CONCEPTIONS DU 'MOT' ET DÉBUTS DE LA TRADUCTION AUTOMATIQUE

Jacqueline LÉON
UMR CNRS 7597, Université Paris VII

RÉSUMÉ : Dans cet article, nous proposons d'étudier de façon comparative les conceptions du mot mises en œuvre par les linguistes lors de leur confrontation avec les premières expériences de traduction automatique dans les années 1950-1960 et qui les ont amenés à définir tant des unités de segmentation de la chaîne écrite que des unités de traduction. Nous examinerons les traditions américaine et française, qui offrent à la fois des points de similitude et d'opposition qu'on peut saisir à travers les aspects suivants : deux conceptions structuralistes distinctes du mot, un décalage de plus de dix ans concernant les débuts de la TA, une implication concrète très différente des linguistes dans le traitement automatique.

MOTS-CLÉS : Structuralisme européen ; Structuralisme américain ; Théorie de l'information ; Traduction automatique ; Mot ; France ; États-Unis.

ABSTRACT : This article gives a comparative study of the various conceptions of the word which were implemented by linguists when they were first confronted with the earliest machine translation experiments in the 1950s and 60s. They were then led to define segmentation units for the written sequence and translation units. The American and the French traditions will be examined — they offer both similarities and oppositions, which can be grasped through the following issues : two distinct structuralist approaches of the word, a gap of more than 10 years as regards the beginning of NLP, a very different concrete implication by linguists in NLP.

KEY WORDS : Structuralism ; Machine translation ; Word ; France ; United States of America.

INTRODUCTION [1]

ON SAIT qu'en linguistique, le mot est une notion complexe et hétérogène dont les différentes dimensions, graphique, phonétique, syntaxique ou sémantique

---

1. Je remercie Sylvie Archaimbault pour sa lecture attentive.

coïncident rarement et n'ont pas de propriétés constantes. Ce statut particulier du mot a conduit les linguistes à le rejeter comme unité linguistique et à lui substituer d'autres termes correspondant aux différents plans d'analyse et de représentation linguistique : *signe, morphème, lexème, monème, lexie, unité lexicale*, voire *mot* lui-même ; ces termes correspondant chaque fois à des options théoriques distinctes [2].

Or il faut noter qu'un certain nombre de ces termes désignant le mot ont été forgés par des linguistes français dans les années 1950-1960, en partie sous l'impulsion de la mécanisation du langage. En effet, les premiers expérimentateurs en traduction automatique, qui, comme on sait, constitue la première tentative en traitement automatique des langues, ont été confrontés à la définition d'une unité de segmentation de la chaîne écrite qui soit également une unité de traduction. Ce qui est frappant, dans le cas de la France, c'est que l'intérêt d'un certain nombre de linguistes a été éveillé par cette recherche d'unité et par sa confrontation avec le mot graphique aisément reconnaissable par la machine. Il en est résulté un renouveau des études lexicales et sémantiques.

C'est ce trajet, des unités de segmentation aux unités lexicales, que nous nous proposons d'étudier dans cet article, au travers d'une analyse comparative. Afin de mettre en perspective ces aspects de la linguistique contemporaine et du traitement automatique des langues, dont les paradigmes sont encore très présents, la comparaison de deux traditions est un moyen d'introduire une certaine réflexivité, nécessaire à toute approche historique. Contrairement à l'époque actuelle, où l'internationalisation des recherches est la règle dans le domaine fortement industrialisé du traitement automatique des langues, les débuts de la mécanisation du langage dans les années 1950-1960 restent nettement marqués par les différentes traditions linguistiques et culturelles.

Les travaux américains nous semblent les plus aptes à être mis en parallèle à la tradition française. Plusieurs points opposent les deux traditions dans les années 1950-1960 tout en permettant la comparaison : deux conceptions structuralistes distinctes du mot, un décalage de plus de dix ans concernant les débuts de la TA, l'intérêt et l'implication concrète des linguistes, une

---

2. Toutefois le débat n'est pas clos et, si l'on se réfère simplement aux travaux de la dernière décennie, il est de nouveau vif parmi les linguistes français. En témoigne la parution récente d'ouvrages ou de numéros spéciaux de revue qui lui sont consacrés : en 1990 le n° 12 de *Modèles linguistiques* intitulé « La notion de mot. Le cas des langues indo-européennes », le n° 10 de *Lalies* paru en 1992 « Le mot », l'ouvrage édité par Fradin et Marandin en 1997 *Mot et grammaires*, enfin le n° 7 de *Langues et langage* « Le mot : analyse du discours et sciences sociales » paru en 1998, pour ne citer que ceux-là. Il serait intéressant d'examiner ce qui motive, à l'heure actuelle, ce regain d'intérêt.

inscription singulière des linguistes intéressés par la TA dans la tradition linguistique de chacun des pays.

Pour chaque tradition nous tenterons de dégager les conceptions du mot sous-jacentes aux premiers travaux de traduction automatique ou bien qui ont émergé de la confrontation des linguistes à la mécanisation du langage.

1. DU MOT-CODE AU MOT-PROCÉDURE. L'HÉRITAGE DE LA THÉORIE DE L'INFORMATION, DE L'ANTHROPOLOGIE LINGUISTIQUE ET DU STRUCTURALISME NÉO-BLOOMFIELDIEN

Nous partirons de deux idées reçues concernant le mot et la traduction automatique aux USA dans les années 50.

Aux États-Unis, les linguistes ne veulent entendre parler ni d'informatique ni de TA. On citera en tout premier lieu, Chomsky, qui, bien qu'embauché au MIT sur un projet de TA par Yngve en 1955, refuse de s'y intéresser (*cf.* Cori, Marandin dans ce numéro même). Les rares linguistes [3] qui s'y impliquent ne s'intéressent qu'à l'analyse syntaxique automatique de la langue source sans étudier ni le passage de la langue source à la langue cible ni la synthèse de la langue cible. Préoccupés pour des raisons politiques et militaires par la seule traduction du russe vers l'anglais, la comparaison entre les langues ne les intéressent pas. La TA, très rapidement réduite à la mise à l'épreuve des langages formels par l'informatique, fait partie très tôt de la linguistique computationnelle.

Le deuxième constat concerne le statut du mot. Les linguistes américains qui étudient les langages formels, comme ceux qui élaborent des modèles d'analyse syntaxique pour la TA adoptent, sans la discuter et indifféremment, la notion de mot ou de morphème, telle qu'elle a été définie par Bloomfield en 1933 et qui reste traditionnellement admise dans la linguistique structuraliste américaine. Dans cette perspective, racine, mot et morphème (signe minimum), formes libres minimales, isolables prosodiquement, sont toutes des unités syntaxiques. Il est donc inutile de distinguer morphologie et syntaxe. Ce sera également la position de la grammaire générative à ses débuts qui ne reconnaîtra un niveau morphologique indépendant de la syntaxe qu'au début des années 70 en acceptant un module morphologique de formation des mots.

Dans les premiers travaux de Chomsky, que ce soit *Structures syntaxiques* ou *Trois modèles de description du langage*, morphèmes et mots sont indifféremment des symboles terminaux. Dans les *Trois modèles* (version française p. 55) on trouve une note qui définit les morphèmes comme les plus petits éléments à fonction grammaticale de la langue, par exemple *boy*, *run*, *ing*

---

3. Le cas de Bar-Hillel, philosophe logicien, et auteur de formalismes comme sa grammaire catégorielle, doit être traité à part. Premier chercheur en TA nommé au MIT en 1951, il fut aussi un des évaluateurs les plus critiques de la TA.

dans *running*, *s* dans *books*. Mais on ne trouve pas de définition du mot et sa nature d'unité linguistique n'est pas discutée. Le passage de la séquence terminale de morphèmes à la séquence de mots n'est pas explicité :

> Pour produire une phrase à partir de la grammaire (structure syntagmatique, structure transformationnelle, morphophonologie), nous construisons une dérivation élargie commençant par Phrase. En passant par les règles F, nous construisons une séquence terminale qui sera une suite de morphèmes, pas nécessairement dans l'ordre correct. Nous passons alors par la suite de transformations $T_i...T_j$ appliquant celles qui sont obligatoires et peut-être certaines de celles qui sont facultatives. Ces transformations peuvent réordonner les séquences, ajouter ou effacer des morphèmes. Elles ont pour résultat la production d'une séquence de mots. Nous passons alors par les règles morphophonologiques, qui convertissent cette séquence de mots en une séquence de phonèmes. (Chomsky 1957, traduction française, p. 52)

Par ailleurs il semblerait, qu'aux États-Unis, les problèmes posés par la segmentation de la chaîne écrite, nécessitée par la TA, n'ait pas éveillé l'intérêt des linguistes.

