Les erreurs dans la traduction automatique du genre dans les couples français-anglais et anglais-français : typologie, causes linguistiques et solutions

Antonia Cristinoi-Bursuc

LES ERREURS DANS LA TRADUCTION AUTOMATIQUE DU GENRE DANS LES COUPLES FRANÇAIS-ANGLAIS ET ANGLAIS-FRANÇAIS : TYPOLOGIE, CAUSES LINGUISTIQUES ET SOLUTIONS

Antonia Cristinoi-Bursuc

Publications linguistiques | « Revue française de linguistique appliquée »

2009/1 Vol. XIV | pages 93 à 108

ISSN 1386-1204
DOI 10.3917/rfla.141.0093

Rev. franç. de linguistique appliquée, 2009, XIV-1 (93-108)

Les erreurs dans la traduction automatique du genre dans les couples français-anglais et anglais-français : typologie, causes linguistiques et solutions

Antonia Cristinoi-Bursuc
Université d'Orléans

Résumé : *En s'appuyant sur les notions de* classes de comportement*, de* marquage *et d'*indices morphosyntaxiques*, cet article montre d'une part que l'on peut prédire en amont, au niveau lexical et pour chaque unité linguistique, le type de problème qui peut surgir lors de la traduction (automatique ou non) du genre dans les couples français-anglais/anglais-français, et d'autre part qu'il est possible de trouver des solutions systématiques et automatisables pour chaque type de problème. Le modèle proposé pour ces deux langues peut être étendu à d'autres langues ou couples de langues mais aussi à d'autres catégories linguistiques, et peut en cela contribuer à l'amélioration du fonctionnement des traducteurs automatiques.*

Abstract: *By means of the notions of* behavioural classes, marking *and* morphosyntactic markers *this paper shows that all the translation (or Machine Translation) problems that arise when translating gender from French into English or vice-versa can be predicted a priori at a lexical level, for all the linguistic items concerned. It also proves that systematic solutions to these problems can be found and implemented. The approach defended here for French and English can be applied to other languages or language pairs, and to other linguistic categories, and could thus contribute to the improvement of Machine Translation systems.*

Introduction

Le monde actuel est un monde du développement technologique, de la multiplication des domaines scientifiques et techniques, un monde « mondialisé », en mouvement permanent, dans lequel la transmission de l'information et la rapidité de sa diffusion sont vitales. Dans un contexte où les moyens de diffusion des connaissances évoluent aussi rapidement que les connaissances elles-mêmes, la traduction joue un rôle primordial, permettant à tous l'accès au savoir. De plus en plus, elle est considérée comme un produit qui doit être évalué, en dehors de la qualité du texte cible, c'est-à-dire de la traduction proprement dite, en termes de coût, de présentation, et surtout de rapidité du résultat, rapidité garantie actuellement

par les outils de traduction assistée par ordinateur (TAO) et par la traduction automatique (TA)[1]. Outre son caractère quasi-instantané, la TA permet l'accès à la traduction à tous et en toutes circonstances, mais présente en revanche le désavantage de ne pas être toujours fiable.

Les évolutions technologiques en matière de traduction, TA et TAO, conduisent à réexaminer le rôle que peut jouer la linguistique dans le développement des outils de traduction, à étudier l'effet de l'avènement de la traductique sur la traduction, ainsi que les attentes en moyens linguistiques qu'elle impose, et aussi à explorer la façon dont la traduction, tout particulièrement la TA, influe à son tour sur la linguistique.

La présente étude, quoique limitée à la question de la traduction automatique du genre dans les couples français-anglais/anglais-français, vise à proposer une nouvelle manière (inspirée par la linguistique contrastive et la morphologie) d'envisager le traitement des erreurs de traduction produites par les logiciels de TA. Le point de départ de cette démarche est le constat que des erreurs récurrentes dans les traductions réalisées par ordinateur seraient facilement évitées si l'on prenait en compte *a priori* les différences structurelles entre langues. La solution serait donc d'anticiper sur les problèmes au lieu de les traiter un par un après coup, comme c'est largement le cas dans les démarches actuelles, essentiellement statistiques. Car si celles-ci effectuent un inventaire *a posteriori* des erreurs de traduction les plus fréquentes (de façon à parvenir rapidement à un taux de résolution des problèmes qui assure l'intérêt commercial d'un programme), afin de les corriger une par une, cette réparation ponctuelle ne prend pas en compte les causes profondes du problème (le plus souvent des différences structurelles entre les langues). Il serait donc souhaitable, si l'on veut véritablement améliorer les logiciels de TA, de repérer plutôt *a priori* les problèmes que soulèvent ces différences interlinguistiques et d'y apporter une solution.

