STUDIA ROMANICA POSNANIENSIA
UAM Vol. 25/26 Poznań 2000

GRAŻYNA VETULANI
Université Adam Mickiewicz

QUELQUES EXEMPLES D'ANALYSE DES CORPUS EN VUE DE LA TRADUCTION

Abstract. Vetulani Grażyna, *Quelques exemples d'analyse des corpus en vue de la traduction* [Some examples of corpus analysis on the way of translation]. Studia Romanica Posnaniensia, Adam Mickiewicz University Press, Poznań, vol. XXV/XXVI: 2000, pp. 317-325, ISBN 83-232-0965-0, ISSN 0137-2475.

Intensive works by (first) computer scientists and (then) linguists in order to obtain Automatic Translation Systems started already in the late 40-ties. Since that time many things have changed in the field of tools and methodologies like for example creation of basic electronic resources as corpora and dictionaries or the lexicon-grammar framework. This paper presents some language engineering tools recently developed (with participation of the Author) for Polish and claims their interest for the Translation Community.

1. AVANT PROPOS

Les linguistes ont pour objet de description les langues naturelles qu'ils étudient à de nombreux points de vue, dans l'espace et dans le temps. On sait à quel point il est difficile de faire un choix de méthodologies actuellement présentes dans les recherches quand il faut initier les étudiants à la linguistique.

Sans oublier la linguistique historique, la linguistique comparative ou la linguistique théorique, aujourd'hui, il faudrait parler aussi des perspectives nouvelles, des recherches consacrées au traitement automatique du langage, à la technologie linguistique, ou, enfin, à la linguistique computationnelle qui utilise les résultats de la recherche linguistique générale pour exprimer ces connaissances de manière formelle (informatisée).

Aujourd'hui, presque tous les textes sont à un moment ou à un autre de leur création (édition, mise en page, correction ou impression) disponibles sur support électronique; il est donc naturel que les spécialistes cherchent à utiliser des méthodes informatiques pour les analyser (traiter). Ces données constituent une excellente source d'information sur la langue et l'analyse automatique accélère considérablement la saisie de cette information.

La vie à l'ère de l'information et le fait que celle-ci se transmette sous forme de textes qui se laissent traiter automatiquement font que les unités d'information constituent des valeurs commerciales. Comme il n'existe pas un seul langage humain, au contraire, il y a beaucoup de langues particulières répandues sur tous les continents, le marché des technologies liées au traitement du langage présente une dimension mondiale. C'est en réponse à ces besoins que s'est développée **l'industrie de la langue**. Elle recouvre tout type d'activités dans lesquelles le traitement du langage par l'homme ou par la machine constitue une part essentielle du service offert (ceci dans les télécommunications, la publicité, le courrier électronique, ou autres). L'industrie de la langue inclut des services bien établis, comme la traduction, la transcription, la rédaction technique, ainsi que d'autres plus récents, comme l'apprentissage des langues assisté par ordinateur (même si les résultats ne sont pas tout à fait satisfaisants). Elle compte de nombreux moyens informatiques de traitement des langues naturelles, comme les dictionnaires électroniques de poche, les traitements de texte, les bases de données terminologiques et enfin les systèmes de traduction automatique. La technologie linguistique implique l'élaboration et l'exploitation **d'outils** permettant de fournir de tels produits et services, et englobe toutes les formes de traitement et de transmission automatisés de la langue, parlée ou écrite. A l'heure actuelle, de très nombreuses technologies sont en jeu et chaque jour, elles deviennent de plus en plus sophistiquées. En même temps, à côté des deux appelations classiques, tels que *software* et *hardware,* fonctionne déjà le terme de *lingware* pour désigner les produits de ces technologies.

2. LANGAGE ET TECHNOLOGIE

2.1. APERÇU HISTORIQUE

(D'après le Rapport de la Commission Européenne, 1996: Langage et la Technologie. De la tour de Babel au Village Global)

1940 – En pleine seconde guerre mondiale, les chercheurs des États-Unis font une première tentative d'analyse automatique de la parole (au départ à des fins de guerre psychologique et idéologique).

1951-1954 – Zellig Harris ouvre la voie de l'analyse de la parole, grâce à son approche de la linguistique distributionnelle.

1954 – Première démonstration publique de traduction automatique, organisée par IBM (l'Université de Georgetown; en raison du climat politique, l'accent est mis sur la traduction de publications techniques russes).

