" de façon à faciliter le choix d'une tête donnée. Par exemple, Parker-Rhodes (1956) propose qu'à chaque chunk soit associé un "word class indicator".

[15] En-dehors des raisons politiques et militaires, renforcées dans un contexte de guerre froide, les Américains étaient persuadés que les Soviétiques étaient très en avance sur eux sur le plan scientifique, non seulement dans le domaine de l'aérospatiale (le premier spoutnik date de 1957) mais aussi dans le domaine de la TA.

[16] L'argument d'économie d'algorithmes est évoqué par ce groupe comme par tous les groupes de TA qui ont envisagé une méthode par langue intermédiaire (cf. Kay, 1959).

un texte scientifique du 17ème siècle en latin "Les Principia Mathematica" de Newton (livre I, proposition LIX, theorème XXII) dont l'ordre des mots est plus proche de celui l'anglais (cf. Masterman 1967).

M.Masterman (1959a:7) souligne la parenté du texte de César avec un texte scientifique en évoquant la façon dont César polit son style "de façon perverse pour impressionner le pentagone comme le font les scientifiques contemporains". De plus, les particularités de ce texte lui permettent d'argumenter contre Bar-Hillel. Il contient des transpositions d'ordre des mots et des ellipses de façon à constituer un réel défi à la TA, contrairement aux phrases proposées par Bar-Hillel. Masterman soutient que ce paragraphe de "La Guerre des Gaules", qui n'a pu être traduit par des méthodes de TA ordinaires, a pu l'être grâce à une méthode par thésaurus.

Quant aux pidgins, ils fascinent beaucoup le groupe de Cambridge, et ceci à plusieurs titres. Ce sont des formes intermédiaires entre les deux langues dont ils sont composés. Bastin (1956), par exemple, se demande dans quelle mesure le pidgin des antilles peut ou non être considéré comme un pont entre le chinois et l'anglais. Par ailleurs, ce sont des langues qui, comme le langage des enfants, sont susceptibles d'apprendre quelque chose de fondamental sur le commencement de la pensée. Telle est du moins l'hypothèse soutenue par Wittgenstein dans le *Blue Book* et que cite Masterman (1959a), justifiant ainsi en partie l'engouement du groupe pour les pidgins.

Dès 1956, Richens, baptise "Mechanical Pidgin" la sortie mot à mot d'une traduction par application automatique d'un dictionnaire bilingue à un texte en langue source[17]. La référence au chinois est, là encore, invoquée. Masterman (1967) signale que le lexique d'une sortie en Mechanical Pidgin est celui de l'anglais et que sa structure rappelle celle du chinois.

4.2. UN LANGAGE FORMEL PRODUIT PAR LA MACHINE

Il est tout à fait intéressant de noter que les membres du CLRU décident de donner un statut à cette sortie mot à mot en langue cible. Masterman (1967) rappelle les expériences faites à la fin des années 50 sur le Mechanical Pidgin considéré comme un langage logique de base. Parce que produit par la machine, c'est déjà un langage formel et les variations introduites par la différence des langues sources sont négligeables.

Très curieusement, les membres du groupe de Cambridge n'hésitent pas à renforcer la cohérence conférée au Mechanical Pidgin par l'apparition de régularités attribuées, à l'époque, aux langues naturelles. Selon Masterman (1967) en effet, le Mechanical Pidgin présente les traits statistiques mis au jour par Zipf pour les langues naturelles, à savoir la division en deux groupes, les mots de contenu, et les mots outils (les pidgin

[17] Richens et Booth ont fourni des traductions en Mechanical Pidgin de vingt phrases rédigées en vingt langues sources différentes prises au hasard dans la littérature de la génétique végétale (cf. Locke et Booth 1955:36-37).

variables ci-dessous). Ainsi on peut dire que le Mechanical Pidgin a un statut de langue intermédiaire à double titre. Dans le processus de TA, tout d'abord, il a une position intermédiaire puisqu'il constitue une première étape de la traduction en langue cible. Sur cette sortie mot à mot sera appliqué le thesaurus de façon à obtenir une sortie dont la qualité idiomatique et l'intelligibilité sera améliorée, par la réduction des ambiguïtés sémantiques.