Je montrerai ici qu'en étudiant en détail la TA du genre dans un nombre restreint de langues, même typologiquement proches, il est possible de dégager des modèles de traitement généralisables et applicables à un grand nombre de langues (voir Cristinoi 2005). La perspective adoptée se situe sur un plan théorique et non dans une optique directe d'implémentation. Il s'agit de valider une démarche, et non de concevoir un logiciel de TA, bien que les exigences et les contraintes de la TA soient prises en compte. Cette perspective permet ainsi de faire apparaître des difficultés de traduction qui passeraient inaperçues pour un traducteur humain, étant résolues, dans ce cas, de façon quasi automatique.

L'objectif immédiat de cette étude est avant tout d'identifier de manière exhaustive les erreurs/problèmes[2] de traduction qui ont pour source le marquage différent du genre en français et en anglais et d'en mettre au jour les causes, afin d'envisager ensuite un traitement automatisable, donc applicable à tous les cas

[1] Voir, par exemple, Boitet (1999), Bowker (2002), Hutchins (2001, 2006), Quah (2006).

[2] Les notions d'erreur et de problème constituent les deux facettes d'une même réalité dans la mesure où une erreur de traduction n'est rien d'autre que la matérialisation d'un problème existant en amont, raison pour laquelle je les considère comme indissociables.

similaires pour ce couple de langues, mais aussi généralisable à d'autres langues. Cela revient plus concrètement à l'élaboration d'un cahier des charges implémentable pour la traduction du genre dans les langues concernées.

1. La question du corpus

Lorsque l'on se donne comme objectif d'étudier les erreurs commises par un traducteur automatique, deux types de démarche sont envisageables : l'une empirique, consistant à recueillir un corpus de textes, dans lequel les occurrences de la question traitée sont identifiées et analysées sur des bases essentiellement statistiques, pour servir d'appui à un raisonnement ; l'autre théorique, supposant une analyse préalable de la question envisagée, suivie de l'élaboration d'hypothèses qui seront ensuite testées.

Dans le premier cas, les problèmes de traduction seront traités *a posteriori,* en partant des erreurs pour trouver leur source et ensuite y remédier si possible, ce qui soulève la question de la pertinence d'un échantillon donné pour l'identification de *toutes* les questions que pose par exemple la traduction du genre, dans la mesure où celle-ci concerne souvent des lexèmes particuliers, comme nous le verrons plus loin. Serait-il raisonnable, dans ce cas, de songer à élaborer un corpus qui devrait pratiquement inclure tous les mots d'une langue ?

Dans le second cas, sans tomber dans les travers d'une démarche statistique (qui a sans aucun doute ses mérites mais qui sera volontairement écartée ici), la question du genre et de sa traduction en français et en anglais sera traitée d'abord hors corpus, afin de vérifier dans quelle mesure les différences structurelles entre ces deux langues ont un impact sur la traduction en général et sur la TA en particulier, et les hypothèses avancées seront corroborées sur corpus. Cela s'avère nettement plus économique qu'une démarche de type correction d'erreurs, surtout si les résultats sont généralisables à d'autres langues.

L'élaboration du corpus destiné au type d'étude mené ici ne saurait donc précéder une réflexion sur la question du genre dans les deux langues, dans la mesure où un grand nombre d'erreurs dans la TA du genre, qu'il est en théorie possible de prévoir (et qui sont confirmées par des tests sur corpus), n'apparaissent pas dans des corpus établis *a priori.* Par conséquent, au lieu de repérer les problèmes de traduction sur corpus et de les analyser ensuite, ceux-ci sont déterminés d'une façon axiomatique et exemplifiés soit par des phrases toutes faites illustrant un cas de figure, accompagnées ou non d'une traduction ou d'une traduction automatique, soit par des phrases tirées de corpus, tel le *British National Corpus* (*BNC*), qui illustrent les questions traitées, accompagnées d'une traduction automatique.

