La traduction automatique, quel avenir ? Un exemple basé sur les mots composés

Annik Baumgartner-Bovier

Département de Linguistique
Université de Genève
<annik.bovier@lettres.unige.ch>

Résumé

La traduction automatique est un domaine en plein développement, ce qu'illustrent bien les systèmes de traduction REVERSO et SYSTRAN. Tous deux sont des produits commercialisés, qui offrent plusieurs paires de langues. Afin de pouvoir comparer efficacement les résultats des deux systèmes, le champ de l'analyse a été réduit aux mots composés et à la traduction de l'allemand en français. REVERSO et SYSTRAN utilisent une méthode de traduction similaire et sont ainsi confrontés aux mêmes difficultés, à savoir la décomposition des termes dans la langue source et leur traduction dans la langue cible. Malgré ces ressemblances évidentes, leur capacité de traduction diffère et met en lumière l'état actuel de la traduction automatique.

1. Introduction

Cet article[1] porte sur la traduction automatique des mots composés allemands en français à l'aide des systèmes de traduction REVERSO et SYSTRAN, qui sont des traducteurs allemand-français et inversement. Les deux systèmes utilisent une technologie ancienne. REVERSO a été créé dans les années 1970-1980 par des chercheurs russes et a été commercialisé en 1998 par l'entreprise Softissimo. SYSTRAN a été créé par Peter Toma dans les années 50 pour la traduction du russe et de l'anglais et en 1964 pour le russe et l'allemand. Depuis lors, il y a certes eu des développements, mais les programmes restent toutefois anciens par rapport à des produits créés actuellement.

La traduction automatique offrant un champ très vaste, une restriction à un domaine s'impose. Les mots composés offrent une étude intéressante de la traduction automatique, car il devrait être possible d'établir, d'une part, des règles déterminant leur formation dans les langues source et cible, et d'autre

[1] Cet article est basé sur mon mémoire de DEA (2002), ayant pour directeur le Professeur Eric Wehrli et pour jury le Professeur Jacques Moeschler.

part, des règles mettant en correspondance ces structures. Par exemple, l'usage des prépositions en français peut-il être expliqué d'un point de vue sémantique ? Est-il possible d'établir des règles de formation sur la base de ces caractéristiques sémantiques ? Est-il tout simplement du domaine du réalisable de concevoir la création de règles gérant les différentes possibilités de traduction de manière automatique ?

Le but de cet article est d'une part d'exposer les capacités et les manques des deux systèmes par le biais d'une comparaison en se limitant à la traduction des mots composés et d'autre part de proposer des solutions pour résoudre les manques relevés dans les deux systèmes. Pour ce faire, le corpus employé est composé de trois types de textes[2] : un texte de mots composés confectionnés de manière *ad hoc*, un texte de mots composés puisés dans des articles du journal *Die Zeit* et un texte de mots composés relevant du registre de la mode et de la couture. Ce corpus permet de traiter à la fois un vocabulaire courant et spécifique. La suite de cet article se déroule comme suit. Tout d'abord, la section 2 présente les deux systèmes de traduction REVERSO et SYSTRAN. La section 3 décrit les mots composés. La section 4 explique la méthode employée pour étudier la traduction automatique et en expose les résultats. La section 5 détaille les problèmes rencontrés lors de la traduction et tente de les résoudre. La section 6 s'intéresse aux problèmes non résolus.

2. Les systèmes REVERSO et SYSTRAN

Les deux systèmes s'inscrivent dans une même tradition : ils utilisent une approche directe par le biais des dictionnaires bilingues et une approche indirecte par le système de transfert, qui s'occupe de la traduction des prépositions, des idiomes standards et du transfert structurel. Le transfert correspond au passage d'une structure d'une langue à celle d'une autre. Ils sont tous deux modulaires et multicibles, c'est-à-dire qu'ils peuvent développer plusieurs paires de langues. Leur méthode de traduction est basée uniquement sur des critères linguistiques, à savoir l'utilisation de règles syntagmatiques et de règles de correspondance. Les premières renvoient aux règles de réécriture, qui produisent la décomposition des phrases en catégories grammaticales, notamment la phrase est réécrite en syntagme nominal et en syntagme verbal ($S \rightarrow NP + VP$). Les secondes mettent en correspondance les règles syntagmatiques des langues source et cible. Cette méthode se distingue des approches basées sur la connaissance, sur les statistiques ou bien encore sur les exemples, car elle est la seule à avoir donné lieu à des produits commercialisables.

[2] Pour un accès aux textes, se référer à Bovier (2002).

Les deux systèmes se différencient toutefois l'un de l'autre par leur mode d'intervention. SYSTRAN est non-interventionniste, c'est-à-dire que les utilisateurs ne peuvent pas influencer l'ordinateur, ils doivent attendre les résultats sans pouvoir intervenir dans la traduction. Ce type de stratégie s'insère dans les débuts de la traduction automatique. Le traitement entièrement automatique est critiqué, car la traduction implique des capacités humaines que l'ordinateur ne possède pas, telles que la connaissance du monde réel.

En revanche, REVERSO est un système interactif, c'est-à-dire qu'il autorise une intervention directe pendant le traitement automatique. Dans le cas d'une intervention interactive, l'ordinateur peut demander à l'utilisateur de compléter des informations linguistiques, de confirmer des décisions ou bien encore de sélectionner un mot ou une traduction parmi un certain nombre d'alternatives. La sélection d'un mot s'effectue lorsqu'une distinction existe entre la langue source et la langue cible, notamment dans le cas du terme français *mur* qui a deux traductions en allemand, *Wand* et *Mauer*. La sélection d'un mot ou d'une phrase a lieu lorsqu'il y a une ambiguïté lexicale ou syntaxique dans la langue source ou dans la langue cible.