Le Mechanical Pidgin, par ailleurs, constitue une première étape vers la formalisation non fondée sur la logique mathématique, mais sur les invariants des langues.

4.3. UN EXEMPLE DE TRADUCTION EN MECHANICAL PIDGIN

Cet exemple est présenté dans Masterman (1967). Le texte source en latin est segmenté en 'chunks' sémantiquement significatifs (* sépare deux chunks dans un mot).

latin: possibil*e est, at non expert*um, omn*es speci*es eiusdem generis ab eadem speci*e ort*um trax*isse
anglais: it is possible, though not proved, that all species of the same genus have been derived from the same species
mechanical pidgin: possible z is however not prove/lacking z all m species / appearance same g genus / son-in-law z from same species / appearance o arise z draw p

Le Mechanical Pidgin comporte des marqueurs (z g o et p): z désigne les formes grammaticalement non spécifiées; g genitif; o oblique; p passé; / alternative au niveau du lexique; + connecte des mots qui forment un syntagme de sortie et - connecte les racines et les flexions. La sortie en Mechanical Pidgin est transformée en une sortie intelligible à l'aide de plusieurs opérations: éliminer les z qui ne servent à rien dans la traduction, transformer les marqueurs en pidgin variables (les auteurs appellent pidginisation cette transformation) et créer des expressions.

g est 'pidginisé' dans la langue cible (l'anglais) en *-ish* quand il est postposé, et en *of the* quand il est antéposé.
p est 'pidginisé' par *-ed* quand il est postposé, et en *did* quand il est antéposé.
species devient *form* et *genus* devient *family* (*son-in-law* ne peut appartenir à un dictionnaire de génétique végétale)

Les expressions sont les 'bits of information' nécessaires à la compréhension d'un texte. Le 'bit of information' véhiculé par l'expression *it + is +possible* qui remplace *possibile est* est qu'un être humain fait l'hypothèse que ce qui suit est possiblement vrai. Les expressions peuvent être classés selon le 'bit of information' qu'ils véhiculent. Cette classification est la base de la construction d'un thesaurus pour la TA. On obtient enfin une sortie en pidgin "intelligible":
it + is + possible however not + prove all form of + the+ same-ish family from same form + draw-to + have

Ce statut de langue intermédiaire accordée à ces sorties mot à mot avec variables, permet de ne pas y voir seulement une traduction fruste et quasiment illisible issue

d'une prétention naïve, mais d'y reconnaître une véritable tentative de formalisation. Il est intéressant de voir comment la formalisation crée un objet qui hésite entre les propriétés des langues naturelles et celles d'un langage formel[18]. Cet objet ambigu est sans doute une des caractéristiques de la traduction automatique. Celle-ci suppose une analyse de la langue source et une synthèse dans la langue cible. Le résultat est donc nécessairement un texte dans une langue naturelle.

Le Mechanical Pidgin fut explicitement abandonné dans un texte de Masterman publié dans le dernier recueil qui clôt cette première phase des recherches en TA un an après le rapport de l'ALPAC (Booth, 1967). Elle reconnaît que les objectifs que se donnaient le CLRU à travers le Mechanical Pidgin n'ont pas pu être atteints: "From the experimental we had done we considered that Mechanical Pidgin Translation had been tested to destruction." (1967:224). Le groupe s'est en effet trouvé rapidement devant une impasse: les ambiguïtés sémantiques pouvaient certes être réduites grâce à la création d'un grand nombre d'expression particuliers mais c'était contraire avec le projet de généraliser l'information sémantique.

5. DU MECHANICAL PIDGIN AU NUDE: LANGUE INTERMÉDIAIRE ET LANGUE UNIVERSELLE

A partir du Mechanical Pidgin, Richens (1956a), a le projet de construire une langue unverselle dans laquelle les particularités structurales de la langue source sont supprimées et qui se compose d'un réseau sémantique d'idées nues, les 'naked ideas', d'où le nom de Nude donné à cette langue. Pour Richens en effet, le réseau sémantique est ce qui est invariant durant la traduction[19].