Ces deux constats, désintérêt des linguistes pour la TA et absence de réflexion linguistique sur la définition d'une unité de traduction expliqueraient, en partie [4], l'arrêt des recherches en TA aux États-Unis, après la publication de l'ALPAC en 1966, au profit de la seule linguistique computationnelle, assimilée à cette époque au développement des langages formels et des analyseurs syntaxiques automatiques [5].

Il n'est pas étonnant que, dans ces conditions, les dictionnaires électroniques, immédiatement utilisables par les analyseurs syntaxiques, aient été considérés comme les seuls résultats positifs de la période. Toutefois, à côté de l'historiographie officielle, cautionnée par les recommandations de l'ALPAC, il est intéressant de voir que, dans les années 50, deux courants importants aux États-Unis étaient impliqués dans la conception de méthodes de traduction, mais que ces méthodes, qui mettaient en œuvre une conception spécifique et originale du mot, eurent tôt fait d'être abandonnées — certaines cependant ressurgirent plusieurs décennies plus tard. Il s'agit de la théorie de l'information conçue par Shannon et Weaver (1948) dans le cadre de la première cybernétique, et d'autre part des méthodes de traduction mises au point par les anthropologues linguistes et les néo-bloomfieldiens qui restèrent en grande partie méconnues.

---

4. Il ne faut pas oublier que l'évaluation négative de la TA tenait également à la faiblesse des résultats obtenus, disproportionnés par rapport aux moyens investis et à l'ambition affichée de certains chercheurs.
5. Sur l'histoire générale de la TA, voir Hutchins (1986) et son article dans le présent numéro.

1.1. Dictionnaires électroniques et « mot-procédure »

Les dictionnaires électroniques ont eu une importance non négligeable dans l’histoire du traitement automatique des langues. Premiers outils opérationnels dans le domaine, ils ont contribué à promouvoir une conception du mot totalement originale, que l’on pourrait nommer « mot-procédure ». Cette organisation du dictionnaire et les stratégies d’analyse qu’elle implique sont au fondement de nombre d’algorithmes de traitements automatiques des langues.

Le premier dictionnaire électronique russe-anglais, objet d’une thèse en mathématiques appliquées et non en linguistique (Oettinger 1955), se distingue des dictionnaires bilingues traditionnels par un certain nombre de traits :
- contrairement aux conventions de la lexicographie, il est élaboré à partir des formes abrégées et non à partir des lemmes ;
- les unités stockées dans le dictionnaire n’obéissant qu’aux contraintes graphiques, seules reconnues par la machine, les expérimentateurs ont été amenés à fabriquer des dictionnaires de racines (appelées bases de mots) et de terminaisons qui n’obéissent pas aux critères fonctionnels discutés par les grammairiens [6]. Ceux-ci reprochent à la segmentation adoptée de n’être valable que pour les langues à suffixation et non pour les langues dont l’information morphologique est portée par des procédés tels que la reduplication, l’apophonie de la voyelle, les mutations consonantiques ou des morphèmes discontinus.

Enfin, ces dictionnaires comportent une dernière propriété de portée considérable. L’information grammaticale, extraite de l’analyse morphologique automatique visant à identifier les bases, va se trouver stockée dans le dictionnaire. Incluse dans des procédures intégrées au dictionnaire, elle devient dynamique. On voit ainsi apparaître ce qu’on pourrait appeler le mot-procédure, qui constitue un apport important du TAL notamment à la théorie des dictionnaires [7].

---

6. Certains expérimentateurs en TA revendiquent d’ailleurs une position spécifique de ‘linguistes pour la TA’. C’est notamment la position de Reifler (1955), germaniste et sinologue, qui considère qu’il n’est pas toujours nécessaire de tenir compte des résultats des linguistes et qui propose parfois un traitement « arbitraire » du matériel linguistique mieux adapté aux objectifs de la TA.
7. Plus tard, dans son *Cours de Morphologie Générale T1*, (1993, p. 123), Mel’©uk proposera d’intégrer cette information grammaticale, qu’il appelle syntactique, dans la définition même du signe. Il redéfinit un signe linguistique comme un triplet X = (Y ; Z ; W) dans le quel Y est un signifié de Z, Z est un signifiant de Y et W est un syntactique de la paire (Y ; Z). Le syntactique, marginal dans un langage formel, est tout à fait systématique dans une langue naturelle.

1.2. Mot-code et théorie de l'information

La théorie de l'information tient les mots graphiques pour des unités de code, les codes étant les langues naturelles écrites. Formé des lettres qui sont des unités de traitement du signal, le mot-code n'est pas une unité linguistique. Les textes écrits, en tant que systèmes de signaux, ont deux propriétés spécifiques par rapport aux autres systèmes. Premièrement ils restent intelligibles malgré le bruit, que constituent par exemple les coquilles dans un article ou les erreurs de transmission d'un télégramme, appelé par Shannon *channel noise*. Cela tient au fait que les langues sont redondantes (selon Shannon, l'anglais écrit est redondant à 50 %, c'est-à-dire déterminé par la structure de la langue). La deuxième propriété des textes, c'est que le bruit peut être aussi sémantique, et pas seulement aléatoire et lié à la transmission du signal ; ce qui permet de comprendre que, malgré l'écart entre le code de l'émetteur et celui du receveur, il y a tout de même compréhension. Les bruits sémantiques qui intéressent particulièrement Weaver sont ceux que l'on rencontre dans la traduction d'une langue dans une autre et qu'on peut quantifier grâce à la théorie de l'information. C'est le cas des ambiguïtés.

Si l'on considère une langue naturelle écrite comme un code, il faut la traiter comme une source d'information discrète, représentable par un processus stochastique, c'est-à-dire un système produisant une séquence de symboles (des lettres ou des mots) selon certaines probabilités. Dans le cas particulier des langues, il s'agit d'un processus stochastique dont les probabilités dépendent d'événements précédents (lettres ou mots qui précèdent) ; on a alors affaire à un processus (ou chaîne) de Markov. On définit ainsi des digrammes ou des trigrammes pour lesquels le choix d'une lettre dépend de la lettre ou des deux lettres précédentes. Weaver suggère qu'un contexte de *2n* mots adjacents (*n* variant selon le type de texte traduit) devrait suffire à lever les ambiguïtés.

À l'exception de quelques expérimentations (*cf.* Kaplan 1950), la méthode préconisée par Weaver n'eut pas beaucoup de succès. Celle-ci était, à l'époque, impossible à mettre en œuvre, ce que Weaver a lui-même reconnu (*Memorandum* p. 21, 1955). Il n'existait pas de corpus suffisamment représentatif et les ordinateurs n'étaient pas assez puissants pour stocker toutes les chaînes de di-grammes et de tri-grammes nécessaires à l'établissement de statistiques.

Les méthodes empiriques, prônées par certains groupes de TA (notamment celui de l'Université de Georgetown) et qui nécessitaient de vastes corpus de textes, rencontraient les mêmes difficultés. Selon ces méthodes, il s'agissait de construire les règles de grammaire et les dictionnaires « manuellement » à

partir de corpus de la langue-source et cumulativement par l'accroissement du corpus [8].

Il faut noter que l'abandon de ces méthodes est dû également à de vives critiques relayées par de fortes sanctions institutionnelles. Bar-Hillel, dans son rapport de 1960 sur la TA, présente des arguments sévères contre l'approche empirique. Chomsky, à la fin du chapitre 2 de *Structures syntaxiques*, s'oppose radicalement aux méthodes statistiques et à l'approche empirique dans son argumentation destinée à distinguer suites grammaticales et agrammaticales.

Or on sait que l'utilisation de corpus n'est pas complètement étrangère aux pratiques des structuralistes neo-bloomfieldiens qui en font une pierre de touche de leur méthode de description. Si l'on suit Murray (1994), ce rejet du corpus par Chomsky participerait de sa volonté de radicaliser la rupture avec ses prédécesseurs et maîtres pour s'ériger en novateur absolu.

1.3. Le mot « découvert ». Les structuralistes américains et la théorie de l'information.

Loin d'être ignorée par les linguistes, la théorie de l'information a été discutée, voire utilisée par les structuralistes des années 50. Hockett (1953) consacre à l'ouvrage *Mathematical Theory of Communication*, un très long compte-rendu de vingt-quatre pages dans *Language* où il déclare avoir suivi, au MIT en 1951, un cours intensif dans le domaine. Il applique la méthode à un exemple en phonémique et en morphologie où l'encodage d'un morphème s'effectue en fonction de son contexte : « *wife* is encoded into */wayv/* if the next morpheme is the noun-plural *-s* and *-s* is encoded into */z/* rather than */s/* when the preceding morphem is *wife* » (p. 87). Il signale des expériences qui montrent que la probabilité d'apparition d'un phonème augmente au milieu d'un mot ou d'un morphème, et diminue aux frontières. Elle est la plus faible à la frontière entre constituants immédiats.