Le choix d'un mode interactif permet donc de simplifier le lexique : en effet, si le système ne doit pas choisir seul l'alternative des équivalents ou résoudre seul l'ambiguïté de la catégorie grammaticale, le lexique n'a pas besoin de contenir d'informations complexes sur la grammaire, la sémantique et la syntaxe. Le système offre une liste d'alternatives à l'utilisateur et c'est ce dernier qui choisit. Les questions sont sous le contrôle de l'utilisateur, puisqu'il construit le dictionnaire qui dirige l'interaction. Ce type de système a la possibilité d'apprendre les corrections au lieu de répéter sans cesse les mêmes questions en insérant un apprentissage probabiliste, qui donne la préférence à une modification qui survient souvent. Le système apprend alors cette modification et corrige par ce biais tout seul certaines erreurs. Cette distinction majeure souligne le caractère vétuste du système SYSTRAN.

Les systèmes REVERSO et SYSTRAN sont semblables sur le mode du fonctionnement, mais différents sur le mode de la technologie. En effet, tous deux utilisent un dictionnaire général et peuvent avoir recours à des dictionnaires spécialisés. Ils disposent d'une syntaxe locale et d'une sémantique très restreinte, à savoir il est par exemple possible de différencier une personne d'un objet inanimé. Tous deux sont considérés comme des systèmes directs de par le rôle des dictionnaires et comme des systèmes de transfert à cause du module de transfert. Cependant, REVERSO et SYSTRAN diffèrent quant à leur technologie, puisque REVERSO est un programme plus récent que SYSTRAN, qui utilise des chaînes d'octets pour reproduire les mots. Ainsi, les deux systèmes ont une même approche, principalement une analyse lexi-

cale, mais ils font intervenir une réalisation computationnelle différente, plus archaïque dans le cas de SYSTRAN. Pour ces raisons, il semble pertinent de comparer ces deux systèmes. Quels peuvent être les impacts d'une technologie plus ancienne sur la traduction ? Quelles sont les différences de traduction entre deux systèmes exploitant une même méthode ?

Avant de répondre à ces questions en analysant les résultats obtenus, il convient d'introduire le domaine de cette étude, à savoir les mots composés.

3. Les mots composés

Le mot composé est un mot, c'est-à-dire qu'il possède les mêmes propriétés que le mot auxquelles il ajoute quelques particularités. Le mot correspond à une combinaison de morphèmes et il dispose d'une certaine mobilité de position dans une phrase. Le mot composé a pour particularité de ne supporter ni modification, ni substitution. Il n'est en effet pas possible d'insérer un élément supplémentaire dans un mot composé, notamment l'expression *une pomme frite* ne peut pas être modifiée en *une pomme très frite* ; on ne peut pas davantage substituer un synonyme à un des termes du mot composé, tel que *les beaux-arts* par *les jolis arts*.

Les mots composés peuvent se définir en ces termes (cf. Catach 1981, 15) :

> « Un mot, quoique formé d'éléments graphiquement indépendants, est composé dès le moment où il évoque dans l'esprit non les images distinctes répondant à chacun des mots composants, mais une image unique ».

Plus exactement, si l'on combine les éléments de la phrase *elle est belle de nuit seulement*, on obtient *belle-de-nuit* (cf. Bouvier 1999, 32). Les trois mots contigus, représentant chacun des concepts distincts, sont interprétés comme un concept unique, ne découlant pas entièrement de la combinaison régulière des concepts qui le composent : la phrase initiale est modifiée en une expression possédant un unique sens.

Un exemple différent permet d'illustrer cette capacité de concept unique : *chien-loup* correspond au concept unique d'un chien. Le premier terme donne le sens à l'expression au détriment du deuxième terme. On parle de concept unique, car le mot composé est lexicalisé. La lexicalisation consiste en l'enregistrement d'un nouveau mot. Par exemple, le mot composé *tasse à thé* renvoie à un référent unique, la tasse. Il s'agira toujours d'une tasse à thé, même si on la remplit de café. Par contre, l'expression *tasse de thé* est associée au fait que la tasse est remplie d'un liquide qui est du thé. Dans ce cas, le mot composé renvoie à un autre référent unique, le thé et non plus la tasse. Le mot composé est une expression qui s'est créée à partir d'une autre unité lexicale : il s'agit d'un processus et non d'un état.

3.1. Les mots composés en français

Les mots composés en français possèdent une graphie très variable : ils peuvent être soudés en un seul mot, notamment *gentilhomme*, inclure des traits d'union comme dans *arc-en-ciel*, introduire une apostrophe comme dans *aujourd'hui* ou encore user d'espace, comme dans *hôtel de ville*. Si la graphie est une des difficultés des mots composés en français, le mode de l'accord en est une autre : la marque du pluriel peut être nulle (*des cache-pot*), simple (*des couvre-lits*) ou double (*des cerfs-volants*).

Les prépositions sont également source de problème. La tendance actuelle est de donner la préférence à la préposition *de*, car celle-ci est la plus utilisée. Cependant, chaque préposition possède des propriétés intrinsèques, qui peuvent être décrites.