Nude s'inspire du Mechanical Pidgin et ses 50 éléments primitifs sont très proches des marqueurs 'pidginisés' du Mechanical Pidgin. Mais contrairement au Mechanical Pidgin qui conserve une certaine parenté avec les langues naturelles - son lexique notamment est celui de l'anglais - Nude renoue avec la tradition des langues universelles des 17ème et 18ème siècles[20]. Débarrassée de toutes les particularités lexicales et syntaxiques des langues naturelles, Nude se veut une langue intermédiaire algébrique purement notationnelle. Ainsi, Richens (1956b) précise que chacun des cinquante éléments de Nude dénote une idée de base telle que 'pluralité', 'animal', 'négation' et qu'il ne peut comporter qu'une seule lettre. La syntaxe de Nude consiste en deux connecteurs et une convention de parenthésage.(Masterman, 1959b). Le

---

[18] Sur la distinction entre langues naturelles et langages formels voir Auroux (1998).

[19] Toutefois le problème principal consiste précisément à extraire ces réseaux sémantiques des textes de base et les auteurs reconnaissent qu'aucune procédure automatique générale n'a pu être élaborée à cette fin.

[20] A travers l'utilisation du Roget's thesaurus dont la première version publiée en 1852 s'inspire directement des travaux de Wilkins au 17ème siècle, les membres du CLRU marquent également leur filiation aux langues universelles.

premier connecteur, les deux points, représente une relation d'ajout à l'élément principal. Le slash est un connecteur verbal non commutatif qui représente la relation de sujet au verbe ou de verbe à objet. Les parenthèses relient un ajout avec son élément principal, un objet à un verbe qui précède et un sujet avec le prédicat. L'exemple de Bar-Hillel *in the inkstand* devient en Nude:

IN! | (man /use) / (IN!: thing) -INKSTAND

Cette langue intermédiaire semble avoir eu un certain succès dans le groupe de Cambridge. Cette tentative de langue universelle a même donné lieu à des pratiques de 'fous de la langue'. M.Masterman (1959b) relate que les auteurs de Nude avaient prévu de résoudre par le marquage les problèmes de polysémie, en particulier le problème des mots qui peuvent avoir à la fois un sens littéral et un sens figuré. Ainsi UP! au sens figuré (élévation dans la société) est marqué par rapport à UP! au sens littéral (élévation dans l'espace) de la façon suivante : (not same :up) (up: (part:folk)):man.

Masterman poursuit en disant que lorsque les membres du CLRU se sont mis à parler Nude, à faire des plaisanteries en Nude, à écrire des lettres en Nude etc., les formules, notamment (not same), sont devenus peu à peu des expressions idiomatiques. Elle en conclut que Nude a eu tendance à se développer de façon idiomatique comme un pidgin naturel et non comme un pidgin artificiel dont l'invariance doit être requise. En note, elle précise qu'en réaction, la langue intermédiaire Lattite, une des versions programmées de Nude, a été émaillée de marqueurs métaphoriques dans l'espoir que si il y en avait assez, les gens n'allaient pas les utiliser eux-mêmes métaphoriquement[21].

Au-delà du caractère ludique de cette utilisation de Nude devenue langue naturelle ou langue auxiliaire, les membres du CLRU découvrent plusieurs problèmes qui se posent pour la TA et l'automatisation du langage. D'une part, ils se rendent compte que la production d'artefacts, bien qu'inévitable dès qu'on a affaire à des intervenants humains, est très importante. On ne peut en effet empêcher les rédacteurs de dictionnaire de développer les sens idiomatiques. D'autre part, se pose le problème de la multiplicité des expressions idiomatiques et de leur non correspondance d'une langue à l'autre, multiplicité qui constitue une limite de la langue intermédiaire. Les membres du groupe s'aperçoivent qu'il faut un dictionnaire spécial pour toutes les locutions, expressions idiomatiques et autres phraséologies. Ce qui réduit à zéro l'efficacité d'une méthode par langue intermédiaire. Par ailleurs cela pose des problèmes technologiques insurmontables liées aux possibilités réduites des machines[22], d'autant plus, rappelons-

---

[21] On est ici en présence de la tension provoquée par la mise en oeuvre d'un des traits du processus de formalisation, relevé par Auroux (1998), selon lequel toute formalisation suppose un blocage, une perte et un éloignement de la vie.