La traduction automatique des phrases n'est pas utilisée dans le but de montrer l'incapacité des traducteurs automatiques mais pour vérifier la façon dont le programme gère les problèmes identifiés, et pour repérer les erreurs qu'il peut commettre. L'utilisation de plusieurs traducteurs automatiques pour rendre la même phrase permet de mettre en évidence les différences de traitement et de déterminer si les erreurs qui s'ensuivent sont du même type. Le choix des traducteurs automatiques a été fait sur la base de leurs performances (SYSTRAN et REVERSO

sont considérés comme les meilleurs traducteurs automatiques sur les couples anglais-français, français-anglais) et de leur accessibilité (INTERTRAN est accessible en ligne pour un grand nombre de langues)[3].

Les phrases retenues illustrent les trois types de situations dans lesquels il est probable de rencontrer des problèmes de traduction du genre, à savoir la traduction des noms nus, la traduction des déterminants, et la traduction des expressions anaphoriques. Ces situations sont strictement liées aux trois fonctions que peut détenir le genre à travers les langues (Cristinoi 2007, 106) : la contribution à l'identification/caractérisation du nom (fonction référentielle/sémantique), la coréférence intraphrastique (par le phénomène d'accord au sein de la phrase) et la coréférence interphrastique (par l'accord anaphorique).

2. Analyse théorique

2.1. Genre et comportement des unités lexicales

Au terme d'un long questionnement sur le genre et la classification nominale à travers les langues (Cristinoi 2007), il est apparu que toute étude linguistique du genre doit répondre aux questions suivantes :

- quel est le nombre de genres présents dans la langue en question ?
- quels sont les éléments pour lesquels le genre constitue un trait inhérent ?
- quels sont les éléments grammaticaux sur lesquels est marqué le genre ?
- quels sont les mécanismes mis en œuvre pour le marquage de cette catégorie ?
- quels sous-types de comportements des catégories d'unités lexicales peuvent être dégagés en fonction des deux critères, genre et marquage du genre ?

Reprendre en détail les réponses à ces questions, même pour deux langues seulement, est impossible dans le cadre limité de cet article, où seules seront reprises les conclusions d'une étude précédente. Une précision s'impose cependant : les questions posées ci-dessus supposent résolu le problème de l'existence ou non du genre dans les langues traitées. Or, si pour le français il n'y a aucun doute à ce sujet, pour l'anglais le débat reste encore largement ouvert. La position adoptée ici est celle de l'existence de la catégorie *genre* en anglais, sur la base de la définition des genres comme des classes d'accord (accord au sens large, s'appliquant aussi à l'anaphore pronominale) proposée par Hockett (1958, 231) et reprise par Corbett (1991, 4)[4].

Les classifications très générales proposées par les grammaires traditionnelles, destinées à l'apprentissage, n'étant pas suffisantes pour la traduction ou pour la TA, une nouvelle classification est proposée, prenant en compte non pas le nombre de genres qui existent dans les langues traitées (parce que les dissymétries ne se situent pas seulement à ce niveau), mais plutôt le comportement individuel des unités

[3] SYSTRAN : version utilisée SYSTRAN 6, 2007 ;
REVERSO : <http://www.reverso.com/index-fr.html>.
INTERTRAN : <http://www.tranexp.com:2000/Translate/result.shtml>.

[4] Pour d'autres discussions de la question du genre, voir notamment Unterbeck & *al*. (2000).

lexicales du point de vue du genre. Cette classification permet de montrer que les problèmes de traduction du genre ne surgissent pas, comme on pourrait le croire, du nombre différent de genres en français (deux genres – masculin et féminin) et en anglais (trois genres – masculin, féminin et neutre), mais du comportement différent des unités lexicales par rapport au marquage du genre et aux informations portées par cette catégorie. Ce dernier point mérite quelques précisions : afin d'étudier le comportement des unités lexicales du point de vue du genre, il convient de prendre en compte le type d'informations fourni par le genre dans les langues étudiées : d'une part des informations référentielles (le genre de certaines unités lexicales constituant un indice du sexe du référent de l'unité lexicale en question), ce que je nommerai désormais *genre référentiel* ; d'autre part des informations grammaticales (concernant l'accord intra et interphrastique), qui seront reprises sous la dénomination de *genre grammatical*.