3.1.1. *La préposition* de

La préposition *de* correspond à la préposition la plus utilisée. Elle désigne notamment la direction, l'origine, le sujet, l'objet, la partie et la totalité. Elle est abstraite, générale et polyfonctionnelle. En effet, l'énoncé *le train de Paris* n'indique pas si le train est en provenance de Paris ou s'il a pour destination Paris. La préposition *de* permet donc de produire ces deux propositions.

La préposition *de* a des caractéristiques bien définies.

- Elle a une valeur intrinsèque étendue : on peut l'utiliser pour indiquer différentes relations, dont la provenance (*lait de vache*, le lait qui est produit par une vache), le matériel (*cheville de bois*, une cheville qui est en bois), le lieu (*une classe de neige*, une classe dans la neige), la relation génitive (*la maison du professeur*, la maison que possède le professeur), etc. La préposition *de* implique une post-détermination, à savoir le second nom (*N2*) s'appuie sur le premier (*N1*), d'où on obtient la relation *N1 de N2*.
- La préposition *de* a une valeur grammaticale : elle est un indice d'infinitif et elle permet de former un complément de nom à valeur adjectivale, tel que *un poète de génie* pour l'expression *un poète génial*.
- La préposition *de* peut être un élément d'article de matière (*de l'air conditionné*) ou un élément d'article indéfini (*des grands-parents*), de même qu'elle peut être elle-même un article de matière (*une cabine d'air conditionné*) ou un article indéfini (*de bons grands-parents*). La préposition *de* peut également être un article de négation (*pas de grands-parents*).

Dans le cas des mots composés, seuls les deux premiers points sont pris en compte, car seule la formation interne du composé nous intéresse. L'article ou l'élément d'article dépend de la structure de la phrase et non de la construction interne du composé. Lors de la présence d'un article dans la structure interne d'un composé, celui-ci est pris en compte ; mais dans la description ci-dessus, où l'article correspond au déterminant d'un groupe nominal et non au déterminant utilisé pour la création d'un composé, l'article est traité lors de l'analyse syntaxique de la phrase. Le premier point est particulièrement pertinent, car il permet de définir les diverses utilisations de la préposition *de*, ce qui offre la possibilité d'établir des règles quant à l'usage de cette préposition dans un composé plutôt qu'une autre.

3.1.2. *La préposition* à

La préposition *à* a quatre fonctions selon Bossang (cf. Wolf 1990, 130) :

- Elle indique l'information du but (*armoire à pharmacie*).
- Elle met en évidence le principe de fonctionnement des objets ou des individus (*avion à réaction*).
- Elle souligne la quantité (*vol à longue distance*).
- Elle exprime les caractéristiques générales des objets (*bombe à retardement, robe à bretelles, patin à roulettes*).

La préposition *à* se distingue de la préposition *de*, car elle indique une fonction spécifique, alors que la préposition *de* marque la cohésion de référence entre les termes. Les expressions *un verre à vin* et *un verre de vin* mettent bien en évidence les différences : le premier cas souligne la spécificité et l'utilisation que l'on fait du verre, alors que le second cas indique uniquement que le verre contient du vin.

Les prépositions *de* et *à* sont les plus fréquentes dans la formation des composés, aussi il est important de bien les différencier. Les définitions exposées ci-dessus devraient permettre d'établir l'utilisation de chacune d'elles. Les autres prépositions intervenant en moindre mesure, leur interprétation sémantique semble moins pertinente[3].

3.2. Les mots composés en allemand

Notre sujet portant sur la traduction des mots composés de l'allemand en français, il ne me semble pas nécessaire de s'intéresser à la reconnaissance des mots composés en français, puisque nous travaillons dans l'autre sens de traduction. La reconnaissance des mots composés en allemand est évidente, puisque cette langue concatène les mots entre eux. Les mots composés sont

[3] Pour une description détaillée des différentes prépositions, se référer à Bovier (2002).

généralement constitués de deux membres, mais également de trois membres et voire plus. Les mots composés peuvent être formés sur la base de constituants de différentes catégories, à savoir de substantifs, d'adjectifs, d'adverbes, de prépositions, d'abréviations ou encore de verbes. On retrouve certes deux types de graphie, à savoir la juxtaposition des mots (*warmblutig*) ou l'usage des traits d'union (*Internet-Ära*), mais ces deux types de constructions se distinguent aisément des mots simples.

D'autre part, le mode de l'accord ne pose pas de difficulté, puisque dans les mots composés, seul le dernier terme prend l'accord. La difficulté ne réside donc pas dans la reconnaissance d'un mot composé allemand, mais bien dans la traduction de ce mot : le fait de concaténer les mots entre eux rend la traduction ardue, car il faut décomposer correctement le mot composé et trouver des règles de correspondance entre les langues. Ce problème sera abordé par la suite.

Les deux langues diffèrent en ce qui concerne la construction des composés. En allemand, les composés peuvent être formés sur la base de deux membres, mais également de trois, quatre ou plus. En français, la structure des composés repose sur deux constituants.

Par ailleurs, l'allemand utilise des éléments de liaison entre les différents membres du composé, qui peuvent être les lettres *–(e)s, -(e)n, -ens, -e* ou bien encore *–er*. Le français a également recours à divers moyens de liaison entre les deux membres du composé, qui sont la soudure entre les deux mots, la présence d'un espace entre les deux mots, le trait d'union ou une préposition qui relie les deux termes. Si le choix de la lettre de liaison en allemand pose des problèmes, il en est de même en français en ce qui concerne le choix entre ces différentes possibilités d'éléments de liaison. Si la soudure et le trait d'union marquent l'aspect lexicalisé du composé, la préposition provoque plus de difficulté. Il faut pouvoir choisir la préposition idéale pour chaque composé.