[22] Rares sont en effet les équipes de TA qui, à l'époque, envisagent de construire des dictionnaires en extension. Seule l'équipe dirigée par Reifler à l'Université de Washington, à partir d'une conception de la TA, plaçant le lexique au centre du dispositif, envisage la confection d'un dictionnaire bilingue anglais-allemand en extension. Il est important de noter que cette position est étayée sur un choix technologique

le, que le CLRU a fait le choix d'utiliser la mécanographie dont la capacité à traiter des données importantes est encore plus limitée que celle des calculateurs électroniques.
Ces expériences et ces écueils font apparaître que le CLRU s'est engagé dans une tentative de formalisation du langage comme totalité plutôt que de formalisation des langues. A travers l'utilisation de Nude comme langue auxiliaire, ils se sont trouvés confrontés à la non transparence du langage et à la limite entre langue naturelle, destinée à être parlée et langue artificielle, outil de formalisation pour la machine.

6. UNE LANGUE INTERMÉDIAIRE FONDÉE SUR UNE GRAMMAIRE DES CONTEXTES EXTRA-LINGUISTIQUES (M.A.K. HALLIDAY)

Halliday est un linguiste, spécialiste de chinois. Il partage avec Margaret Masterman et l'ensemble du groupe l'idée que ce n'est pas la grammaire qui est au centre du processus de traduction automatique, mais le lexique et le lexique en contexte. Il envisage de construire une langue intermédiaire, fondée sur une description universelle de la syntaxe et se référant aux catégories de la grammaire contextuelle. Cette interlangue grammaticale n'est pas une langue universelle; c'est un ensemble de systèmes de relations grammaticales identifiées dans le cadre d'une grammaire de contexte. Ce qui est tout à fait particulier à son approche c'est que le contexte ainsi invoqué est extra-linguistique et non intra-linguistique.
Halliday (1956) rejette la méthode par transfert[23], adéquate lorsqu'il s'agit de traduire deux langues, mais insatisfaisante dès qu'on veut traiter un nombre quelconque de langues. Par ailleurs, il considère que la mise au point de catégories universelles pour la traduction, même pour un groupe limité de langues, demanderait trop de temps. En pratique, dit-il, il faut trouver des compromis: faire une analyse descriptive de chaque langue qui soit à la fois autonome et orientée vers les besoins de traduction. Le problème principal consiste à trouver les points de recouvrement optimum entre langue source et langue cible. Et comme la traduction, bien qu'elle soit une relation mutuelle, est avant tout un processus unilatéral, ce qui est important c'est le choix des formes de la langue cible. Selon Halliday, le traitement autonome de la langue cible réduit l'indétermination (the loss of determination) de la traduction. Halliday reformule les deux étapes de traduction adoptée unanimement par tous les membres du CLRU en considérant que la première étape qui fournit une sortie mot à mot exploite le sens littéral des mots (primary meaning), alors que la seconde étape effectue la traduction à partir d'un système déterminé contextuellement.

---

de mémoire 'illimitée', la mémoire photoscopique de King, qui devait constituer l'élément principal de la machine à traduire MarkI (cf. Léon 1998).

[23] Dans la méthode dite "grammaire de transfert", mise au point par Yngve (1957) au MIT, le processus de traduction comporte quatre étapes: reconnaissance de la structure du texte d'entrée, transcription de cette structure dans un langage de transition spécifique, transfert de ce langage dans le langage de la langue cible, construction du texte de sortie.