Dans une approche *traductionnelle* du genre, il s'agit en effet de dégager parmi les noms des séries lexicales nommées *classes de comportement* (unités lexicales qui partagent le même comportement du point de vue du genre), afin de déterminer les problèmes qui peuvent surgir pour chaque classe lors de la traduction[5]. La répartition des noms au sein de ces classes s'effectue sur la base de leur comportement en termes d'accord (syntaxique et sémantique), de leur genre grammatical et de leur genre référentiel (ou de la pertinence de ce paramètre). Cette classification vient ainsi compléter celles qui sont proposées par les grammaires de référence traditionnelles[6].

Quatre classes de noms ont été ainsi répertoriées dans les langues étudiées en fonction des différences de comportement :

- *Classe 1* : noms qui possèdent un genre grammatical soit identique au genre référentiel soit totalement arbitraire et qui suivent un seul schéma d'accord, du type *boulanger* ou *table* en français, *waitress* ou *shirt* en anglais.

- *Classe 2* : noms qui possèdent un genre grammatical, correspondant à deux genres référentiels possibles et qui suivent un seul schéma d'accord. En français et en anglais, ce type de comportement s'applique uniquement aux animés (noms d'animaux dont le sexe n'est pas important) et l'accord sémantique et l'accord syntaxique se font de la même manière.

(1) *La girafe* (F) a passé sa journée à manger des feuilles d'arbres et ensuite *elle* (F) est partie chercher un point d'eau.

(2) The *swallow* (N) flew away and *it* (N) came back two hours later.

- *Classe 3* : noms qui possèdent un genre grammatical correspondant à deux genres référentiels et suivent deux schémas d'accord (l'accord syntaxique et l'accord sémantique étant différents).

(3) *Le professeur* est *entré* dans la classe. *Elle* portait une robe blanche.

Les noms de ce type sont quasi absents en anglais. Des noms comme *baby* ou *ship* pourraient avoir éventuellement ce type de comportement :

[5] Sur la classification nominale, voir les études réunies par Craig (1986) et Senft (2000).
[6] Par exemple, pour le français, Riegel & *al*. (1994), pour l'anglais, Quirk & *al*. (1985).

(4) We went to see the *baby* and we admired *it*. *Her* name was Mary.

Les noms de cette catégorie sont essentiellement humains (avec très peu d'exceptions).

- *Classe 4* : noms que l'on peut qualifier d'*ambigenres*, pour lesquels la même forme correspond à deux genres grammaticaux, qui renvoient chacun à un genre référentiel différent et suivent deux schémas d'accord (cette fois-ci accord syntaxique et accord sémantique sont identiques). On peut également considérer ces noms comme *indéterminés* du point de vue du genre, puisque, contrairement aux cas précédents, il n'y a, pour les noms nus, aucune indication du genre. Cette classe est constituée, dans les deux langues étudiées, exclusivement de noms représentant des animés humains.

(5) *Le/la journaliste* est *parti/e* au Moyen-Orient.

(6) The *teacher* came into the room and *her/his* dog followed *her/him*.

Si du point de vue du fonctionnement de leurs membres, ces classes sont uniformes, ce n'est pas le cas si l'on se place dans une perspective de traduction. La classe qui pose problème est la classe *1*, parce que les éléments qui la composent ont un comportement spécifique en traduction. Deux catégories d'éléments peuvent être dégagées au sein de cette classe : ceux pour lesquels le genre n'apporte pas d'information référentielle (*classe 1a*) ; et ceux pour lesquels le genre apporte une information référentielle, précisant le sexe, comme *boulangère* (*classe 1b*).

2.2. Problèmes de traduction du genre selon les classes de comportement

A l'aide de cette nouvelle classification, on peut dresser un relevé exhaustif des problèmes de traduction générés par l'appartenance des unités lexicales à des classes différentes dans les deux langues. A un niveau purement théorique on a un total de 25 combinaisons possibles (chacune des classes *1a*, *1b*, *2*, *3*, *4* en LS (Langue Source) pouvant en principe se combiner à chacune des classes *1a*, *1b*, *2*, *3*, *4* en LC (Langue Cible)) dont il faudrait étudier l'impact sur la traduction. En pratique, les caractéristiques sémantiques des classes rendent certaines combinaisons impossibles :

- la classe *1a* (*chaise*/*chair*) étant composée d'éléments pour lesquels la distinction de sexe n'est pas pertinente (donc plus simplement de tous les inanimés), tandis que les autres classes ne contiennent que des animés, avec des degrés variables de différenciation sexuelle, toute combinaison (incluant la classe *1a*) autre que *1a*-*1a* est impossible ;

- la classe *2* (*moineau*/*sparrow*) réunissant globalement des êtres vivants non humains et les classes *3* (*bébé*/*baby*) et *4* (*journaliste*/*journalist*) des humains, les combinaisons *2*-*3*, *3*-*2*, et *2*-*4*, *4*-*2* sont également impossibles.