Enfin, le français a deux autres difficultés à gérer, à savoir l'ordre des mots et le pluriel. La disposition des mots peut modifier le sens d'un composé, aussi il faut respecter la place de l'adjectif et du nom qu'il accompagne. En allemand, l'adjectif est soit concaténé au nom, soit pré-posé au nom : dans les deux cas, le sens du composé reste le même. Quant au pluriel, il est source d'ambiguïtés en français : si la réforme de l'orthographe résout les problèmes générés par les tendances sémantiques (certains termes ayant une valeur sémantique particulière sont des invariables au singulier ou au pluriel, or les nouvelles règles annulent ce principe), il faut toutefois être capable de dissocier les catégories grammaticales, car elles ne reçoivent pas les mêmes marques du pluriel : notamment seuls les substantifs et les adjectifs ont une marque d'accord. En allemand, le pluriel ne pose pas de problème, car les mots

composés reçoivent la marque du pluriel du dernier terme concaténé, de la même manière que si ce terme est un mot simple.

3.3. Classification du mot composé en français et en allemand

Le problème de la classification des mots composés a été soulevé par Otto Jespersen dans son œuvre *Analytic Syntax*, qui a tenté de classifier les mots composés de différentes langues (cf. Bouvier 1999, 6). Cette classification est basée non pas sur la syntaxe, mais sur la qualité du matériel lexical. Or, les mots peuvent être de plusieurs catégories et à l'intérieur même d'une catégorie, il y a des mots formés par des règles de formation de mots différents, notamment *chemin de fer* et *fer à repasser* appartiennent à la même catégorie des substantifs formés à partir d'une préposition, mais dans un cas on a *N* + *Prép* + *N* et dans l'autre cas *N* + *Prép* + *V*.

On peut proposer la classification suivante :

> Regrouper les mots composés selon leur structure interne et leur structure de surface. Par exemple, les termes *chemin de fer* et *tasse à thé* relèvent de la structure interne *N* + *complément* et de la structure de surface *N* + *Prép* + *N*.

Ce type de classification convient à la traduction automatique, car il faut une analyse non seulement de la structure interne des expressions, mais également de la structure de surface pour pouvoir établir des règles précises de correspondance entre les langues.

Les mots composés allemands peuvent être des noms d'objets (*Autoradio*), d'événements (*Hundertjahreskrieg*) ou de lieux (*Waffenplatz*). Ils peuvent également relever d'autres catégories grammaticales, par exemple celle de l'adjectif (*mordsmässig*) ou encore celle du verbe (*hilferufen*). La complexité des mots composés provient de cette diversité de catégories, et donc de constructions différentes, et de la difficulté à effectuer des correspondances littérales entre le français et l'allemand due à des constructions fondamentalement divergentes.

Les systèmes de traduction REVERSO et SYSTRAN se heurtent notamment à ces difficultés : utilisant tous deux une méthode basée sur le lexique (ils disposent d'un dictionnaire général et de dictionnaires particuliers), bien que leur support computationnel soit différent, ils sont confrontés aux mêmes problèmes, à savoir la décomposition et l'identification des catégories grammaticales des mots composés allemands et la correspondance littérale française.

4. Méthode de traduction

La méthode de traduction utilisée se base sur la comparaison entre les deux systèmes. Les points analysés sont le lexique et la grammaire. En ce qui concerne le lexique, nous avons étudié sa densité, ses capacités de correspon-

dance entre les mots composés et ses possibilités de composition et de décomposition de nouveaux termes. La densité du lexique réfère au nombre de mots codés, bien traduits, mal traduits ou inconnus. Les capacités de correspondance renvoient au nombre de mots codés dans les deux langues et les capacités de composition à la production de nouveaux composés à partir du codage interne. Quant à la grammaire, nous avons analysé les règles de formation des composés et les règles de correspondance pour la traduction des composés. Les règles de formation renvoient à la question de savoir comment les mots composés sont formés en français et en allemand, c'est-à-dire quelle est leur structure de surface. Les règles de correspondance mettent en relation d'équivalence les règles de formation des deux langues.

Les exemples suivants illustrent la grammaire des deux systèmes :

(1) Règle de formation en français : N2 de N1

(2) Règle de formation en allemand : N1 + N2

(3) Règle de correspondance : N1 + N2 = N2 de N1

Les règles de formation de chacune des deux langues sont mises en correspondance et sont inscrites dans les systèmes.

La comparaison représente une mise en correspondance des résultats obtenus par le système REVERSO avec ceux du système SYSTRAN. On peut noter que REVERSO EXPERT est un traducteur professionnel, alors que SYSTRAN est un traducteur disponible sur Internet et par conséquent moins performant que REVERSO. En effet, il existe des traducteurs professionnels SYSTRAN, mais ceux-ci ne traitent pas la paire de langues allemand-français. Toutefois, même si les deux systèmes n'appartiennent pas à la même catégorie de traduction, l'impossibilité d'obtenir un traducteur SYSTRAN plus performant autorise cette comparaison.

Etant donné les résultats obtenus dans les analyses des systèmes REVERSO et SYSTRAN[4], le lexique de REVERSO est plus riche et plus complet que celui de SYSTRAN. Il en est de même pour la grammaire : REVERSO contient davantage de règles de correspondance que SYSTRAN et celles-ci sont souvent plus précises que celles de SYSTRAN.