1956 – À l'université d'été de Dartmouth, Herbert Simon et ses collègues proposent des modèles artificiels permettant à la machine de simuler des exercices de raisonnement et d'apprentissage humains (ainsi, on a ouvert la voie à l'intelligence artificielle).

1957-1964 – Noam Chomsky développe l'approche de Zellig Harris par une théorie générale de la langue, qui inspirera l'utilisation de différents modèles grammaticaux dans le traitement automatique du langage.

1962 – Première conférence sur la traduction automatique organisée au MIT par Y. Bar Hillel. Le thème principal est la production et l'exploitation automatique des dictionnaires. Les participants reconnaissent de façon générale que, dans l'état actuel des connaissances, une traduction automatique complète et de bonne qualité est irréalisable.

1966 – Publication du rapport ALPAC (Automatic Language Processing Advisory Council, Conseil Consultatif pour le Traitement Automatique du Langage). Discréditant l'idée qu'une traduction automatique immédiate coûterait moins cher qu'une traduction humaine, il entraîne l'arrêt des investissements massifs consentis au cours de la période précédente. (Quelques petites équipes nationales de recherche aux USA, en URSS, en Allemagne, au Japon et en France continuent à recevoir des subventions, généralement d'origine militaire.)

1970-1975 – Le développement du traitement automatique de la parole se poursuit. Les premiers systèmes de reconnaissance de la parole et les premiers synthétiseurs vocaux font leur apparition.

1980 – Le débit plus rapide des nouvelles machines et leur fiabilité accrue ouvrent des perspectives réalistes pour le traitement numérique de la parole en temps réel.

1976-1983 – La communauté européenne décide d'expérimenter des systèmes de traduction automatique disponibles sur le marché. C'est SYSTRAN qui est choisi (il est maintenu et développé jusqu'à aujourd'hui) et la politique adoptée consiste à investir dans le développement des dictionnaires et dans les machines de réécriture des programmes.

1982-1992 – Lancement du projet EUROTRA (EUROpean TRAnslator). Le projet visait à renforcer le niveau d'expertise en linguistique computationnelle en Europe. Chaque pays devait élaborer une description électronique de sa langue. En fin de parcours, on espérait aboutir à un système assurant des traductions entre les neuf langues communautaires. Ce projet a considérablement renforcé la recherche en ingénierie linguistique et a donné une base solide pour la poursuite des travaux.

2.2. RÉSULTATS OBTENUS

– Au Danemark, mise au point du système PaTrans, qui traduit actuellement les brevets d'anglais en danois et fait l'objet de nouveaux travaux en vue de traduire d'autres combinaisons linguistiques.

– Utilisation de systèmes de traduction automatique dans les grandes sociétés industrielles qui produisent de gros volumes de documentation dans des formats clairement définis, en utilisant un vocabulaire limité, spécifique au domaine traité et, éventuellement, avec un nombre restreint de structures syntaxiques. La traduction effectuée permet en général de saisir l'information, mais, dans la plupart des cas, non pas sous une forme publiable.

– Quelques succès dans l'automatisation de la traduction, comme TAO (traduction assistée par ordinateur). Des outils puissants ont été proposés au service du traducteur. Il s'agit par exemple de dictionnaires électroniques et de bases de données terminologiques, de mémoires de traduction (réduction du coût de la traduction).

Certes, toutes ces recherches ont donné de nombreuses connaissances utiles. Toutefois, la leçon que l'on a tirée de ces expériences a été telle que le domaine est plus complexe et plus difficile que prévu, il reste encore à accomplir un immense travail fondamental.

2.3. SITUATION ACTUELLE

Bien que la traduction automatique (comparable au niveau humain) reste aujourd'hui un rêve, ceci ne signifie pas que la technologie n'a rien à offrir au processus de la traduction. Au contraire, on s'intéresse à nouveau à la traduction automatique. Des technologies très nombreuses et de plus en plus puissantes sont expoitées pour traiter le langage. Ceci est certainement le résultat direct d'un énorme progrès technologique et du fait qu'il existe déjà, pour beaucoup de langues des descriptions linguistiques formelles suffisamment précises (du moins pour certains fragments de langue) pour créer des ressources de base, et, notamment, des dictionnaires électroniques morphologiques et syntaxiques.