En 1955, Harris mettra cette méthode en œuvre, bien que ne citant pas Hockett, pour segmenter une chaîne transcrite en phonèmes. Par exemple,

---

8. On notera que ce n'est qu'au début des années 1990, suite au succès du traitement du signal dans la reconnaissance de la parole, que les méthodes probabilistes connaîtront un regain d'intérêt, de même que les méthodes empiristes sur corpus et les mémoires de traduction. Celles-ci bénéficièrent, outre de la puissance des ordinateurs, de la disponibilité de grands corpus bilingues, comme les *Canadian Hansards*. Dans un grand corpus de textes bilingues, on calcule la probabilité qu'un mot dans une phrase dans une langue donnée corresponde à aucun, un ou deux mots dans la traduction. On établit un glossaire des mots équivalents dans les deux langues qui est constitué de listes de possibilités de traduction de chaque mot avec une probabilité correspondante. Par exemple *the* (en anglais) a une probabilité de 0.610 d'être traduit (en français) par *le* et de 0.178 d'être traduit par *la*. Toutefois on peut déplorer que le statut des statistiques, outil ou modèle du texte, des langues ou du langage ne soit pas explicité dans ces méthodes.

l'énoncé transcrit /hiyzkwiker/ *He's quicker* sera segmenté selon les points : /hiy.z.kwik.er/. Toutefois, précise-t-il, seule l'application de méthodes morphologiques traditionnelles sont capables de déterminer si les segments obtenus sont des morphèmes ou des mots. De ce fait la méthode reste limitée du point de vue linguistique.

Dans une telle méthode, le mot code a perdu la réalité empirique propre au mot : il n'a plus de correspondance avec un segment perceptible intuitivement par les locuteurs. Comme il est peu probable qu'Harris ait partagé le modèle de langue-code et du texte comme résultat d'un processus stochastique préconisé par Weaver, on ne sait pas très bien à quel modèle des langues ou du langage référer les propriétés statistiques invoquées dans la découverte des morphèmes.

*1.4. Une définition opérationnelle du mot pour la traduction : l'*informant word *des anthropologues linguistes*

Dans ce paragraphe, nous voudrions revenir sur le préjugé, évoqué plus haut, selon lequel les linguistes américains ne se sont pas intéressés à la traduction automatique. Ce qui était vrai pour les chomskiens ne l'était pas pour leurs aînés.

Des structuralistes, comme Harris et Hockett, et des anthropologues linguistes comme Voegelin ou Garvin se sont associés pour publier en 1954 un numéro spécial sur la traduction dans l'*International Journal of American Linguistics*. Y sont exposées des méthodes de traduction des langues amérindiennes dont il est explicitement dit qu'elles pourraient être programmées sur ordinateur. Or, bien que ce numéro d'*IJAL* ait été signalé dès 1954 par la revue *Journal of Machine Translation* créée la même année, aucun de ces articles ne sera cité par les expérimentateurs en TA. Par ailleurs, il est tout à fait curieux qu'aucun des auteurs de ce numéro, à l'exception de Paul Garvin, ne poursuivra de travaux en TA.

Il est intéressant de voir que les auteurs du numéro d'*IJAL* se réfèrent à la méthode utilisée par les anthropologues linguistes pour traduire les langues amérindiennes et décident d'en systématiser le processus en explicitant linguistiquement les étapes intermédiaires entre la traduction interlinéaire, dite « mot à mot », et la traduction libre. Voegelin définit une procédure en huit étapes qu'il nomme 'Multiple Stage Translation'. Le premier problème à résoudre, dit-il, est l'identification des unités de la première étape dite de traduction mot à mot. Comme il n'est pas sûr que les langues amérindiennes aient toutes des unités mots, Voegelin propose une procédure utilisant une définition opérationnelle du mot, qu'il appelle *informant's word*. Le texte à traduire enregistré sur magnétophone est découpé en contours, c'est-à-dire en étendues de parole délimitées par des éléments prosodiques comme les pauses et les intonations montantes ou descendantes. L'informateur a pour consigne de lire et de traduire « mot à mot » cette étendue de parole. L'*informant's word*

est donc l'étendue de parole la plus petite qu'un locuteur natif peut prononcer de façon isolée, et traduire, à partir d'une étendue de parole plus longue délimitée prosodiquement. Voegelin fait remarquer qu'il y a un très fort consensus entre informateurs sur la délimitation de l'unité mot ainsi définie.

Souvent, pourtant, l'*informant's word* ne correspond pas au mot ou au morphème du linguiste. Ainsi en *shawnee*, comme dans la plupart des langues algonquines et agglutinantes, le verbe commence par une particule et finit par un suffixe. Alors que, pour le traducteur linguiste, tous les morphèmes compris entre la particule et le suffixe font partie du même mot, pour l'informateur, la particule est prononcée et traduite comme un mot séparé. En *hidatsa*, comme dans les langues sioux en général, c'est le contraire et l'informateur fusionnera en un seul mot plusieurs mots considérés comme différents par le linguiste.

La seconde étape consiste à identifier les morphèmes ; la traduction morphème par morphème restant, pour le linguiste, un outil précieux d'analyse de la langue considérée. Les autres étapes, par la mise en oeuvre d'opérations d'addition, soustraction, substitution et réarrangement de mots, aboutissent à un texte dit en 'traduction libre'.

Dans cette procédure de traduction en plusieurs étapes, on se trouve en présence de deux définitions du mot. Une définition opérationnelle permettant de définir l'unité de segmentation de la langue source pour la traduction. Comme le mot, cette unité de traduction est un segment signifiant et isolable par les locuteurs dans la chaîne parlée. Pourtant elle ne correspond pas nécessairement au mot empirique en anglais. L'unité linguistique reste le morphème et garantit l'homogénéité de la description linguistique pour les deux langues.

Certaines idées de Voegelin semblent avoir été reprises par Paul Garvin — qui était son doctorant — auteur d'un article sur la traduction du *kutenai* dans le numéro spécial d'*IJAL*, et engagé dans des travaux de traduction automatique dans le groupe de Georgetown. En 1956, Garvin préconise de distinguer les *sensing units*, mots graphiques, unités de segmentation du texte à traduire et les *translation units* dont il distingue deux types : les unités de sélection (les unités lexicales), dont le sens ne peut être prédit de la somme de ses parties, et les unités d'arrangement (les morphèmes) dont la distribution ne peut être établie à partir de ses composants. Il justifie cette distinction en argumentant que, contrairement à l'informateur humain qui a tout de suite affaire au sens et à qui le mot peut servir à la fois d'unité de segmentation de la chaîne parlée et d'unité de traduction, la machine ne reconnaît que des formes, suites discontinues de lettres ou de séparateurs mais pas les unités de sens. Comme la traduction nécessite qu'on ait affaire au sens et parfois au continu, il faut distinguer deux types d'unités, distinctes du mot.

Il est intéressant de voir que Garvin transfère à la machine la tâche de traduction habituellement dévolue à l'anthropologue linguiste. Pour cela il

équipe la machine de procédures et d'unités spécifiques qui pallient le fait qu'elle ne comprend pas ce qu'elle traduit.