5. Les problèmes rencontrés et leurs solutions

Etant donné que REVERSO et SYSTRAN exploitent la même méthode, ils sont confrontés aux mêmes difficultés, tant au niveau du lexique que de la grammaire. Différentes solutions peuvent être apportées en fonction des deux

[4] Pour une description détaillée des analyses de chacun des systèmes, se référer à Bovier (2002).

systèmes. Aucune liste exhaustive ne sera présentée, seules les difficultés prédominantes seront expliquées[5].

5.1. Les mots inconnus produits par REVERSO et SYSTRAN

Les mots inconnus proviennent de la difficulté des systèmes à les décomposer. Lorsque l'un des composants est inconnu, la traduction échoue, car le système ne trouve pas le terme dans le lexique. Le cas des abréviations illustre ce phénomène pour les deux systèmes.

Lorsque les termes sont tous reconnus par le système, la traduction ne s'effectue toutefois pas toujours, comme les structures suivantes l'exemplifient :

(4) V1 + V2 = V2 + V1

REVERSO et SYSTRAN produisent des mots inconnus, car tous deux ne possèdent pas de règle de correspondance pour traiter ces expressions. On peut noter que cette règle est la seule qui fait défaut à REVERSO. En principe, l'insertion de cette règle de correspondance devrait éviter la lecture en tant que mot inconnu, mais il est possible qu'un certain nombre de ces composés doivent être codés dans le lexique à cause d'une modification de sens.

(5) Adj + V = Adj-V

REVERSO ne produit que des mots inconnus, notamment dans le cas du verbe *hochkämmen* (*retrousser*). SYSTRAN ne parvient pas davantage à traduire ces expressions, sauf dans le cas des adjectifs *hoch* et *gut*, où seuls les adjectifs sont traduits. Cette différence ne démontre pas l'efficacité de SYSTRAN par rapport à REVERSO, puisque la traduction demeure incorrecte, mais elle met en évidence que le problème ne vient pas du découpage, car l'adjectif est identifié. L'impossibilité de SYSTRAN de pouvoir traduire ces expressions bien décomposées semble provenir de l'absence de règle de correspondance. La difficulté de traduction des deux systèmes souligne des aspects différents de la complexité de la traduction automatique : REVERSO n'arrive pas à décomposer les expressions, alors que SYSTRAN manque de règles de correspondance.

Le système REVERSO n'arrive pas à traiter cette structure. Une explication possible serait la confusion entre le nom et l'adjectif, notamment *hoch (Das Hoch), schwarz (Schwarze)* et *schlecht (Schlechte)*, ou la confusion entre l'adverbe et l'adjectif, comme *klar*. Le système, ne parvenant pas à établir la catégorie grammaticale du terme, ne saurait dès lors pas quelle règle de correspondance utiliser. Cependant, si l'on inscrit les mêmes verbes cités ci-dessus au participe passé, le système les traduit correctement. Par exemple,

[5] Pour une description complète des problèmes relevés au niveau du lexique et de la grammaire, se référer à Bovier (2002).

hochgespielt se traduit par *très joué* et *klargedacht* par *pensé clairement*. Le fait que le verbe soit à l'infinitif empêche la décomposition. Ce phénomène reste inexplicable. La seule solution est de coder les composés dans le lexique, puisque les adjectifs sont déjà codés dans le dictionnaire.

Pour SYSTRAN, il ne s'agit pas d'un problème de découpage, puisque certains adjectifs sont traduits. On ne peut pas davantage définir ces adjectifs en tant que termes les plus courants, puisque *arm* et *reich* sont des adjectifs usuels et ne reçoivent pas de traduction. Il n'est également pas possible qu'ils aient un codage spécial, puisque, comme l'illustre la structure *Adj1 + Adj2*, dans le cas des adjectifs *hell* et *hoch*, le système les traduit de manière irrégulière, notamment *hochgross* reçoit la traduction *grand*, alors que le composé *hochschön* est inconnu. Ces différences ne s'expliquent pas, elles mettent en évidence les défauts de SYSTRAN.

La première solution consiste à inscrire les règles de correspondance pour chacune des structures : ce procédé devrait permettre de combiner les deux composants ensemble, notamment l'adjectif avec le nom, l'adjectif avec le verbe ou les deux adjectifs ensemble. La seconde solution concerne le codage des adjectifs en interne : par exemple, *hoch* se traduit par *très*. A l'aide du codage interne, les adjectifs qui doivent recevoir un sens différent dans un composé proposent des traductions correctes. Dans le cas des adjectifs exigeant une préposition, comme *arm* qui se traduit par *pauvre en*, il convient d'inscrire la préposition s'attachant à l'adjectif dans le lexique. Ces trois solutions devraient permettre d'éliminer la proportion de composés inconnus.

(6) Adv + N = N + Adj ou Préfixe + N

Si l'adverbe n'est pas codé en interne, la traduction échoue. Les deux systèmes produisent des mots inconnus. Il en est de même pour les structures *Adv + V* et *Prép + N*, où l'adverbe et la préposition doivent avoir un codage interne. Une différence réside toutefois entre REVERSO et SYSTRAN au niveau de la structure *Prép + Adj* : SYSTRAN ne parvient pas à traduire les expressions adjectivales comprenant la préposition *anti-*, alors qu'il produit des traductions correctes dans le cas d'un composé nominal. REVERSO n'est pas confronté à cette difficulté.