Les systèmes de traduction automatique s'améliorent. On espère qu'ils continueront à se développer au cours des prochaines années et finiront un jour par produire des textes qu'on ne pourra pas distinguer de ceux rédigés par des traducteurs humains. La plupart des experts estiment cependant que ce n'est guère vraisemblable dans un futur proche, sauf pour des systèmes conçus pour des données très spécifiques.

Les systèmes actuels donnent de meilleurs résultats lorsque le texte original a été révisé au préalable par l'homme pour veiller à ce qu'il «convienne» à la compétence du système. Ainsi, il est envisageable que l'utilisateur averti «collabore» avec le système en s'adaptant aux limites de celui-ci. Cette attitude semble être tout à fait bien fondée, d'autant plus qu'elle est souvent observée dans la communication humaine. Elle peut donc constituer un choix méthodologique important pour la recherche des systèmes réellement utilisables.

3. INTÉRÊT POUR LES CORPUS

3.1. GRAND REGAIN DES CORPUS

«Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage» [B. Habert, A. Nazarenko, A. Salem, 1997 (J. Sinclair, 1996, p. 4)].

De vastes corpus de textes électroniques sont aujourd'hui disponibles pour beaucoup de langues. Souvent, il s'agit tout simplement de collection ou de rassemblement de textes électroniques, de recensement d'énoncés produits librement – on a alors «du texte» plutôt qu'un corpus à proprement parler dont on ne sait pas toujours très bien de quels usages langagiers il est représentatif.

L'utilisation des corpus électroniques n'est pas un phénomène neuf. B. Habert, A. Nazarenko, A. Salem (id., 1997) notent qu' «en France, un fonds de quelques 160 millions de mots a ainsi été patiemment constitué à l'Institut National de la Langue Française depuis les années soixante et constitue une base textuelle désormais accessible en ligne: *Frantexte.* Ce fonds a servi en particulier à la rédaction des dix-sept volumes du *Trésor de la Langue Française.*

La nouveauté réside dans l'enrichissement des corpus, l'accroissement de leur taille et dans l'accessibilité effective des corpus et des outils. D'abord, les corpus ne sont plus des suites de mots «nus», c'est-à-dire de simples chaînes de caractères, mais ils sont annotés (ou encore enrichis). Nous entendons par là l'ajout d'information, de quelque nature qu'elle soit: morphologique, syntaxique, sémantique, prosodique, critique... Le niveau d'annotation progresse régulièrement. Les années quatre-vingt ont été consacrées à l'étiquetage morpho-syntaxique. La décennie actuelle voit se développer les corpus arborés. Les annotations sémantiques émergent et vont se répandre. Ensuite, la taille de ces corpus ne cesse de croître. [...] Il y a juste 10 ans, le corpus *Brown*, avec son million de mots, était considéré comme un grand corpus [...]. Aujourd'hui, de nombreux centres de recherche disposent de données textuelles de millions voire de milliards de mots».

La nature des phénomènes étudiés peut parfois réclamer des données très vastes. L'expérience montre (id., p. 145) qu'un corpus d'un million de mots est bien trop restreint pour étudier la flexibilité des expressions toutes faites et un corpus de 20 millions de mots s'avère trop petit pour trouver un nombre suffisant d'occurences de toutes les expressions idiomatiques.

3.2. L'UTILITÉ DES CORPUS

À partir des faits linguistiques rassemblés dans les corpus annotés, de grande taille, variés et assortis d'outils d'exploration puissants, on peut par exemple:

– développer des dictionnaires et des grammaires descriptives,
– tester des hypothèses,
– confronter un modèle postulé aux réalisations effectives,
– observer plus finement les phénomènes langagiers,
– remettre en question une partie des postulats de la linguistique basés sur l'intuition individuelle ou sur des observations isolées,
– apprécier l'importance de différentes réalisations (aussi bien dans les langues vivantes que mortes, comme en Ancien Français, par exemple, où l'on n'arrive pas toujours à porter des jugements d'acceptabilité).

4. ANALYSE DE CORPUS

Le texte peut être analysé automatiquement. «L'analyse automatique présente certaines analogies avec de nombreux traitements informatiques: compression ou cryptage de textes, traitements statistiques, compilateurs, etc. Ici et là, la première étape des traitements consiste toujours à découper le fichier de données (texte) afin d'identifier les unités minimales de traitement (mots). Cette étape préliminaire s'appelle l'*analyse lexicale*» (M. Silberztein, 1993, p. 1).