1.5. Mot et classes de mots

Les classes de mots sont aussi convoquées pour établir des méthodes de traduction. Hockett (1954) propose une traduction du chinois en anglais au moyen des constituants immédiats. Les étapes de la traduction sont figurées dans son tableau p. 314. Celui-ci est un tableau (*chart*) de la structure en constituants immédiats de la phrase chinoise (1ère ligne) :

<table>
<tr><td>jè
proxi
mal</td><td>-i
thing
state</td><td>gen
with,
and</td><td>ni
thou</td><td>yi
one</td><td>gùng
to-
ge-
ther</td><td>gai
must
owe</td><td>wo
I,me</td><td>de
nom
ina
li-
zer</td><td>hai
still-
yet</td><td>chà
differ
fall
short</td><td>de
nom
inali
zer</td><td>dwo
much
more
many</td><td>ne
suspen-
sive</td></tr>
<tr><td colspan="4"></td><td colspan="2">altogether</td><td colspan="5"></td><td colspan="4"></td></tr>
<tr><td colspan="4"></td><td colspan="3">altogether owe(s)
altogether must</td><td colspan="3"></td><td colspan="4"></td></tr>
<tr><td colspan="4"></td><td colspan="4">owe(s) me altogether</td><td colspan="4"></td><td colspan="3"></td></tr>
<tr><td colspan="4"></td><td colspan="5">you owe me altogether</td><td colspan="4"></td><td colspan="3"></td></tr>
<tr><td colspan="4"></td><td colspan="6">what you owe me altogether</td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="2">this</td><td colspan="8">and what you owe me altogether</td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="9">this and what you owe me altogether</td><td colspan="2">yet differ(s)</td><td colspan="2"></td><td colspan="3"></td></tr>
<tr><td colspan="11">this is still less than what you owe me altogether</td><td colspan="2"></td><td colspan="3"></td></tr>
<tr><td colspan="12">the extent to which this is still less than what you owe me altogether</td><td colspan="2"></td></tr>
<tr><td colspan="13">this is still a lot less than what you owe me altogether</td></tr>
<tr><td colspan="13">This is STILL a lot less than what you owe me altogether</td></tr>
</table>

Hockett considère comme essentiel de rendre disponible pour le linguiste les résultats intermédiaires entre la traduction morphème par morphème (2e ligne) et la traduction libre (dernière ligne). Pour tous les segments du texte, il donne le dénominateur commun de sens à tout contexte possible. Il ne s'agit pas d'une option de traduction mais d'une option d'interprétation. Ainsi pour *wo*, il donne les sens *I* et *me*. En traitant *wo* comme un constituant immédiat, la construction dans laquelle il apparaît permet de décider que c'est la glose *me* qui va être retenue.

Son utilisation des constituants immédiats comme contexte morpho-syntaxique pour résoudre les ambiguïtés préfigure la méthode de *mutual-pinpointing* développée par le groupe de TA de Washington dirigé par Erwin Reifler (Reifler 1955, Micklesen, 1956). Reifler donne l'exemple de l'article *den* en allemand qui peut soit être accusatif masculin singulier, soit datif pluriel. S'il est suivi de *Männern*, qui n'est que datif pluriel, l'ambiguïté est levée (*den* et *Männern* jouent ici le rôle de *mutual pinpointers*). La conception du mot ici revendiquée étant celle de « classes de formes » développée par Fries.

Il est également intéressant de mentionner l'article de Harris « Transfer grammar » bien que la question de l'unité de segmentation, mot ou classe de mots, n'y soit pas posée. Harris cherche à mesurer la différence entre les

langues et il propose une méthode, la grammaire de transfert, susceptible de mesurer les différences de structures grammaticales et d'établir le minimum de différence ou le maximum de ressemblance entre deux systèmes linguistiques. La différence entre deux langues est définie sous forme d'instructions grammaticales servant à générer les énoncés d'une langue à partir des énoncés d'une autre langue. Cette méthode peut être utilisée à des fins d'enseignement des langues étrangères en faisant l'hypothèse qu'on peut acquérir une langue en n'apprenant que les différences entre la nouvelle et l'ancienne. Transformées en instructions pour la machine, les instructions grammaticales peuvent également servir à mettre au point une procédure de traduction automatique.

Enfin, le cas de Sydney Lamb constitue une exception parmi les acteurs de la TA dans la mesure où il est un des rares à se préoccuper de l'unité de segmentation. Linguiste, arrivé plus tardivement dans le domaine de la TA et formé par les anthropologues linguistes, Murray B. Emeneau et Mary Haas, élèves de Sapir, il revendique leurs méthodes de recueil des données et de travail sur corpus (*cf.* Lamb 2000). En TA, il part donc du texte et se pose la question de l'unité de segmentation. Mais c'est surtout la recherche d'une unité syntaxique qui l'intéresse. Proche des néo-bloomfieldiens et notamment de Hockett, il veut mettre en œuvre une syntaxe par des méthodes distributionnelles et ne semble pas chercher à donner un statut linguistique aux unités de segmentation (Lamb 1962).

Pour conclure, il est intéressant de voir que, même pour les structuralistes américains, pour qui seul le morphème a une réalité linguistique, le mot devient une unité dès lors qu'ils sont confrontés à une tâche pratique, celle de segmenter et traduire un « texte » parlé dans une langue autre que des langues européennes. Cette question de la segmentation a été reprise pour la TA, de façon distincte par Garvin et Lamb, tous deux formés aux méthodes des anthropologues linguistes. Toutefois, leurs travaux ont abouti au développement d'algorithmes d'analyse syntaxique, plus qu'à une réflexion sur les unités de texte ou de traduction.

2. ENTRE MOT-SIGNE ET MOT-VOCABULAIRE : LA *TA*, LES LINGUISTES FRANÇAIS ET LE RENOUVEAU DES ÉTUDES LEXICALES

Concernant la situation en France au début des années 60, on peut faire trois constatations :

1) ce n'est que tardivement que les Français ont commencé à s'engager dans des projets de TA ;

2) alors que peu de linguistes étaient impliqués directement dans des projets de TA, beaucoup avaient des fonctions d'évaluation ou de consultation et suivaient le domaine de très près ;

3) contrairement aux Américains, la question de la définition du mot était au cœur des discussions sur la mécanisation du traitement du langage. Celles-ci rassemblaient non seulement des linguistes, mais des mathématiciens, des ingénieurs, des traducteurs et des documentalistes. En témoignent plusieurs colloques dans la période 60-65 dont un colloque sur le mot organisé fin 62 par l'ATALA (Association pour la Traduction Automatique et la Linguistique Appliquée).

On peut suggérer deux raisons expliquant l'intérêt des linguistes français pour la question du mot : l'ancrage, dans le structuralisme européen, de la linguistique française 'la plus avancée', et l'attachement des linguistes plus traditionnels à l'étude du vocabulaire, de la dialectologie et de la stylistique. On peut penser que la TA a permis de dépoussiérer un certain nombre de questions et que la confrontation de certains linguistes avec la TA, autour de la question du mot, a ouvert la voie à un renouveau de la lexicologie, notamment à une certaine forme de sémantique lexicale.

2.1. Les linguistes et les débuts de la TA en France

Rappelons que les centres de TA en France sont créés une dizaine d'années environ après les Américains et les Britanniques ; le retard français en informatique [9] expliquant en partie cet intérêt tardif. Outre l'ATALA, créée en avril 1959 autour d'Emile Delavenay qui a joué un rôle moteur dans le développement de la TA en France, il existe deux centres de TA créés par le CNRS : le CETA, crée en décembre 1959 comprenant deux pôles, l'un à Paris, l'autre à Grenoble ; et le Centre de l'Université de Nancy créé en mai 1960.

Très peu de linguistes font partie des centres de recherches de TA. Maurice Gross, qui a reçu une formation d'ingénieur et Yves Gentilhomme, de formation mathématique, font partie du centre parisien. À sa dissolution fin 1962, ils se consacrèrent à l'étude des langages formels. Parmi les linguistes de formation littéraire, seuls Bernard Pottier et Guy Bourquin, fondateurs du Centre de Nancy, s'investirent réellement dans des projets concrets de TA.

---

9. Sur les raisons de l'intérêt tardif des Français pour la TA, *cf.* Léon 1998.

En revanche les linguistes qu'on pourrait qualifier d''institutionnellement établis', s'ils ne sont pas directement acteurs du domaine, s'y intéressent de très près. Certains sont membres fondateurs de l'ATALA comme Marcel Cohen, David Cohen ou Antoine Culioli. D'autres font partie des comités d'évaluation des centres de Paris et Grenoble. C'est le cas, pour ne citer que les plus prestigieux, d'Émile Benveniste, Georges Gougenheim, Michel Lejeune, André Martinet, Bernard Quemada, des slavistes comme Jean Train et Marc Vey, auxquels se joindront plus tard Bernard Pottier et Jean Fourquet.

Il est à noter que, même dans le *BSL*, organe de la Société de Linguistique de Paris qui dédaignait cette linguistique pour ingénieurs comme la qualifiait Martinet, rend compte sur les recherches en TA grâce à Georges Mounin et aux slavistes, notamment René Lhermitte, concernant la TA en URSS.

2.2. Les débuts de la mécanisation du langage en France : les études du vocabulaire

Mais ce qui caractérise la France, contrairement aux États-Unis et à la Grande-Bretagne, c'est que le début de la mécanisation du langage n'est pas passée par la traduction automatique mais par l'automatisation du vocabulaire. Comme le signalent Chevalier et Encrevé (1984, p. 72), l'essor de la lexicologie a commencé avec la figure de Mario Roques et son *Inventaire Général de la Langue Française* créé en 1936. Il a été jalonné par le colloque de 1957 à Strasbourg « Lexicologie et lexicographie françaises et romanes » qui a abouti à la création du TLF en 1960. Lors de ce colloque, a été évoqué explicitement l'apport prometteur des machines mécanographiques et électroniques dans l'accélération des dépouillements et des classements du lexique.