Pour résoudre ce problème, le codage interne semble à nouveau être un moyen efficace. Cependant, dans le cas de SYSTRAN, la structure *Adv + V* produit des composés inconnus, malgré le codage interne des adverbes. Par exemple, dans la structure *Adv + N*, l'adverbe *wieder* est codé par le préfixe *re-* dans un composé : *Wiederdruck* est traduit par *repression*, au lieu de *réimpression*, mais le découpage s'effectue correctement. En revanche, l'expression *wiederentwickeln (redévelopper)* est inconnue, bien que le verbe *entwickeln* soit connu du système. Il en est de même pour la structure *Adv + Adj*. On peut donc en déduire que les règles de correspondances pour ces

deux types de structures ne sont pas inscrites dans le système. Il suffit de les ajouter, afin que le système génère des traductions correctes.

Cependant, le cas de la préposition *anti-* pose un problème : étant donné qu'elle produit des composés connus lors de la structure *Prép + N* et des composés inconnus lors de la structure *Prép + Adj* chez SYSTRAN, on peut mettre en doute le postulat du codage interne. Si le codage ne s'applique qu'aux prépositions concaténées à un nom, il faut élargir le codage aux adjectifs. S'il existe des règles propres aux prépositions, telles que *anti + N = anti + N* ou *inter + N = inter + N*, il faut en inscrire pour chaque préposition et pour les adjectifs également. Dans tous les cas, étant donné qu'il existe deux prépositions qui offrent la possibilité de décomposition, il doit être possible d'augmenter ce résultat à l'aide d'une des deux solutions proposées, à savoir soit le codage de la préposition en interne, soit des règles permettant de traiter chaque préposition séparément.

Les deux systèmes se heurtent aux mêmes problèmes, même si SYSTRAN semble avoir un degré de décomposition des composés supérieur. Comme nous l'avons vu, sa capacité de décomposition des mots n'améliore pas ses résultats, car les composés, formés d'un verbe et d'un autre composant, modifient souvent leur sens et ne correspondent plus à la traduction littérale du mot. Il faut les coder dans le lexique. Les problèmes relevés mettent en évidence les difficultés que la traduction automatique doit parvenir à résoudre.

5.2. L'absence du choix de la préposition dans les mots composés traduits par REVERSO et SYSTRAN

Un autre problème rencontré dans le cadre des systèmes REVERSO et SYSTRAN renvoie au choix des prépositions dans la composition des composés en français. Tous deux utilisent une unique préposition, à savoir *de*. C'est pourquoi, un grand nombre de composés sont incorrects, tels que *Werkzeugkasten (boîte à outils)* ou *Nähkasten (boîte à ouvrage)*.

En ce qui concerne les difficultés engendrées par l'absence du choix des prépositions dans les règles de correspondances, on peut apporter trois types de réponses différentes : insérer plus de données dans la rection, introduire de la sémantique dans le programme ou coder tous les termes dans le lexique. Il n'est en effet pas possible d'effectuer de correspondance sur la structure même des composés.

La première solution est relativement peu lourde pour le système, car il suffit de modifier la fonction de rection. Si l'on inscrit pour chaque nom allemand entré dans le lexique la préposition à utiliser en français, même lorsque aucune préposition n'est employée en allemand, le système n'aura plus qu'à appliquer la bonne préposition dans le composé. Par exemple, si le

nom *Trip* se voit associer la préposition *à* dans le cas d'une ville et des pays masculins (*au Japon*), la préposition *en* dans le cas d'un pays féminin (*en Chine*) et la préposition *de* dans tous les autres cas (*voyage d'affaire*), la traduction d'*Honolulu-Trip* en *voyage à Honolulu* s'effectuera correctement.

Il en est de même pour *Anspruch*, qui demande la préposition *à* suivie d'un nom en français, *le droit au pouvoir (Machtanspruch*), et la préposition *de* suivie d'un verbe, *le droit d'acheter (Anspruch zu kaufen)*. Les composés *bankintern (interne à la banque)* et *strukturschwach (déficient en structure)* trouvent également leur résolution dans le codage des prépositions dans la rection, notamment *intern* demande *à* et *schwach en.* La même solution se retrouve dans l'emploi de la préposition *envers* qui accompagne le nom *Feindlichkeit (hostilité)* et dans l'emploi de la préposition *sur* dans le cas de *Debatte (débat)*.

Cependant, le terme *Feindlichkeit* peut aussi utiliser la préposition *de*, notamment dans *hostilité des étudiants*. Dans un cas, on décrit l'objet envers lequel porte l'hostilité, dans l'autre cas, on dépeint l'agent de l'hostilité. Il est donc difficile sans usage de la sémantique de définir quelle est la préposition à employer dans ce cas. Le même problème se retrouve avec *Debatte*, qui peut référer au sujet du débat (*débat sur la pollution*) ou aux agents du débat (*débat des hommes politiques*). *Nähkasten (boîte à ouvrage)* et *Werkzeugkasten (boîte à outils)* utilisent le même nom, *Kasten*, qui a besoin de la préposition *à*. Cependant, à nouveau, on peut utiliser ce terme avec une autre préposition, comme dans *boîte de chocolat*. Dans un cas, la préposition *à* réfère au rôle du récipient et dans l'autre cas, la préposition *de* renvoie au contenu.

La sémantique permet de gérer ces différences, mais il semble difficile à la fonction de rection de les contrôler : il faudrait développer davantage le rapport entre les quelques choix sémantiques disponibles pour les noms dans le programme (*nom propre, monnaie, pays, ville, nom abstrait, substance et autre)* et les prépositions à inscrire. Il faudrait aussi affiner les choix sémantiques, en ajoutant d'autres catégories, dont *agent, patient, cause, contenu, récipient, bénéficiaire,* etc.

