Meta
Journal des traducteurs
Translators' Journal

META

REFTEX —un progiciel pour la traduction assistée par ordinateur

Poul Sören Kjaersgaard

Volume 34, Number 3, septembre 1989

1. Actes du Colloque Les terminologies spécialisées : Approches quantitative et logico-sémantique et 2. Actes du Colloque Terminologie et Industries de la langue

URI: https://id.erudit.org/iderudit/003322ar
DOI: https://doi.org/10.7202/003322ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Kjaersgaard, P. S. (1989). REFTEX —un progiciel pour la traduction assistée par ordinateur. *Meta*, *34*(3), 496–501. https://doi.org/10.7202/003322ar

REFTEX — UN PROGICIEL POUR LA TRADUCTION ASSISTÉE PAR ORDINATEUR

POUL SÖREN KJAERSGAARD
Université d'Odense, Danemark

INTRODUCTION

Il existe deux approches différentes pour l'utilisation des ordinateurs dans le processus de traduction. L'une qui consiste à vouloir faire traduire par l'ordinateur un texte, éventuellement suivi d'une phase de postédition humaine (traduction automatique ou traduction automatique assistée par l'homme), et l'autre qui consiste à vouloir faire exécuter par l'ordinateur des éléments fastidieux du processus de traduction. Cette dernière approche, souvent désignée traduction humaine assistée par la machine (THAM/THAO) recouvre une large gamme de processus différents, tant pour le degré d'intervention de la machine que pour les méthodes.

Quant aux méthodes, la plupart des systèmes THAM ont ceci de commun qu'ils s'appuient sur un dictionnaire ou une banque de terminologie pour un ou plusieurs domaines spécialisés. Dans ces dictionnaires, on retrouve le(s) équivalent(s) d'une ou de plusieurs langue(s) ainsi que des définitions ou des explications des mots-vedettes.

Dans le système REFTEX, l'approche est différente en ceci qu'on met l'accent sur le contexte du mot ou de l'expression à traduire. Ceci implique qu'il n'y a aucun dictionnaire — au sens propre de ce mot — dans le système, mais un nombre de textes déjà traduits, dits textes de référence ou en anglais ***reference texts*** d'où d'ailleurs le nom de l'approche.

ÉLÉMENTS CARACTÉRISTIQUES DU LOGICIEL REFTEX

L'objectif de REFTEX est de fournir au traducteur humain un auxiliaire supplémentaire, se greffant sur ceux dont il dispose normalement.

Le logiciel sera notamment utile dans deux situations : *primo*, lorsque le traducteur ne trouve pas une traduction utile dans les dictionnaires ; *secundo*, lorsque le dictionnaire bilingue ne donne aucun critère de distribution des équivalents qu'il propose.

La caractéristique principale par rapport à d'autres systèmes de logiciels de THAM est le remplacement des dictionnaires par des concordances bilingues, c'est-à-dire des extraits de textes équivalents en deux langues, la langue-source et la langue-cible.

L'approche employée dans le logiciel REFTEX constitue une combinaison de l'usage (traditionnel) de concordances bilingues en philologie et de l'usage des textes déjà traduits, principe dont on se sert, par exemple, dans les services de traduction des Communautés européennes. Le logiciel a finalement été conçu comme un système interactif, dont le traducteur se servira au cours du processus de traduction. Ce mode conversationnel fonctionne, dans la mesure du possible, en langage naturel.

LE LOGICIEL DE L'APPROCHE REFTEX

Le logiciel consiste en deux programmes : ARBORAL et REFTEX.

Le programme ARBORAL

Le premier, ARBORAL, est un programme-support qui transforme un texte de référence, c'est-à-dire, un texte original en langue-source et sa traduction en une ou plusieurs langues-cibles, en une nouvelle structure de données équivalente, permettant ainsi la reconstruction du texte original, mais qui contient des informations supplémentaires, telle la fréquence de chaque mot-forme et la position de l'occurrence suivante, ce qui facilitera la construction d'une concordance.

La structure de données est organisée comme deux enregistrements, le premier correspondant à une arborescence binaire à laquelle sont ajoutées des informations supplémentaires (fréquence absolue et position dans le texte de la première occurrence de chaque mot). Cet enregistrement constitue, si l'on veut, un dictionnaire des mots-formes d'un texte, c'est-à-dire tous les mots différents, y compris les formes fléchies.

Le second enregistrement est une structure de liste contenant des informations sur chaque mot-occurrence du texte. Ainsi, il contient un pointeur indiquant la position de chaque mot-occurrence dans l'arborescence/le dictionnaire du premier enregistrement, un pointeur indiquant s'il existe encore des occurrences du mot plus loin dans le texte, et si oui, leur position. Enfin, cet enregistrement indique pour chaque mot-occurrence la position du premier mot du paragraphe qui contient le mot.