Cet essor de la lexicologie par sa mécanisation est également marqué 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, suivie par un colloque international sur la mécanisation des recherches lexicologiques qui a eu lieu en 1961, toujours à Besançon, et d'un colloque à Strasbourg en 1964 intitulé « Statistiques et analyse linguistique ». À ce dernier colloque participent de nombreux linguistes, G. Bourquin, E. Coseriu, J. Dubois, F. François, G. Gougenheim, A. J. Greimas, P. Guiraud, R. Martin, H. Mitterand, G. Moignet, R. Moreau, Ch. Muller, B. Pottier, B. Quemada, G. Straka. Les communications portent sur l'application de méthodes statistiques à la stylistique, à la philologie, à la dialectologie et à l'enseignement des langues [10].

---

10. *Le Français élémentaire* publié en 1954 sous la direction de Georges Gougenheim, était fondé sur des techniques de dénombrement du vocabulaire.

Ainsi, la mécanisation de la lexicologie s'inscrit en droite ligne des préoccupations de nombre de linguistes de l'époque, à savoir l'étude du vocabulaire français et la stylistique.

Plus rares sont les travaux consacrés à l'analyse structurale de la langue et du lexique. Jean Dubois, utilise des méthodes statistiques et notamment un dictionnaire inverse pour évaluer le rôle de la suffixation dans la formation du lexique en français. D'autres (Robert Martin et Charles Müller) étudient les variantes morphologiques (pronoms relatifs en anglais, auxiliaires *can* et *may*, plus-que-parfait et passé antérieur en français).

Les méthodes statistiques ne font pourtant pas l'unanimité et le débat est vif. On citera notamment un article de A. J. Greimas dans *Le Français Moderne* (1962-1963) critiquant le livre de Pierre Guiraud « Problèmes et méthodes de la statistique linguistique » et dans lequel il lui reproche de considérer les mots graphiques comme les seules unités constitutives du style, au détriment de la structure linguistique. Certains linguistes déclarent ne pas pouvoir faire confiance à leurs collègues mathématiciens puisqu'ils ne sont même pas d'accord entre eux : Mandelbrot critique certains calculs de Guiraud et Herdan dénie à la loi de Zipf une quelconque valeur scientifique.

Enfin, malgré les efforts louables des mathématiciens statisticiens René Moreau et Daniel Hérault, qui tentèrent de promouvoir les langages formels et la théorie de l'information, en organisant dès mars 1960, au sein du *Séminaire de Linguistique Quantitative* de Jean Favard, un enseignement de mathématiques pour linguistes auxquels participèrent notamment Maurice Gross et André Lentin, la plupart des linguistes français qui s'intéresseront à la linguistique quantitative — terme, qui à l'époque recouvrait aussi bien les statistiques que les langages formels ou les modèles probabilistes de la performance proposés par la théorie de l'information — se limiteront aux études statistiques de dénombrement appliquées à la philologie et la stylistique sans toujours éviter le risque de confusion entre statistiques comme outil ou statistiques comme modèle du langage.

L'étude statistique du vocabulaire dans les textes met en œuvre une conception du texte comme ensemble de formes à dénombrer. C'est la forme graphique qui est l'unité de base, unité constitutive du style d'un auteur, unité de sens d'un texte. Même si des procédures de lemmatisation viendront affiner cette approche, l'univocité entre l'apparition d'une forme et son interprétation reste au cœur de l'analyse sémantique d'un texte. Ce que contestera plus tard l'analyse de discours (*cf.* Pêcheux 1969).

2.3. Le mot et le signe

Le mot n'a cessé de défier les linguistes structuralistes européens qui ont tenté régulièrement de s'en débarrasser. Il était incontournable pour Saussure (CLG154) : « Le mot, malgré la difficulté qu'on a le définir, est une unité qui s'impose à l'esprit, quelque chose de central dans le mécanisme de la langue. »

Dans son célèbre article de 1949, Togeby analyse les difficultés qu'il y a à définir le mot comme unité linguistique et tente de contourner le problème et utilisant l'opposition plan de l'expression / plan du contenu développée dans le cadre de la glossématique de Hjelmslev.

Ce caractère incontournable du mot pour les structuralistes européens a été noté par Hockett (1951) dans un compte rendu des conférences données par Martinet en 1946 à Londres. Il en rend compte en comparant les méthodes d'analyse phonologique mises en œuvre par l'École de Prague et par les structuralistes américains. Jusqu'en 1940, dit-il, la phonologie ne se donnait comme points de référence que les mots et les frontières de mots. Ainsi dans l'approche de Troubetskoy et de Martinet, on reconnaît d'abord les mots et ensuite seulement on procède à l'analyse phonologique. Dans l'approche américaine récente, en revanche, on fait d'abord l'analyse phonologique de l'énoncé dans son entier et ensuite on reconnaît les mots.

La TA a ravivé cette question du mot en lui donnant un nouvel enjeu, celui de définir des unités pour la machine qui soient aussi des unités linguistiques et des unités de traduction. Il s'agissait de faire coïncider forme graphique, unité syntaxique et unité sémantique.

Témoignent de ce renouveau d'intérêt, le colloque sur le mot organisé par l'ATALA en 1962 [11], la prise en compte du mot dans les définitions des langages formels et des unités pour la TA en mathématiques appliquées, et surtout le renouveau de la lexicologie au travers de la sémantique structurale [12].

Dans leur ouvrage de 1967 sur les grammaires formelles, Gross et Lentin prennent le soin de définir les unités de base et de distinguer langages formels et langages de programmation d'une part, application aux langues naturelles d'autre part. La citation ci-dessous, extraite de leur ouvrage (p. 13), est intéressante en ce qu'elle soulève la distinction entre mot, unité linguistique, et mot, unité du langage formel. Elle a le mérite de tenter d'éclairer une difficulté des langages formels (*cf.* Auroux 1998) à savoir que l'utilisation du mot comme unité dans le vocabulaire terminal suppose une continuité entre langue naturelle et langue formelle qui n'est pas explicitée dans la théorie. Ce point n'est d'ailleurs pas abordé dans la grammaire générative par exemple.

> Quand nous aurons en vue l'étude des langages de programmation, nous dirons de l'ensemble de base qu'il est un alphabet abstrait formé de lettres

---

11. Parmi les résumés de communications publiés par le numéro de *La Traduction Automatique* de septembre 1963, figurent ceux de Maurice Coyaud, Bernard Pottier, Jacques Perriault, Claude Dubois, Georges Gougenheim et Jean Fourquet.

12. Sémantique structurale et études du vocabulaire étaient étroitement liées. Ainsi au cours du 8e congrès des linguistes à Oslo 1957, S. Ullmann a assigné comme but à la sémantique structurale la description intégrale de la structure totale du vocabulaire.

> abstraites et qu'il est représenté (dans des circonstances données) par un alphabet concret formé de lettres concrètes ; nous dirons encore que les séquences finies sont des mots abstraits, représentables par des mots concrets.
> […] dans les applications aux langues naturelles, il arrive que l'ensemble de base soit formé de lettres ou qu'il soit formé de phonèmes, et que les séquences soient des mots. Il arrive aussi que l'ensemble de base soit un vocabulaire dont les éléments sont des mots (au sens linguistique), les séquences de mots reçoivent alors des noms tels que syntagmes, phrases etc. Là encore il convient de distinguer entre le mot graphique et ce à quoi il réfère : il est bien connu que le mot 'chien' ne mord pas.
> De même il convient de distinguer entre la lettre 'a' considérée comme symbole de base et le mot monogramme 'a' par exemple dans la phrase : 'le mot vide a pour degré zéro'.
> Par contre dans les théories où l'on a en vue les seules propriétés formelles des mots, sans référence à un sens quelconque, il n'y aucune raison de distinguer entre une lettre et le mot monogramme formé de cette lettre : ce sera le cas dans la théorie algébrique du monoïde libre.

Même chez les acteurs informaticiens de la TA, la question du mot-signe est présente. Ainsi dans la définition des unités pour la TA donnée par un membre du centre de TA parisien, le CETAP, c'est la métaphore du signe qui est convoquée. La forme graphique constitue le signifiant et l'ensemble des informations grammaticales le signifié.