Pour ces raisons, il semble plus facile d'insérer directement des données sémantiques dans le programme. Cette seconde solution est plus lourde pour le système, car le programme doit alors non seulement regarder si le terme est dans le lexique, mais également contrôler sa valeur sémantique. Il s'agit de créer un vaste réseau sémantique, qui contient un maximum de classes sémantiques différentes, regroupées en fonction des prépositions. Par exemple, le terme *voyage* demande les prépositions *à* ou *en*, le terme *débat* la préposition *sur* ou *de*, etc. Il faut construire des réseaux sémantiques performants, qui contiennent tous ces termes, notamment introduire la classe sémantique *voyage* et lui associer les prépositions en fonction des règles qui s'y rappor-

tent (*à* = ville et pays masculin ; *en* = pays féminin), de même que la classe sémantique *discussion* qui réfère à la préposition *sur* dans le cas du patient et à la préposition *de* dans le cas de l'agent. Il faut donc introduire la grammaire des cas, qui permet de différencier les fonctions des prépositions, selon une distinction entre l'agent, le patient, le bénéficiaire, etc.

Une autre différenciation doit s'opérer entre les rôles de la fonction (préposition *à*) et du contenu (préposition *de*). Par exemple, la préposition *de* s'utilise avec toute boisson ou nourriture (*verre de lait, verre de bière, assiette de soupe,* etc.) et avec tout récipient ou ustensile de cuisine (*bidon de lait, tasse de lait, bol de lait, assiette de lait, plaque de chocolat, rouleau de réglisse,* etc.). Une exception existe cependant, l'ustensile *couteau* qui ne s'emploie qu'avec la préposition *à (couteau à viande*), car il réfère obligatoirement à la fonction. Il faut donc ajouter une catégorie sémantique *ustensile de cuisine* et la relier à la classe sémantique *boire* et *manger* (*avaler, ingurgiter, dévorer*, etc.). Il suffit alors de catégoriser en tenant compte soit de la fonction, soit du contenu. Par exemple, *rouleau* peut créer *rouleau à pâtisserie* (la fonction est mise en évidence) et *rouleau de réglisse* (le contenu est souligné).

S'il est possible de construire un réseau sémantique, en mettant en parallèle les prépositions, il faut que le système soit capable de reconnaître un agent d'un patient, une fonction d'un contenu. Pour ce faire, le système doit posséder des connaissances du monde, ce qui semble impossible. Il n'est en effet pas possible de créer des classes sémantiques *fonction* ou *contenu*, qui permettraient de différencier les emplois, puisque chaque terme appartient aux deux classes. Le système peut proposer à l'utilisateur les deux solutions, à savoir un composé avec *de* et un composé avec *à*, et l'utilisateur choisit en fonction du contexte. Ce procédé implique que l'utilisateur doit connaître la langue cible. D'autre part, REVERSO et SYSTRAN ont opté pour une autre politique, à savoir ne pas générer un grand nombre de choix pour l'utilisateur, mais bien offrir une traduction complète. Le réseau sémantique semble donc un moyen lourd et pas totalement efficace pour gérer les prépositions.

La dernière solution possible reste à coder tous les termes erronés dans le lexique. Actuellement, c'est la solution qui a été choisie, en notant tous ces nouveaux termes dans le dictionnaire général. Le nombre de mots n'étant pas trop exhaustif, cette solution demeure la plus économique. Cela évite de surcharger le système avec des données sémantiques et permet de ne pas modifier la fonction de rection, en la rendant difficile, voir inutilisable, de par une trop grande complexité.

6. Les problèmes non résolus

Les difficultés rencontrées par les deux systèmes et présentées dans la section précédente n'ont pas reçu de solution satisfaisante. Les problèmes liés à la décomposition des composés allemands et à la gestion des prépositions en français ont été contournés en utilisant le codage dans le dictionnaire, mais aucune explication n'a été fournie, permettant d'éliminer entièrement ces mauvaises productions. Un autre problème reste entier : l'insertion d'un adjectif dans un mot composé.

Les expressions *des italienischen TV-Publikums* et *des italienischen Fernsehpublikums (du public italien de la télé)* sont traduites par *du public de télé italien*. Les mots composés *TV-Publikum* et *Fernsehpublikum (public de télé)* n'acceptent pas l'insertion de l'adjectif *italienisch (italien)* entre les termes *TV* et *Publikum* ou *Fernseh* et *Publikum*, car les expressions sont codées dans le lexique et ne peuvent pas être modifiées. Étant donné qu'il n'est pas possible de coder les syntagmes eux-mêmes, de par une lecture trop dépendante du contexte, la seule solution réside à insérer une règle stipulant que l'accord de l'adjectif dans un composé de deux noms doit s'effectuer avec le premier nom allemand, même s'il porte un autre accord. Cela a comme conséquence que le sens du texte est modifié. Dans l'exemple du *public italien de la télé*, le sens n'est que très peu influencé, mais dans d'autres cas, il peut en être autrement. Le sens peut même devenir incompréhensible. Ainsi, cette solution ne semble pas très productive, au contraire, elle pourrait nuire à la traduction. C'est pourquoi, il convient d'être très attentif en codant des expressions dans le lexique. Ce cas problématique ne reçoit donc pas de réponse satisfaisante. Ce type de problème illustre les limites de la traduction automatique.