4.1. QUELQUES EXEMPLES DE TRAITEMENT POUR LE POLONAIS

4.1.1. DICTIONNAIRES ÉLECTRONIQUES

Pour reconnaître les unités minimales, on a souvent besoin d'un dictionnaire électronique conu en vue de traitements sur ordinateur. Il s'agit d'une base de données linguistiques et des programmes permettant de la traiter. Pour certaines langues, de tels dictionnaires morphologiques ou syntaxiques (de taille différente) existent déjà, pour d'autres, ils sont en réalisation.

Pour le polonais, il existe actuellement des dictionnaires électroniques morphologiques (POLEX et GRAMLEX à Poznań et POLLEX à Varsovie) qui sont le résultat des dernières années de travail. Ainsi, beaucoup d'analyses automatiques sont déjà possibles. Vu la nature du polonais qui est une langue hautement flexionnelle (p. ex. pour le nom, il y a 7 cas, 3 genres, 2 nombres plus les *pluralia tantum*), les dictionnaires électroniques morphologiques du polonais contiennent énormément de classes différentes qui ne sont nullement représentées dans les manuels de grammaire ou dictionnaires traditionnels. Ceci est dû au fait que les dictionnaires traditionnels sont adressés à des humains, dotés de la compétence linguistique qui leur permet d'interpréter les nombreux écarts et exceptions par rapport aux règles (ou normes). On considère par contre que, faute de systèmes dotés de la compétence (et de l'intuition) des linguistes, les ressources supposées interprétables par les machines doivent obéir à des critères de précision très exigeants.

4.1.2. RECONNAISSANCE SEMI-AUTOMATIQUE DES TERMES COMPOSÉS POLONAIS

Le fait que le vocabulaire des langues naturelles contienne des mots simples et des mots composés complique leur reconnaissance automatique. Les composés du type *woda sodowa, dom spokojnej starości, służba zdrowia, pas ruchu, opieka zdrowotna* ou *Stalowa Wola* (dits *juxtapositions*, c.-à-d. *zestawienia* en polonais et non des *fusions*, c.-à-d. *zrosty* ou des *compositions*, c.-à-d. *złożenia*) exigent un traitement spécial parce que, dans l'analyse, il s'agit:

– de la segmentation de la phrase où l'analyseur doit faire la différence entre p. ex. *vase de Chine* employé comme un nom composé et *vase* p. ex. provenant ou reçu *de Chine*, entre *cordon bleu* employé dans le sens de *cuisinère* et *cordon* (de couleur) *bleu*,

– de la reconnaissance des règles de flexion (pour les langues slaves) à l'intérieur de ces groupes de mots.

Les unités composées ont des caractéristiques sémantiques (et, bien souvent, syntaxiques) spécifiques; elles sont non-compositionnelles. Pour cela, elles devraient être répertoriées dans les dictionnaires, surtout, si l'on vise des applications informatiques. Or, ce n'est pas le cas. Pire, cela ne peut être jamais le cas à 100% parce que les mots composés constituent une classe ouverte, surtout dans les sous-langages techniques. Il en résulte que les dictionnaires existants ne peuvent être qu'une source très insuffisante des composés.

L'expérience réussie pour le polonais (J. Martinek & G. Vetulani, 1997) montre que, grâce aux logiciels et aux corpus existants, on peut acquérir de telles unités semi-automatiquement, donc très vite et de manière efficace. Pour l'extraction des termes, des outils d'aide sont pourtant nécessaires. Dans notre recherche consistant en un recensement de 2500 composés, nous nous sommes servi d'un logiciel conu par Martinek (EXTRA, 1997) exploitant l'idée de Bourigaut de présegmentation de la phrase en segments «cohérents» qui, à priori, peuvent être considérés comme des phrases nominales. La relation de cohérence entre deux formes fléchies voisines a lieu si, et seulement si, elles sont susceptibles de coexister dans le même groupe nominal. La suite maximale des mots cohérents est présentée pour acceptation ou rejet par un expert. C'est finalement à l'expert de qualifier la structure comme un composé ou non. L'assistance humaine est indispensable.

4.1.3. TRAITEMENT DES STRUCTURES DU TYPE LEXIQUE-GRAMMAIRE

Si un mot ne se comprend que dans le cadre d'une phrase, ceci signifie qu'une entrée de dictionnaire ne doit pas être constituée d'une unité lexicale, mais d'un niveau minimal d'analyse (d'une phrase élémentaire) permettant de rendre compte du fonctionnement d'une unité. *Grosso modo* il s'agit de ne pas séparer le lexique de la grammaire. Ce n'est malheureusement pas le cas de la plupart des dictionnaires traditionnels polonais.