Une fois construite, la structure des données sera sauvegardée dans un fichier à partir auquel le programme REFTEX pourra accéder.

Avant l'exécution du programme ARBORAL, il faut procéder à une brève préédition du texte de référence, qui consiste en l'insertion manuelle de jalons, c'est-à-dire des délimitateurs de chaque paragraphe, ainsi qu'un numéro l'identifiant de manière univoque. Un paragraphe contient normalement deux ou trois périodes, mais rien n'empêche qu'il soit plus long. L'important est qu'on arrive à définir un paragraphe suffisamment large pour que l'utilisateur du programme, le traducteur, puisse se faire une idée du contexte du mot dont il cherche la traduction. Dans un second temps, ces jalons sont insérés dans le texte parallèle, la version traduite du texte original. Ceci est un point capital dans l'approche choisie, car sans cette préédition, il serait impossible d'obtenir la sortie de passages parallèles. Les modes d'expression ainsi que l'ordre des mots de deux langues, voire de langues apparentées, sont si différents qu'il ne sert à rien de demander à l'ordinateur d'imprimer les n'ièmes lignes du texte-source et du texte-cible — comme on l'avait d'ailleurs fait dans les essais primitifs d'informatisation de concordances.

Le programme REFTEX

Le second programme du logiciel, qui porte d'ailleurs le nom de l'approche est le programme dont se sert le traducteur humain. À la limite, celui-ci n'a même pas besoin de connaître l'existence du programme-support, ARBORAL.

Dans la phase initiale du programme, le traducteur entre les noms de la paire de textes de référence, dans laquelle il juge opportun de chercher des traductions aux mots et expressions qu'il doit connaître pour les besoins de la traduction en cours.

Après la saisie de ces textes, le programme lui demande d'entrer le premier mot dont il veut connaître la traduction. Si ce mot est présent, l'ordinateur affiche sur l'écran le premier passage le contenant ainsi que le passage parallèle de la langue-cible. En fonction de ses connaissances des deux langues concernées et de ses connaissances pragmatiques, le traducteur aura donc à décider si la proposition contenue dans le passage cible est appropriée à la traduction en cours. Le critère de cette décision sera normalement le degré de parallélisme entre le passage-source et le contexte du mot à traduire. Si le traducteur juge acceptable la traduction proposée, il passera au problème suivant. Si, par contre, le premier passage affiché ne contient pas une proposition de traduction que le tra-

ducteur juge appropriée, il passera aux éventuels passages suivants, et ainsi de suite jusqu'à ce que le texte-source soit épuisé. Dans ce dernier cas, le programme sauvegardera le mot dans une matrice afin de permettre ultérieurement sa recherche dans un texte de référence alternatif. Le traducteur ayant épuisé sa liste de problèmes, le programme vérifiera s'il existe des mots, soit non repérés, soit repérés dans un contexte non satisfaisant. Dans la négative, REFTEX aura contribué à résoudre les problèmes du traducteur. Dans le cas contraire, le programme cherchera à repérer ces mots dans une autre paire de textes de référence.

PROBLÈMES DE MÉTHODE

L'informatisation des concordances pose deux problèmes au linguiste : le repérage des formes fléchies d'un mot ou d'un lemme et la polysémie inhérente à toute langue naturelle.

L'ordinateur, incapable du moindre raisonnement intellectuel et à fortiori incapable de procéder à des rapprochements sémantiques, ne décèle aucun lien entre deux mots, tels *maison — maisons*, le premier étant une chaine de six caractères, et le second en possédant sept. Le traducteur humain, par contre, ne se soucie guère de cet aspect lorsqu'il doit apprécier la traduction d'un mot, puisque le nombre de mots ayant des signifiés différents au singulier et au pluriel (tels ciseau-ciseaux) est infime.

Pour rendre l'ordinateur capable de trouver toutes les formes fléchies d'un mot, il faudra par conséquent ou bien organiser la structure des données de façon que toutes les formes apparentées soient détectées, ou bien concevoir la procédure de repérage de façon qu'on atteigne l'objectif visé. Il existe en principe, trois approches. À partir d'un *tronc commun*, la chaîne de caractères commune à toutes les formes fléchies, on peut donc repérer toutes les occurrences de *maison* et de *maisons* dans un texte donné.

Cette approche risque cependant de fournir des informations non pertinentes, notamment dans le cas des mots dérivés et des mots composés, aspect caractéristique des langues germaniques.