> Un mot (ou forme linguistique) est un couple (F, S). F est le signifiant de ce mot, sa forme graphique, et S est le signifié de ce mot (représenté pratiquement par l'ensemble des informations morphologiques, syntaxiques ou sémantiques associées à ce mot et par une définition dans le cas d'un dictionnaire monolingue ou par une liste d'équivalents de la langue-cible dans le cas d'un dictionnaire bilingue).
> Constituer un dictionnaire automatique, c'est enregistrer la suite $F_1S_1....F_nS_n$ dans la mémoire d'un calculateur électronique et établir un programme-machine (programme de consultation) pour effectuer l'opération de recherche d'une forme F_i et l'opération de lecture des informations Si. (*cf.* Dupuis 1962, p. 110.)

2.4. Idiomatismes et lexies

La confrontation avec la TA a été particulièrement importante pour deux linguistes français, A. J. Greimas et B. Pottier, qui feront partie tous deux des fondateurs de la revue *Langages* en 1966. Greimas est l'un des membres du groupe des lexicologues comprenant aussi Georges Matoré, Bernard Quemada, et Pierre Guiraud. Il a lu Shannon et Weaver ainsi que Bar-Hillel. Bernard Pottier, qui a suivi les cours de Guillaume et de Martinet, est un des rares linguistes impliqués dans des travaux concrets de TA. Leur réflexion sur le mot, à partir d'une automatisation de la traduction, sera à l'origine de modèles en sémantique lexicale tout à fait inédits.

Contrairement à la plupart des Américains impliqués dans la TA qui ne se préoccupaient pas du mot, Bar-Hillel (1955), en sa qualité de philosophe et de logicien, et de surcroît non-Américain d'origine, se pose la question du traitement automatique des mots composés [13], en particulier des expressions idiomatiques. Il propose un algorithme de traduction automatique de groupes formés de plusieurs formes graphiques, formes dont le nombre varie selon les langues et dont le sens ne peut être traduit de façon univoque. Par exemple *red herring* en anglais doit être traduit par *fausse piste* en français et par *Finte* en allemand

À la suite de Bar-Hillel, dans un article consacré aux idiotismes bilingues et monolingues, aux proverbes et aux dictons, Greimas (1960, p. 41) pose le problème plus général de la comparaison des langues au niveau sémantique :

> En ce qui concerne les idiotismes bilingues, on peut avoir deux positions extrêmes :
> i) la langue, considérée comme la structure des structures, résultat d'une évolution originale et unique, est idiomatique en tant que telle ;
> ii) un panchronisme à outrance qui, comme l'a formulé Weaver sous forme de paradoxe, le chinois n'est pas du chinois mais de l'anglais codé en chinois.
> Dans le premier cas, la comparaison entre deux langues serait impossible. Dans le second cas, au contraire, rien ou presque rien ne serait idiomatique, et la comparaison consisterait dans l'établissement de listes bilingues de structures et de syntagmes parallèles et équivalents.

Entre ces deux extrêmes, la comparaison entre deux langues, inhérente à toute tentative de traduction automatique, nécessite de définir les niveaux et les unités de comparaison. Celles-ci étant difficiles à définir *a priori*, Greimas propose de les déterminer sur la base d'une classification conçue comme la première étape d'un modèle sémantique général du lexique.

Lorsqu'il s'agit de langues dont les caractéristiques structurelles ne sont pas trop différentes, c'est au niveau de la proposition et de ses éléments constitutifs (groupes de mots) que se font les équivalences dans la mesure où, dit Greimas en se référant à la grammaire syntagmatique, un mot ou un groupe de mots peut être équivalent à une proposition entière [14].

---

13. E. Reifler (1955b) s'est intéressé au mot composé en allemand, mais son objectif était pratique et non linguistique (*cf.* note 6 ci-dessus). La machine devait être capable de déterminer les frontières internes des noms composés afin de les décomposer et de les interpréter. Pour ce faire il préconise la méthode du facteur X commun aux deux parties du mot. Ainsi 't' dans *Wachtraum* est commun à *Wach / traum* (rêve éveillé) et à *Wacht / raum* (salle de garde).

14. Il faut noter, qu'à la même époque, Halliday, dans un article intitulé « Linguistique générale et linguistique appliquée » et paru en 1962 dans les *Études de Linguistique Appliquée*, préconise la phrase comme unité de traduction. Plus on s'éloigne de la phrase, dit-il, moins il en reste. Lorsqu'on est arrivé à la plus petite unité, au morphème,

Dans la comparaison entre deux langues, c'est l'inégalité des dimensions syntagmatiques d'un groupe de mots donnés qui constitue son caractère idiomatique. Le fait qu'en russe un seul « mot », un seul élément discret, suffise pour rendre 'le vieillard' alors que le français en a besoin de deux et l'anglais de trois 'the old man', constitue le caractère idiomatique concret de chacune des trois langues. Lorsque les dimensions et les structures ne correspondent pas, le problème se pose de comparer les structures syntagmatiques non au niveau de leurs signifiants, mais au niveau de leurs signifiés.

Les idiotismes intralingues posent un autre problème, celui des unités de langue et des unités de discours (*id.*, p. 60) :

> Ce qui est comparé à l'intérieur d'une langue — car le jugement idiomatique ne peut être que le résultat d'une comparaison — c'est une sorte d'état idéal de la langue avec sa situation historique, originale. Une langue théorique, une sorte d'esperanto, celle qui n'aurait jamais été soumise à la praxis de la communication interhumaine, serait une langue où les valeurs lexicales ne se trouveraient réalisées que comme des racines, où la définition morpho-syntaxique du mot 'plein' correspondrait à sa définition sémiologique. De nombreuses langues ont d'ailleurs le sentiment très net que l'unité syntagmatique lexicale par excellence est le mot. La praxis historique de la langue déborde largement les cadres morpho-syntaxiques du mot créant des unités lexicales de type différent...

Pour Greimas, la TA pose la question de la comparaison des langues au niveau de leurs signifiés. Une telle entreprise ne peut être menée que par l'élaboration d'un modèle sémantique qui prend en compte le niveau discursif [15].

En tant que linguiste et praticien de la TA, Pottier considère que le lexique est un des domaines les plus complexes pour la TA, tant au plan formel (notamment les mots composés, comme Greimas) que sur le plan sémantique (polysémie).

Dans ses articles sur la définition des unités pour la TA, Pottier (1962a, 1962b) soutient que le mot ne peut être défini uniquement par sa forme (une suite de signes graphiques séparés par un blanc), sa signification ou le rapport forme / signification. Le mot ne peut être qu'une *unité de comportement*.

Un élément formel (mot-graphique par exemple) n'est une unité que s'il peut fonctionner librement. De ce point de vue *prendre la mouche* est soit une suite de trois mots (chaque forme gardant son autonomie fonctionnelle et sémantique) soit une suite formant une unité dans la langue. Pottier propose de nommer cette unité de langue une *lexie*. Les lexies sont les unités linguistiques

---

tout reste d'équivalence entre deux langues disparaît. Le morphème est intraduisible, le mot un peu moins.

15. Il faut mentionner l'étude des unités sémantiques complexes, comme l'étude de Jean Dubois (1962) sur le vocabulaire politique et social en France de 1869 à 1872 qui préfigure les recherches françaises en analyse du discours.

de base de la construction syntaxique. Elles peuvent être simples ou complexes (*pierre, bateau-mouche, chemin de fer,* etc.). Ce sont les lexies qui doivent être introduites dans les vocabulaires fondamentaux et les dictionnaires de TA. D'ailleurs nombre d'expérimentateurs français en TA reprennent le terme.

La lexie se caractérise par un certain nombre de critères fonctionnels dont :

> - la non-séparabilité : [16] u*ne pomme de terre froide / **une pomme froide de terre* ou *le cheval de course alezan* (inséparable) / *le cheval alezan de Jean* (séparable).
> - non qualification et non quantification : **prendre la grosse mouche.*

Pottier (1962c) s'attaque également au problème du transfert de sens des lexies d'une langue à l'autre dans le cadre d'une procédure de traduction automatique. Sur le plan sémantique, les unités lexicales n'étant pas équivalentes d'une langue à l'autre, il propose de doter chaque mot du lexique de la langue d'entrée d'indices sémantiques correspondant aux traits pertinents.

Soit son exemple (1962c, p. 201) :

	avec bras	avec dossier	haut 45 à 50 cm	au théâtre	cheval	bicyclette
	traits constitutifs			traits circonstanciels		
chaise	0	1	1	1	0	1
fauteuils	1	1		1	0	0
selle	0	0	0	0	1	0
silla	0				1	0
butaca	1	1	1	1	0	0
sillon	1	1		0	0	0
sillin	0	0	0	0	0	1

Deux objets (lexique) sont comparables d'une langue à l'autre s'il n'apparaît pas d'exclusion 0 / 1
chaise 001 est compatible avec *silla* 0
fauteuil 11 est compatible avec *butaca* 111 et *sillon* 11.