Une description universelle des langues pour la TA ne peut se développer qu'en séparant rigoureusement le traitement des traits monolingues et celui des traits interlangues. En outre il faut distinguer les "chunks" qui peuvent être identifiés par l'analyse grammaticale, les opérateurs, et ceux qui, comme les argument exigent des informations lexicales pour être identifiés. Halliday précise que cette distinction opérateurs / arguments est une distinction arbitraire exigée par la TA et non motivée linguistiquement.

Les opérateurs sont identifiés selon une fonction oui/non. Autrement dit, pour chaque opérateur, l'analyste se pose un certain nombre de questions extrêmement simples auxquelles il peut répondre sans hésitation par oui/non ou bien les deux/aucune. Ces questions se réfèrent aux êtres dans le monde (division de l'humanité en deux sexes, des êtres en animés/inanimés) et à leur coordonnées spatio-temporelles. Chaque terme est donc référé à une description extra-linguistique et non à son système interne dans la langue d'où il est issu.

Masterman (1957) présente la méthode de Halliday en donnant l'exemple de *la,* en français, difficile à désambiguïser automatiquement puiqu'il désigne soit un article défini feminin, soit un pronom féminin à l'accusatif. Pour construire la langue intermédiaire, il faut se poser non pas la question 'est-ce que *la* appartient à un système de genres' parce qu'on sait que les genres n'ont pas de correspondants dans toutes les langues, mais 'est-ce que *la* peut dire quelque chose sur le sexe?'. En posant une telle question on remplace une référence à un contexte intra-linguistique (celui du français) par un contexte extra-linguistique beaucoup plus stable: la division de l'humanité en deux sexes. Sa méthode de questions oui/non, destinée à identifier les opérateurs, peut être utilisée comme méthode de classification des synonymes sous les entrées du thésaurus. D'où la convergence établie entre interlangue et thesaurus. Par ailleurs le caractère dichotomique du thesaurus et de la méthode de classification se prête parfaitement à la méthode des treillis.

Pour Halliday il s'agit d'appliquer la méthode de la linguistique descriptive à l'analyse des contextes extra-linguistiques. En prenant au sérieux l'analogie entre système intra-linguistique et extra-linguistique, et en traitant le premier comme une extension du second, Halliday prétend fonder, à des fins pratiques, une méthode d'analyse grammaticale universelle.

Selon Masterman (1957), le projet de Halliday consiste à élaborer un thesaurus syntaxique. Halliday (1956) propose de faire figurer dans le thesaurus en plus des groupements purement lexicaux (noms ou verbes), des groupements partiellement grammaticaux. Il présente notamment une classification des prépositions en anglais, considérées comme opérateurs partiels, et qui peuvent être traitées comme un simple groupement lexical.

La question qui se pose avec le modèle de Halliday c'est que l'automatisation de la traduction implique un abandon d'une description purement linguistique au profit de la recherche de catégories universelles. On voit ainsi apparaître trois types de catégories, des catégories linguistiques, des catégories du langage, qui, parce qu'elle doivent être universelles, ne peuvent être grammaticales, et des catégories, comme les opérateurs et les arguments, qui sont créées artificiellement à des fins de TA et que l'on pourrait qualifier de formelles.

7. LA TA, UN IMPÉRATIF PRATIQUE POUR LES LINGUISTES (MARTIN KAY)

7.1. TA ET FORMALISATION MATHÉMATIQUE LIMITÉE

Martin Kay fait partie du groupe à partir de 1958. Dans son article de 1959, intégré à la réponse à Bar-Hillel, il se propose d'établir la contribution de la linguistique à la TA, et partant de définir les limites de la formalisation.

Il critique le rapport Bar-Hillel en citant Hockett, partageant avec celui-ci l'idée que l'intérêt d'une description grammaticale dépend en partie de l'usage qu'on veut en faire. Notamment il reconnaît avec Bar-Hillel que la notion de 'sentencehood' est extrêmement difficile à établir[24], mais il ajoute que, comme cette notion a peu d'intérêt pour la TA, il est inutile d'essayer de la définir.