7. Conclusion

Les systèmes de traduction REVERSO et SYSTRAN font intervenir la même méthode pour la traduction des termes composés : le codage interne, qui permet d'améliorer la qualité de la traduction, les règles de formation et de correspondance, qui mettent en relation la syntaxe d'une paire de langues, et le lexique, qui contient le vocabulaire. Des dictionnaires spécialisés et des traitements internes, qui gèrent les cas particuliers, enrichissent également les deux systèmes.

De par l'utilisation de la même méthode, les systèmes se heurtent aux mêmes problèmes : production de mots inconnus ou mal traduits, impossibilité de choisir la bonne préposition ou d'insérer un déterminant dans un mot composé, ainsi que mauvaise disposition de l'adjectif dans un composé codé dans le lexique. Ces problèmes ont reçu une solution dans la section 5, notamment ajout de règles de correspondance, diminuant la proportion de mots

inconnus, ajout de codages internes résolvant certaines maladresses, insertion de sémantique dans le programme pour gérer les prépositions et enfin ajout de règles s'appliquant aux prépositions et aux déterminants. Toutes ces difficultés mettent en évidence d'une part les améliorations possibles de la traduction automatique et d'autre part ses limites.

REVERSO et SYSTRAN sont toutefois confrontés à des problèmes différents, malgré leur stratégie similaire : SYSTRAN, ayant développé moins de règles de correspondance et une couverture lexicale plus réduite, produit davantage de mauvaises traductions. Ainsi, REVERSO opte pour l'encodage d'un maximum de termes, alors que SYSTRAN se contente d'un lexique moins riche. REVERSO propose une grande quantité de règles de correspondance, afin de traiter le plus d'expressions possibles, alors que SYSTRAN restreint le nombre de règles de correspondance. Pour toutes ces raisons, SYSTRAN est un système de traduction moins performant que REVERSO.

Un point négatif peut toutefois être relevé en ce qui concerne REVERSO : la décomposition des termes. Plusieurs composés demeurent inconnus, alors que les composants des expressions sont reconnus par le système. REVERSO ne parvient pas à décomposer certains mots composés en fonction de leurs membres. SYSTRAN n'est pas confronté à cette difficulté : ses traductions sont souvent maladroites, voire incorrectes, mais il découpe facilement l'expression composée. Les causes, qui sont à la base de cette incapacité de décomposition de certaines structures, particulièrement lorsque l'un des membres est un verbe, un adverbe ou une préposition, demeurent inconnues. Dans tous les cas, le système nécessite le codage interne des composants problématiques, tels que les verbes, les adverbes et les prépositions, pour fournir une traduction correcte.

Les mots composés allemands, qui peuvent concaténer deux termes, mais aussi trois et plus, illustrent bien les problèmes liés à la décomposition des termes. Les membres peuvent être de diverses catégories grammaticales, notamment un nom, un adjectif ou un verbe, etc. Étant concaténés les uns aux autres, le système doit les découper pour pouvoir choisir la bonne règle de correspondance. En effet, la traduction française est basée sur les règles de correspondance. L'insertion des catégories grammaticales, de même que l'utilisation d'une préposition, d'un espace ou d'un trait d'union formant un mot composé, sont gérées par les règles de formation françaises. Si les règles de formation et de correspondance sont indispensables à la traduction, la possibilité de décomposition est elle aussi fondamentale.

La traduction automatique est un domaine de recherche actuelle, c'est pourquoi les entreprises commercialisent leurs produits, comme dans le cas de REVERSO et SYSTRAN, bien qu'il reste encore de nombreuses améliorations à effectuer pour obtenir un niveau de traduction professionnel. Ce-

pendant, on peut douter de parvenir un jour à créer un système aussi performant qu'un traducteur humain, de par la complexité du langage humain. Les systèmes de traduction ne sont d'ailleurs pas conçus dans ce but : ils permettent de comprendre un texte écrit dans une langue étrangère et produisent des traductions plus ou moins correctes de textes commerciaux ou autres.

Bibliographie

BOUVIER Y.-F. (1999), *Comment reconnaître et classifier les mots composés*, mémoire de licence, Département de linguistique, Université de Genève.

BOVIER A. (2002), *Comparaison des systèmes de traduction automatique REVERSO et SYSTRAN : les mots composés de l'allemand en français*, mémoire de DEA, Département de linguistique, Université de Genève.

CATACH N. (1981), *Orthographe et Lexicographie : les mots composés*, Paris, Fernand Nathan.

HABERT B. et al. (1993), « Traitements automatiques de la composition nominale », *T.A.L.* 34 : 2, Paris, ATALA.

HENZEN W. (1965), *Deutsche Wortbildung*, Tübingen, Max Niemeyer Verlag.

HUTCHINS J. & SOMERS H. (1992), *An Introduction to Machine Translation*, Cambridge, Academic Press Limited.

NIRENBURG S. (1987), *Machine Translation : Theoretical and Methodological Issues*, Cambridge, Cambridge University Press.

ORTNER L. et al. (1991), *Deutsche Wortbildung. Vierter Hauptteil : Substantivkomposita*, Berlin, Walter de Gruyter et Co.

PÜMPER-MADER M. et al. (1992), *Deutsche Wortbildung. Fünfter Hauptteil : Adjektivkomposita und Partizipialbildungen*, Berlin, Walter de Gruyter et Co.

REICHLER-BEGUELIN M-J. et al. (1996), *Les rectifications de l'orthographe du français*, Neuchâtel, Institut Romand de Recherches et de Documentation Pédagogiques.

WOLF B. (1990), *Nominal Kompositionen im Deutschen und Französischen,* Münster, Kleinheinrich Verlag.