Troisième aspect auquel s'attaque Pottier : résoudre les problèmes de polysémie en établissant un degré de lexicalisation, à savoir le degré de figement en langue des lexies. En TA, le traitement des expressions idiomatiques est facilité par leur classement dans le dictionnaire selon leur degré de lexicalisation. Ainsi *cheval vapeur* ou *cheval de frise* est plus lexicalisé que *cheval de course* (qui peut avoir dans une autre langue, être représenté par un mot unique). Pour mener à bien ce travail, il faut réaliser de nombreux

16. C'est l'École de Prague et notamment Jakobson et Martinet qui ont insisté sur la non-séparabilité des parties constituantes du mot.

dépouillements-machine afin de faire ressortir les associations lexicales de discours qui correspondent à des lexicalisations en langue.

2.5. Le mot « revisité »

On a vu que les travaux de Greimas et surtout ceux de Pottier sont directement liés à une réflexion sur l'automatisation de la traduction. On peut dire que cette réflexion a suscité, au travers de la prise en compte des groupes de mots sur un plan lexical, un ensemble de travaux où la question du mot comme unité linguistique est de nouveau soulevée, et où apparaissent de nouvelles unités rendant compte de la composition nominale, comme les synthèmes de Martinet et les synapsies de Benveniste.

En effet, on peut avancer l'hypothèse que la notion de synthème proposée par Martinet en 1967 et qui constitue le principal remaniement théorique des *Éléments de linguistique générale* dont la première édition date de 1960, est issue indirectement des travaux en TA. Il suffit pour cela d'en retracer la genèse.

Il est intéressant de constater que Martinet publie en 1965 dans la revue Diogène un article intitulé « Le mot » alors qu'il y prêtait peu d'intérêt jusqu'alors et que cet article sera le premier d'une série de trois visant à promouvoir le synthème comme unité linguistique. Suivront en 1967 l'article « Synthème et syntagme » et en 1968 « Mot et synthème ».

Dans les *Éléments de linguistique générale*, Martinet ne consacre que trois paragraphes au « mot » (toujours entre guillemets), qu'il déclare difficile à définir et à délimiter comme unité linguistique pour toutes les langues. Il lui préfère la notion de syntagme autonome qu'il définit comme « une combinaison de deux ou plus de deux monèmes [17] dont la fonction ne dépend pas de sa place dans l'énoncé (*en voiture*, *avec mes valises*) » (1960, p. 109).

Il est à noter que l'article de 1965 qui remet le mot sur le métier, même si c'est pour le rejeter comme unité linguistique, paraît trois ans après le colloque sur le mot organisé par l'ATALA et la parution des premiers travaux en TA sur le lexique.

Martinet justifie ce renouveau d'intérêt pour le mot en disant que, jusqu'à présent, le linguiste ne se posait guère la question de savoir s'il existait des critères permettant pour toute langue et dans tous les cas, d'identifier et de délimiter un segment de la chaîne comme un mot déterminé. Tout en ne donnant pas les raisons pour lesquelles les linguistes commençaient à se poser ce type de question précisément au début des années 60, il évoque à plusieurs reprises dans son texte les traitements automatiques du mot. Il évoque

17. On sait que Martinet a conçu le monème, unité de la 1e articulation, en opposition au morphème néo-bloomfieldien dont il conteste le caractère exclusivement segmentable sur le plan linéaire. Il donne l'exemple du morphème *-orum* dans *dominorum* qui regroupe en fait deux monèmes amalgamés et non segmentables 'génitif' et 'pluriel'.

notamment la segmentation d’un texte écrit en formes graphiques et décrit très précisément, en la critiquant sans citer son auteur, la procédure de Harris de segmentation de la chaîne en morphèmes par des méthodes probabilistes. Sa conclusion évoque les travaux en TA en déplorant que la traduction automatique conduit à définir des unités uniquement pour le traitement du texte écrit au détriment de l’énoncé oral. Or c’est bien la traduction automatique de textes écrits qui soulève le problème de la définition d’une unité de traduction et de son adéquation du mot graphique [18].

Dans ce texte, et dans les textes suivants, Martinet continue à refuser à donner un statut scientifique au terme « mot » qui ne fonctionne pas pour toutes les langues. Il réitère sa proposition d’utiliser le syntagme, en tant que groupement de monèmes comme unité plutôt que le mot. Toutefois, si l’on en croit Martinet lui-même, le syntagme est loin d’être un concept linguistique clair : « On n’a aucun intérêt à poser, entre le monème et l’énoncé complet minimum qui est la phrase, une unité contraignante de celles dont fait nécessairement partie tout segment de l’énoncé. Libre au linguiste de délimiter des syntagmes là où son exposé gagnera en clarté. » (1965, p. 53.)

Or cette idée de considérer le syntagme comme unité rejoint celle de Greimas et Pottier dans leur recherche et la définition d’une unité sémantique pour la traduction automatique dès le début des années 1960.

En 1967, Martinet propose le synthème pour rendre compte des mots composés : « Nous proposons de désigner au moyen du terme synthème les unités linguistiques dont le comportement syntaxique est strictement identique à celui des monèmes avec lesquels ils commutent, mais qui peuvent être conçus comme formés d’éléments sémantiquement identifiables [19]. » (1967, p. 6)

Il garde la distinction entre syntagme et synthème, en introduisant les notions de monèmes libres et conjoints [20], tout en reconnaissant, dans son article de 1968 qu’il est difficile de faire la différence entre syntagme et synthème, d’une part, et entre synthème et monème unique.

Il cite explicitement les lexies de Pottier pour s’en distinguer. En définissant la lexie à partir des mots graphiques, Pottier intègre les désinences dans la lexie, alors que pour Martinet ce sont des signifiants de monèmes. Ainsi *mangeait* est une lexie simple pour Pottier et, pour Martinet, un syntagme

---

18. Plus tard, Martinet déclare réserver le terme de ‘mot’ pour désigner le mot graphique (« Que faire du mot ? » dans *Mot et parties du discours, la pensée linguistique 1*, 1986, dir. Pierre Swiggers et Willy Van Hoecke, Leuven Paris, p. 75-84.)

19. Dans son texte de 1986, les synthèmes reçoivent une définition plus large, dans la mesure où ils recouvrent à la fois les dérivés, les composés et les figements

20. Ainsi, *donnerons*, *dominorum* et *sur la table* sont des syntagmes comportant chacun trois monèmes libres. *Indésirable* et *pomme de terre* sont des synthèmes comportant trois monèmes conjoints.

groupant un monème lexical *mang* et un monème grammatical ε. Il cite également les « synapsies » mises au jour par Benveniste pour la terminologie technique. Dans son article « Formes nouvelles de la composition nominale » paru dans le *BSL* en 1966, Benveniste caractérise les synapsies, telles *modulation de fréquence* ou *avion à réaction*, par un certain nombre de traits :

- la nature syntaxique, et non morphologique, de la liaison entre les membres,
- l'emploi de joncteurs *de* et *à*
- ordre déterminé / déterminant des membres
- forme lexicale pleine et choix libre de tout substantif ou adjectif
- absence d'article devant le déterminant
- possibilité d'expansion pour l'un ou l'autre membre
- caractère unique et constant du signifié

Benveniste précise que l'extrême flexibilité paradigmatique fait de la synapsie l'instrument par excellence des nomenclatures. Martinet cite Benveniste pour signaler que toutes les synapsies sont des synthèmes [21].

Pour conclure, il est intéressant de voir que la réflexion sur l'automatisation de la traduction a conduit directement ou indirectement un certain nombre de linguistes français à s'interroger sur la définition d'une unité de segmentation de la chaîne écrite et à la manière de faire coïncider cette unité avec une unité syntaxique, une unité sémantique pour la traduction d'une langue à une autre, ou bien une unité de discours. Même si ces approches sont essentiellement centrées sur le nom et la composition nominale, elles ont contribué à conférer au lexique un statut d'objet linguistique (alors que la grammaire générative ne s'y intéressera pas avant le début des années 80). On voit comment la demande sociale, comme les besoins en terminologie, ou l'accomplissement d'un objectif pratique comme la traduction automatique, peuvent susciter une réflexion linguistique et participer à l'élaboration de nouveaux champs.