La formalisation ne peut être totale et porter sur tous les aspects du langage, mais elle doit être effectuée en fonction d'un objectif précis. Kay reprend l'analogie faite par Martin Joos entre formalisation en linguistique et cartes géographiques. Les cartes ne peuvent prétendre représenter en taille réelle tous les éléments (arbres, collines etc.) d'un lieu géographique donné. Les éléments représentés doivent l'être en fonction d'une utilisation particulière. De même, pour une application en TA, on n'a pas besoin de toute l'information qu'il est possible d'extraire d'une langue donnée. En effet, la TA, contrairement à la linguistique, ne s'intéresse pas à ce qui commun entre le texte à traduire et les autres textes de la même langue. On a surtout besoin de savoir ce qui est différent dans ce texte; ce qui fait que ce texte est ce texte et pas un autre et ce qu'il essaye de dire (Kay, 1959).

Masterman (1959b) cite Kay pour dire qu'à partir du moment où on pose la question la plus fondamentale qui soit à savoir 'qu'est-ce qui est dit ici?' on doit trouver d'autres dispositifs de description (apparatus) que ceux fournis par la linguistique.

En plus d'être limitée et d'avoir des objectifs précis la formalisation doit être mathématique. C'est la seule condition, ajoute Kay, pour que la traduction puisse être

---

[24] Spark-Jones (1959a) montre les changements de position de Bar-Hillel concernant la notion de phrase. Dans son texte 'a quasi-arithmetical notation for syntactic description', il considère qu'une méthode fondée sur une analyse en constituants immédiats, donc contigus, est suffisante pour déterminer le caractère syntaxique d'une phrase donnée. Puis il doute que même une telle analyse complétée par des opérations transformationnelles qui prennent en compte des unités non contigues mais syntaxiquement associées peut analyser toutes les structures de phrases d'une langue.

automatisée. Il critique ainsi la position de certains qui considèrent que la logique doit être externe à la formalisation du langage.

7.2. LANGUE INTERMÉDIAIRE ET SÉPARATION ENTRE GRAMMAIRE ET LEXIQUE

D'un point de vue pratique, dans le processus de TA, il y a complémentarité entre les routines syntaxique et sémantique. Ainsi la procédure sémantique sera mise en oeuvre au cas où la procédure syntaxique échoue à fournir une analyse unique et à lever les ambiguïtés. Martin Kay souligne que cette position s'oppose à la position communément admise chez les linguistes, pour qui le contenu informationnel des langues ne peut être formalisé. Cette attitude, dit-il, tient aux mêmes raisons pour lesquelles la linguistique structurale a abandonné la sémantique.

Pour la TA, une formalisation du langage doit rendre compte de la séparation entre grammaire et lexique, indispensable à une TA de qualité. Kay considère d'ailleurs que la grammaire transformationnelle de Harris et de Chomsky est adaptée à la TA parce qu'elle se situe précisément entre grammaire et lexique, et qu'elle est particulièrement adaptée à la synthèse du texte en langue cible

Comme ses collègues du groupe de Cambridge, Martin Kay considère que la langue intermédiaire est une étape capitale dans le processus de TA. Toutefois, contrairement aux autres, et en particulier à Halliday, il n'écarte pas la grammaire. La langue intermédiaire est le résultat d'un calcul formel. Elle doit correspondre à la formalisation de la réduction de l'information grammaticale et lexicale à une forme commune. La langue intermédiaire est le produit d'un calcul de l'information sémantique à partir des systèmes lexicaux et grammaticaux de la langue source. Le thesaurus, intégrant des entrées lexicales et grammaticales est la langue intermédiaire la plus utile pour la TA. Il considère que la synthèse en langue cible doit être formalisée différemment de l'analyse de la langue source et qu'il faut privilégiée la synthèse.