Lors d'une étude systématique sur les prédicats nominaux du polonais (recherche menée à la base d'un dictionnaire traditionnel, actuellement achevée au niveau du rassemblement des données), nous sommes arrivée à un recensement des structures dans lesquelles le prédicat apparaît en cooccurrence avec le verbe support et les arguments qu'il sélectionne. Comme les dictionnaires sont mal illustrés par des exemples de phrases attestées dans les textes, nous avons souvent manqué d'information grammaticale nécessaire (p. ex. sur le verbe ou sur la nature et ou le nombre d'argu-

ments). Pourtant, le savoir sur le choix de ces éléments peut être très important (un choix donné d'éléments = une structure = un sens). Les structures formées ainsi (souvent d'après l'intuition du chercheur), peuvent être soumises maintenant à l'analyse automatique sur des corpus authentiques afin de les confronter avec la réalité linguistique.

4.1.4. RECONNAISSANCE DES COLLOCATIONS

De la liste des structures, nous pouvons facilement extraire une liste de mots co-occurrents immédiatement, c'est-à-dire une liste de collocations (associations habituelles lexicales) utiles aussi bien pour l'enseignement que pour la traduction. On arrive ainsi à constituer un dictionnaire spécifique comportant des suites du type: *Nom Prédicatif + Verbe Support* comme:

palnąć kazanie
wszcząć alarm
pleść bzdury
zapłonąć miłością
wymierzyć policzek
uciąć sobie drzemkę
snuć domysły, etc.

Cette méthode peut être aussi utile dans tout type de classification lexicale d'une langue donnée.

4.1.5. CONCORDANCES

Les programmes de traitement de corpus recherchent dans les textes des structures morpho-syntaxiques données, et les présentent sous formes de *concordances.* Par concordance nous entendons «[...]un index de mots présentés avec leur contexte. Une fois réalisée, l'indexation des mots d'un texte, d'un auteur, d'une époque fournit des renseignements sur les références des mots et éventuellement sur leur fréquence; on offre à l'utilisateur la possibilité d'étudier parallèlement les divers emplois du même vocable» (*Dictionnaire de linguistique et des sciences du langage*, 1994, p. 108).

Cette technique basée sur l'analyse des concordances (l'observation des phénomènes dans les textes) est un outil classique aujourd'hui. Elle est utilisée pour des études stylistiques ou des recherches documentaires. Les logiciels existants permettent d'extraire très vite les concordances. Ces outils, à condition d'être basés sur un *lemmatiseur* (lemme: p. ex. *grand* pour toutes les formes fléchies: *grand, grands, grande, grandes*, autrement dit la 'forme de base' pour les réalisations distinctes d'un même phénomène), permettent de trouver les contextes pour toute forme fléchie d'un mot. Cette technique peut être très importante du point de vue de la recherche orientée

sur la traduction. Tout d'abord, parce que les dictionnaires monolingues polonais sont pauvres en exemples d'usage des mots. Souvent, les structures dans lesquelles les mots fonctionnent sont absentes ou incomplètes.

Dans le cadre du travail sur les prédicats nominaux du polonais, les concordances constituent un outil très important pour tester leurs présence et fonctionnement dans les textes.

5. CONCLUSION

Tous les exemples cités concernent les techniques d'acquisition des données de dictionnaires (d'unités simples ou composées, de structures entières, etc.), le premier outil de chaque interprète ou enseignant.

BIBLIOGRAPHIE

Dubois, J., Guespin, L., Giacomo, M. C. & Marcellesi, J.-B., Mével J.-P. (1994), *Dictionnaire de linguistique et des sciences du langage*, Larousse, Paris.

Habert, B., Nazarenko, A., Salem, A. (1997), *Les linguistiques de corpus*, Armand Colin, Paris.

Langage et Technologie. De la Tour de Babel au Village Global, (1996), Luxembourg: Office des publications officielles des Communautés européennes.

Martinek, J. & Vetulani, G. (1997), *A method of computer aided acquisition of compound terms*, projet COPERNICUS 621, GRAMLEX.

Silberztein, M. (1993), *Dictionnaires électroniques et analyse automatique de textes*, Masson, Paris, Milan, Barcelone, Bonn.