CONCLUSION

On remarquera que les linguistes structuralistes français, pour lesquels le statut du mot restait une question non résolue, ont été amenés lors de leur confrontation avec le traitement automatique à définir l'unité de base à partir du mot graphique. La forme graphique, facilement reconnaissable par la machine, pouvait être considérée comme assez proche de la définition du mot empirique, à savoir un élément signifiant du langage spontanément senti comme distinct. C'est une des raisons pour laquelle un certain nombre de ces

21. Or tous les synthèmes ne sont pas des synapsies. Notamment le critère d'absence d'article devant le déterminant de la synapsie n'existe pas pour le synthème : *Ministre du commerce*, cité dans Martinet (1986), est un synthème et non une synapsie.

linguistes, ancrés dans une tradition d'étude du vocabulaire, se sont attachés à rendre compte des groupes de mots graphiques dont il était important, en vue d'un traitement par la machine, d'évaluer le degré de figement. Même Martinet qui s'intéressait davantage à la décomposition interne en monèmes s'est trouvé entraîné dans l'étude de la composition nominale.

Cette identification de l'unité de base au mot graphique ne va pourtant pas de soi. C'est ce que montre une étude comparative. Pour le courant dominant nord-américain en TA, nul n'est besoin de définir une unité de segmentation. Morphologie et syntaxe ne sont pas distinguées, et les unités nécessaires à l'analyse syntaxique sont directement fournies par l'application d'un dictionnaire dont les entrées, on l'a vu, n'ont rien à voir avec des unités linguistiques. En revanche, pour les autres courants, comme la théorie de l'information ou le structuralisme, il est nécessaire de définir une unité de segmentation et de traduction. Mais cette unité n'est pas identifiée *a priori* au mot graphique. Au contraire, l'unité doit être découverte par la machine ou identifiée par les locuteurs. La définition de l'unité n'est en aucun cas assimilée à l'alternance de lettres et de séparateurs : tous les caractères sont traités de façon semblable dans la procédure d'Harris, et ce qu'il cherche ce sont des morphèmes. Dans l'identification de l'*informant's word*, ce sont les traits prosodiques qui fournissent la première segmentation, nécessaire au repérage des unités de traduction des langues amérindiennes. Voegelin ne cherchera d'ailleurs pas à faire coïncider ces unités avec les mots ou morphèmes décrits pour les langues occidentales. Toutefois, comme les Français, il a besoin d'une définition opératoire du mot pour accomplir son objectif pratique de traduction.

Cette référence au mot empirique pour définir une unité de traitement semble donc incontournable dès qu'il s'agit de réaliser une tâche concrète, comme la traduction, la traduction automatique ou plus généralement le traitement automatique du langage.

On remarquera enfin que, quel que soit le pays, les linguistes engagés directement dans des programmes de TA n'y ont participé tout au plus qu'une dizaine d'années. Mais il serait faux de dire que cette expérience a été inutile. Pour les Américains, hors courant chomskien, elle a permis des avancées dans le domaine des algorithmes d'analyse syntaxique ; la question de l'unité de segmentation restant toutefois subordonnée à l'analyse syntaxique sans qu'un véritable intérêt pour le lexique puisse émerger. Quant à la linguistique structurale française, elle s'est enrichie d'un nouvel objet, le « mot-composé », jusque là l'apannage des grammairiens et des lexicographes.

reçu avril 2001

adresse de l'auteur :
Université Paris 7
UMR CNRS 7597
Tour Centrale 8e étage, Bureau 801
2, place Jussieu

75251 – Paris Cedex 05
email : jacqueline.leon@linguist.jussieu.fr

Histoire Épistémologie Langage 23/I (2001) : 81-105 © SHESL, PUV

RÉFÉRENCES

AUROUX, S. (1998). *La Raison, le Langage et les Normes*, Paris, Presses Universitaires de France.

BAR-HILLEL, Y. (1955). « Idioms », Locke, W. N. ; Booth, A. D. (eds), *Machine Translation of Languages*, 14 essays, 183-193, Boston, MIT et John Wiley.

BENVENISTE, E. (1966). « Formes nouvelles de la composition nominale », *Bulletin de la Société de Linguistique*, 82-95.

CHEVALIER, J.-C., ENCREVÉ, P. (1984). « La création de revues dans les années 60. Matériaux pour l'histoire récente de la linguistique en France », *Langue Française* 63, 57-102.

CHOMSKY, N. ([1956] 1968). « Trois modèles de description du langage », *Langages* 9, 51-76.

CHOMSKY, N. ([1957] 1969). *Structures syntaxiques*, Paris, Éditions du Seuil.

DUBOIS, J. (1960). « Les notions d'unité sémantique complexe et de neutralisation dans le lexique », *Cahiers de lexicologie* 2, 62-66.

DUPUIS, L. (1962). « Un système morphologique, compromis entre les facilité de la compilation, les recherches syntaxiques et l'adaptation à de futurs programmes de TA », in *Automatic Translation of Languages*, Papers presented at NATO Summer School held in Venice, July 1962, 109-122, Oxford, Pergamon Press.

GARVIN, P. L. (1956). « Some linguistic problems in Machine Translation », *For Roman Jakobson*, 180-186.

HARRIS, Z. S. (1954). « Transfer grammar », *International Journal of American Linguistics*, 20 / 4 Translation Issue, 259-270.

HARRIS, Z. S. (1955). « From phoneme to morpheme », *Language* 31, 190-222.

HOCKETT, Ch. F. (1951). Review de Martinet *Phonology as functional phonetics : Three lectures delivered before the University of London in 1946*. London, London University Press 1949. *Language* 27, 337.

HOCKETT, Ch. F. (1953). « Analyse critique de Shannon Cl. E. et Weaver W. », *The Mathematical Theory of Communication* in *Language* 29 / 1, 69-93.

HOCKETT, Ch. F. (1954). « Translation via Immediate Constituents », *International Journal of American Linguistics* 20 / 4 Translation Issue, 313-315.

HUTCHINS, W. J. (1986). *Machine Translation, past, present, future*, Chichester, Ellis Horwood ltd.

KAPLAN, A. (1950). « An experimental study of ambiguity in context », *Mechanical Translation* 1 / 1-3.

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.

LAMB, S. M. (2000). « Translation and the structure of language », Hutchins, W. J. (ed.), *Early Years in Machine Translation*, 177-195, Amsterdam / Philadelphia, John Benjamins.

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

MARTINET, A. (1960). *Éléments de linguistique générale*, Paris, Armand Colin.

MARTINET, A. (1965). « Le mot », *Diogène* 51, 39-53.

MARTINET, A. (1967). « Syntagme et synthème », *La Linguistique* 2, 1-14.

MARTINET, A. (1968). « Mot et synthème », *Lingua* 21, 294-202.

MICKLESEN, L. R. (1956). « Form classes : structural linguistics and MT », *For Roman Jakobson*, 344-352.

MURRAY, S. O. (1994). « Theory groups and the study of language in North America », *Studies in the History of the Language Sciences 69*, Amsterdam, Philadelphia, John Benjamins Publishing company.

OETTINGER, A. G. (1955). « The design of an automatic Russian-English technical dictionary », Locke, W. N. ; Booth, A. D. (eds), *Machine Translation of Languages*, 14 essays, 47-65, Boston, MIT et John Wiley.

PECHEUX, M. (1969). *Analyse automatique du discours*, Paris, Dunod.

POTTIER, B. (1962a). « Le mot, unité de comportement », colloque ATALA *Le mot pour la Traduction Automatique et la linguistique appliquée*, 8 décembre 1962.

POTTIER, B. (1962b). « Introduction à l'étude des structures grammaticales fondamentales », *la TA* III-3, 63-91.

POTTIER, B. (1962c). « Les travaux lexicologiques préparatoires à la traduction automatique », *Cahiers de lexicologie* 3, 200-206.

REIFLER, E. (1955a). « The Mechanical Determination of Meaning », Locke, W. N. ; Booth, A. D. (eds), *Machine Translation of Languages*, 14 essays, 136-164, Boston, MIT et John Wiley.

REIFLER, E. (1955b). « Mechanical Determination of the Constituents of German Substantive Compounds », *Mechanical Translation* 2 / 1, 3-14.

SHANNON, Cl. ; WEAVER, W. (1948). *The Mathematical Theory of Communication* 1948, Urbana III.

TOGEBY, K. (1949). « Qu'est-ce qu'un mot ? », *Travaux du cercle linguistique de Copenhague*, Recherches structurales V, 97-111.

VOEGELIN, C. F. (1954). « Multiple Stage Translation », *International Journal of American Linguistics* 20 / 4, Translation Issue, 271-280.

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