Il prend l'exemple de l'aspect en russe. L'analyse en constituants immédiats de *I finished reading the book before dinner* et sa traduction en russe fournit une analyse formelle non équivoque de la phrase avec le parenthésage suivant:

((I (finished reading)) (the book))(before dinner)
(Ia (procital knigi)) (do objeda)

Il y a un problème, dit-il, parce qu'on obtient deux mots russes, et non deux syntagmes comme en anglais. En revanche, si l'on considère que la sortie peut être une transformation de:

(Ia (procital perfective)(knigi)) (do objeda)

La correspondance entre l'anglais 'finished' et le perfectif russe est ainsi établie.Les traits grammaticaux comme l'aspect, le mode, les cas sont intégrés dans le thesaurus au même titre que les items lexicaux. Cet exemple montre à la fois qu'une analyse en constituants immédiats n'est pas un formalisme adapté à la traduction et qu'il faut intégrer des traits grammaticaux dans le thesaurus au même titre que des traits lexicaux.
Le travail de Kay constitue donc une réflexion sur la place de la linguistique dans la TA à partir de l'idée que la validité d'une description linguistique doit toujours être liée à son utilité, que cette linguistique n'est pas la même selon les objectifs poursuivis. La TA nécessite une formalisation mathématique du langage, différente de la linguistiques, qui doit avoir des objectifs limités et précis.

CONCLUSION

L'originalité du CLRU tient au fait qu'il a mené une réflexion plurielle sur la formalisation du langage ou des langues en vue de l'automatisation de la traduction, réflexion qui est et restera tout à fait exceptionnelle parmi les acteurs de la TA ou du traitement automatique des langues. Elle se traduit dans la diversité de conceptions du thesaurus, tour à tour langue intermédiaire ou dictionnaire des contextes, contenant des informations exclusivement lexicales ou bien grammaticales et lexicales. Pourtant le thesaurus est ce qui unifie la réflexion du groupe. Et c'est sans doute parce qu'il est au coeur d'un dispositif matériel, et que son élaboration est assujettie à la mise en place d'une nouvelle technique, que cet objet peut être commun à l'ensemble du groupe et stabiliser une partie des réflexions.

Une critique commune les oppose à la conception de la formalisation de Bar-Hillel ou de Chomsky qui proposent un modèle logico-mathématique a priori de formalisation du langage. Pourtant il est intéressant de rnoter que co-existent dans le groupe deux positions contradictoires. La philosophe logicienne Margaret Masterman et la plupart des membres du groupe conçoivent la formalisation à partir d'une appréhension globale du langage. C'est ce que traduit l'idéal de traduction idiomatique défendu par Masterman et qui place les chercheurs devant l'impossibilité de stabiliser la formalisation. Le langage, pris dans sa totalité, ne veut pas se faire mettre en formule. Même sous une forme abstraite, comme celle Nude, il peut être parlé et produire de nouvelles métaphores. Dans ce cadre, un linguiste comme Halliday se voit contraint de fabriquer, sur le modèle des grammaires descriptives, une grammaire de l'extra-linguistique.

Inversement Kay considère que la formalisation linguistique ne peut être que limitée, en fonction d'objectifs précis. De plus, même s'il accepte l'idée d'une formalisation du sens, partagée par l'ensemble du groupe, il n'adhère pas au 'tout sémantique' promu en réaction au 'tout syntaxique' du structuralisme ambiant. Il est d'ailleurs le seul à intègrer des éléments grammaticaux dans le thesaurus. Tous toutefois considèrent que,

les objectifs n'étant pas les mêmes, la formalisation pour la TA n'est pas celle de la linguistique.Chez Halliday, par exemple, les éléments non-lexicaux, les opérateurs, sont des catégories non linguistiques, artificiellement créées pour la TA.

Cette prise en considération du sens explique en partie que le groupe ait eu comme descendance principale, non des travaux en linguistique computationnelle comme la plupart des groupes américains, mais des travaux en intelligence artificielle.