On pourra alternativement développer un analyseur morphologique, capable de segmenter les mots en racines et flexifs.

La troisième solution est d'initier le repérage à partir du lemme, la forme canonique, tel le singulier masculin des adjectifs. C'est la solution qui a été choisie dans REFTEX.

À partir de l'adjectif *fort*, on aura *forte-forts-fortes*. À partir du nom espagnol *pais*, on aura *paises* et ainsi de suite. Il s'est avéré possible de créer un algorithme permettant la généralisation des formes fléchies de noms, adjectifs, participes et verbes (les classes productives) en français et en espagnol.

Le second problème reflète l'ambiguïté inhérente à toute langue naturelle. En cherchant à repérer l'adjectif *ferme*, on risque en effet de buter sur des occurrences du nom *ferme* ainsi que sur des occurrences du verbe *fermer*.

Dans REFTEX, on a préféré utiliser une approche pragmatique, qui permet au traducteur humain de restreindre le champ de recherche en ajoutant au mot recherché un mot avec lequel il entre en collocation. Cette approche, ayant quelques traits communs avec les règles locales dans les essais de traduction automatique de première génération, permet notamment la recherche de mots composés, et dans certains cas, elle permet aussi la désambiguïsation entre verbe et nom ou bien entre article et pronom personnel.

Si le mot colloqué n'est pas présent dans le texte, ou qu'il ne remplisse pas les exigences du traducteur, celui-ci peut en choisir un autre plus ou moins synonyme, si bien que la solution s'assimile à un dictionnaire de synonymes. Si, par exemple, le texte de

référence contient le mot *sidérurgie*, mais non la collocation *problèmes de la sidérurgie*, le traducteur pourra choisir alternativement *difficultés de la sidérurgie*.

L'inconvénient de cette solution est évidemment qu'elle ne résout pas toutes les ambiguïtés. Pour autant qu'elle soit réalisable, une solution complète demanderait soit un haut degré de préédition, marquant chaque ambiguïté d'un code, soit un analyseur sémantique. Son avantage, par rapport à l'analyseur sémantique, est sa simplicité relative (deux procédures récursives), et par rapport aux dictionnaires, l'approche permet de repérer aussi bien des collocations lexicalisées, tel *chemin de fer*, que celles qui ne le sont pas telles *les problèmes de la sidérurgie*.

REMARQUES MÉTHODOLOGIQUES

Puisque REFTEX est une approche différente par rapport à l'approche dictionnairique, plus répandue en THAM, quelques remarques méthodologiques sur la différence entre ces deux méthodes s'imposent.

Les unités enregistrées dans le dictionnaire et la concordance sont différentes. Dans la concordance, on enregistre le mot-forme, chaque fois qu'il apparaît, alors que dans le dictionnaire, c'est le lemme, le mot-type qui en principe n'est enregistré qu'une fois. Les unités ne coïncident que dans le cas des mots-types invariables dépourvus de paradigme.

À ces différences formelles s'en ajoute une autre, plus réelle : la description du sens des mots. Que le traducteur se serve de l'une ou de l'autre approche, il vise en effet le même objectif : traduire le sens d'un texte original dans une autre langue.

Dans l'approche contextuelle qu'est REFTEX, le sens du mot est décrit à l'aide du contexte, et éventuellement à l'aide de la traduction dans une autre langue. L'aspect caractéristique de cette approche est le raisonnement inductif qui consiste à établir une règle générale, le sens d'un mot et sa traduction, à partir de cas spéciaux (la syntaxe et la combinatorique du mot).

Dans l'approche dictionnairique, le sens du mot est décrit à l'aide d'autres mots (paraphrase) ou à l'aide des équivalents dans une autre langue.

La différence entre les deux méthodes se résume dans la question de savoir si les mots d'une langue possèdent chacun un sens indépendant ou bien si le sens est influencé par le contexte. Dans cette perspective, la différence se ramène en effet à la dichotomie saussurienne de la langue (dictionnaire) et de la parole (concordance).

Les deux méthodes possèdent, chacune, des avantages et des inconvénients. S'il n'est ni exhaustif, ni cohérent (non contradictoire) au sens mathématique, le dictionnaire fournit normalement une idée assez précise du sens des mots. De tels renseignements sont même indispensables dans le cas où le traducteur ignore complètement le mot.