RÉFÉRENCES BIBLIOGRAPHIQUES

- ARCHAIMBAULT Sylvie, LÉON Jacqueline. 1997. "La langue intermédiaire dans la Traduction Automatique en URSS (1954-1960). Filiations et modèles", *Histoire Epistémologie, Langage* vol.17:2. 105-132.
- AUROUX Sylvain. 1998. *La raison, le langage et les normes*, PUF, Paris.
- BASTIN E.W. et NEEDHAMR.M. . 1959. "A new research for analysing language" in Essays on and in *Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- *Essays on and in Machine Translation by the Cambridge Language Research Unit*. 1959. MASTERMAN Margaret , PARKER-RHODES A.F., SPARCK JONES Karen, KAY Martin, MAY E.B., NEEDHAM R.M., BASTIN E.W., WORDLEY C., F.H., ELLIS, MCKINNON WOOD R., 8 essais dédicacés à Yehoshua Bar-Hillel, (non publiés), juin 1959
- BAR-HILLEL Yohoshua . 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.
- BASTIN E.W. . 1956. "Résumé de son intervention au Cambridge Language Research Group, Meeting at King's College, Cambridge, England, August 2-4 1955", *Mechanical Translation*, vol.3:1. 2-7 ;
- BOOTH A.D. (ed.). 1967. *Machine Translation*, North Holland Publishing Company, Amsterdam.
- BRONOWSKI J.. 1956. "The theory and philosophy of language", *Mechanical Translation*, vol.3:1. 12-13.
- CLRU. 1956. "Cambridge Language Research Group, Meeting at King's College, Cambridge, England, August 2-4 1955", *Mechanical Translation*, vol.3:1. 2-7.
- HALLIDAY M.A.K., 1956, "The linguistic basis of a mechanical thesaurus" *Mechanical Translation*, vol.3:3. 81-88.
- HUTCHINS William John. 1986. *Machine Translation, past, present, future,* Ellis Horwood ltd.
- HUTCHINS William John. 1995. 'The whisky was invisible', or persistent myths of MT", *MT News International* 11:17-18.
- KAY Martin. 1959 . "The relevance of linguistics to MT" in Essays on and in *Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- *Language and Machines. Computers in translation and linguistics*. 1966. A report by the Automatic Language Processing Advisory Committee (ALPAC), National Academy of Sciences, National Research Coucil.
- LÉON Jacqueline 1998. "Les premiers outils pour la Traduction Automatique: demande sociale, technologie et linguistique (1948-1960)", *Actes du colloque international FRACTAL, Linguistique et Informatique: théories et outils pour le*

traitement automatique des langues, Besançon, 10-12 décembre 1997, *BULAG* 23. 273-295.
- LOCKE W.N. and BOOTH A.D. (eds.). 1955. *Machine Translation of Languages*, 14 essays, MIT et John Wiley.
- MASTERMAN Margaret .1957. "The Thesaurus in Syntax and Semantics" *Mechanical Translation*, vol 4:1-2. 35-44.
- MASTERMAN Margaret . 1959a. "Fictitious sentences in language" in *Essays on and in Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- MASTERMAN Margaret . 1959b. "What is a thesaurus?" in *Essays on and in Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- MASTERMAN Margaret . 1967. "Mechanical pidgin translation, An estimate of the research value of 'word-for-word' translation into a pidgin language, rather inot the formal form of an output language" in *Machine Translation*, A.D. Booth ed. .195-227.
- MEL'CUK Igor Alexandre 1961. "Some problems of MT abroad, USSR", *Reports at the conference on Information Processing, MT and Automatic Text Reading*, Academy of Science, Institute of Scientific Information, n°6, Moscou .1-44
- PARKER-RHODES A.F. 1956. "An electronic computer program for translating Chinese into English", *Mechanical Translation*, vol.3:1. 14-20.
- RICHENS R.H. 1956a. " Preprogramming for Mechanical Translation", *Mechanical Translation*, vol.3:1. 20-28.
- RICHENS R.H. 1956b. "General program for mechanical translation between any two languages via an algebraic interlingua" résumé *Mechanical Translation*, vol.3, n°2:37
- SPÄRCK JONES Karen .1959a. "The possible use of decision procedures in MT", in *Essays on and in Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- SPÄRCK JONES Karen . 1959b. "Note on the inappropriateness of the division made by Bar-Hillel between FAHQMT and man-machine partnership translation MMPT, in *Essays on and in Machine Translation by the Cambridge Language Research Unit*. rapport non publié
- YNGVE Victor H. .1957. "A Framework for Syntactic Translation" *Mechanical Translation*, vol.4.3 .59-65.