Le problème de la description dictionnairique est qu'il est parfois difficile de définir le sens d'un mot et à fortiori d'indiquer précisément la distribution d'une série d'équivalents. En définissant un mot à l'aide d'autres mots, à leur tour définis, on risque en effet d'arriver à une régression à l'infini. Le problème se voit accentué quand il s'agit de définir des critères de distribution aux équivalents d'un mot. Alors que certains équivalents sont utilisés en fonction de critères syntaxiques ou sémantiques formellement décidables (tels *savoir / connaître* pour le verbe anglais *know*, d'autres dépendent de la situation, ce qui rend la tâche des lexicographes extrêmement compliquée.

L'avantage de l'approche contextuelle est avant tout qu'elle permet au traducteur de faire un choix entre plusieurs équivalents proposés en fonction du contexte.

L'approche qui s'appuie sur le contexte pose cependant, à son tour, de nouvelles questions, notamment l'importance qu'on peut lui accorder ainsi que la question de savoir s'il a partout et toujours la même fonction.

L'importance du contexte linguistique n'est qu'une hypothèse qui reste à vérifier. À coté du contexte linguistique, d'autres facteurs, tel le savoir partagé du destinateur et du destinataire et le contexte pragmatique (la situation dans laquelle le mot est énoncé) jouent certainement un rôle.

L'argument qu'on avance en faveur du contexte linguistique est que s'il est licite de s'appuyer sur celui-ci dans les énoncés écrits, c'est qu'il est le seul moyen dont disposent le destinateur pour faire passer son message et le destinataire pour le déchiffrer. En combinant le contexte linguistique avec son savoir pragmatique, l'homme réussit à comprendre un texte. L'ordinateur, au contraire, qui ne possède pas ce savoir, n'y parvient pas.

La seconde question qui se pose est de savoir si le contexte est également important dans n'importe quel (sous-)vocabulaire. En répondant par la négative, l'approche contextuelle sera moins utile.

Aucune réponse définitive n'a été apportée à cette question. Mais il semble plutôt raisonnable de supposer que plus le sous-vocabulaire est spécialisé, moins le contexte joue un rôle. Il est aussi probable que le nombre d'équivalents sera inférieur. Dans pareils cas, l'utilité d'une approche contextuelle se réduit à sa capacité de repérer les collocations non lexicalisées.

CONCLUSION

L'approche du progiciel REFTEX s'inscrit dans une tradition qu'on retrouve dans plusieurs projets américains de THAM, dont ITS (Interactive Translation Systems), ALPS et Weidner.

Le linguiste américain, Martin Kay, avait recommandé, dans un article de 1980, «The Proper Place of Men and Machines *in Language Translation*», de concevoir et de développer des logiciels capables de tirer profit des capacités spécifiques de l'homme et de l'ordinateur. Il avait notamment prôné la création d'un *work bench*, c'est-à-dire une station de travail apte à intégrer différents auxiliaires de traduction. C'est dans cette optique, le logiciel REFTEX pourrait devenir une composante de la station de travail du traducteur.

L'approche contextuelle du REFTEX permettrait également d'atteindre d'autres objectifs ou de faciliter leur réalisation. Ainsi, un logiciel du type REFTEX pourrait contribuer à normaliser les traductions en ce sens qu'il facilite le travail des réviseurs linguistiques. On peut également supposer que le logiciel avec quelques modifications, pourrait fournir la matière première aux lexicographes.

Enfin, si l'on accepte la validité de l'approche contextuelle, ce logiciel pourrait constituer le point de départ de l'élaboration de règles de traduction.

Notice technique

Le logiciel REFTEX a été programmé dans le langage PASCAL (le dialecte Poly-Pascal). Il existe, comme c'est indiqué dans le texte, deux programmes indépendants ARBORAL et REFTEX, de 67 kilo-octets et de seize kilo-octets respectivement en version non compilée. Les programmes tournent sur des micros dont le système d'exploitation est le DOS, donc sur tout PC compatible IBM.

À l'heure actuelle, il existe des versions danoise, française et anglaise des programmes en ce qui concerne les textes-guides, les questions posées au traducteur, etc. Les programmes sont utilisables pour n'importe quelle paire de languages, pouvu qu'il existe des textes de références entre elles.

La version actuelle des programme est à considérer comme un prototype. Ceci signifie que le programme reste en principe ouvert à des améliorations futures. C'est

aussi la raison pour laquelle la version actuelle du programme est soumise à des contraintes quant à la longueur des textes de référence. Normalement, elles ne peuvent dépasser cinq à six pages, chacune. L'origine de cette contrainte est à trouver dans le maniement des structures de données par l'ordinateur, et on travaille actuellement à réduire, voire éliminer, cette contrainte.

Des informations ainsi qu'une documentation plus amples seront disponibles en s'adressant à l'auteur. En lui envoyant une disquette, il est possible d'obtenir une version du programme pour usage non commercial.