Restent donc 13 possibilités réelles pour la combinaison français-anglais (dans les deux sens), qui sont à étudier une par une, afin de rechercher des régularités et d'établir une typologie des problèmes rencontrés. Trois d'entre elles seront retenues ici pour illustrer la démarche.

(i) Classe 1a – classe 1a

En apparence, cette combinaison ne pose pas de problème particulier à la traduction, puisque l'information codée par le genre est strictement grammaticale, permettant ainsi à un analyseur syntaxique de faire le lien entre les différents éléments du GN (groupe nominal) et de maintenir ce lien dans la LC, si nécessaire. Un cas problématique se rencontre néanmoins lorsque les genres de l'unité lexicale concernée ne sont pas les mêmes dans les deux langues : il s'agit de la traduction de certains pronoms anaphoriques. Si un analyseur syntaxique moyen est parfaitement capable de résoudre les problèmes d'accord lorsqu'une unité lexicale possède des genres différents dans les langues impliquées dans la traduction, gérer ce problème à un niveau macrosyntaxique (anaphore textuelle) est beaucoup plus difficile, d'autant plus que certains anaphoriques sont parfois non marqués[7] du point de vue du genre (comme *l'* dans l'exemple ci-dessous), ce qui provoque des déséquilibres en termes de la quantité d'information fournie.

(7) J'ai acheté une table. *Elle* était belle. Je *l'*ai trouvée dans un magasin.

SYSTRAN: I bought a table. *It* was beautiful. I found *it* in a store.
REVERSO: I bought a table. *She(it)* was beautiful. I found *her(it)* in a store.
INTERTRAN: J'ai buy a table. Her was beautiful. I l'ai found in a store.

La traduction effectuée par REVERSO montre bien le type de difficulté que nous avons évoqué, puisqu'elle propose les alternatives possibles, tandis qu'INTERTRAN s'avère incapable de gérer ce genre de problème.

L'exemple suivant illustre le fait que le traducteur automatique gère parfaitement le problème au niveau du GN, mais beaucoup moins bien à l'échelle d'une phrase complexe :

(8) *A* plaque on a wall is admittedly not much to look at for those with a passion for working machinery and dramatic industrial landscapes, but *this one* marks the site of the foundation of that industry which has had a greater impact on civilization than *any other*. (BNC - B0A 1632)

SYSTRAN : *Une* plaque sur un mur n'est évidemment pas beaucoup pour regarder pour ceux avec une passion pour les machines fonctionnantes et les paysages industriels dramatiques, mais *celui-ci* marque l'emplacement de la base de cette industrie qui a eu un plus grand impact sur la civilisation que *tout autre*.

(ii) Classe 2 – classe 1b

Les noms de la classe *2* étant non-spécifiés par rapport au trait sexe, on se retrouve en fait généralement face à une situation de type : un nom de classe *2* en LS correspond à deux noms de type *1b* pour lesquels le sexe est spécifié, car il est rare pour le type *1b* d'avoir une seule variante correspondant à un seul sexe[8].

La traduction est donc rendue difficile par *l'absence dans la LS d'une information concernant le sexe du référent,* spécifié dans la LC. Les solutions à ce type de problème seront évoquées plus loin. Au niveau anaphorique, on note les mêmes problèmes que ceux qui ont été évoqués jusqu'ici.

[7] Pour plus de précisions sur le caractère marqué/non marqué des pronoms, cf. Cristinoi 2007.
[8] Par exemple des noms de métier destinés habituellement à un seul sexe, comme *sage-femme* ; cela dit, il n'existe pas de nom de classe *2* dans les deux langues désignant ce type de réalité.

Langue	Classe	Unité lexicale	Genre grammatical	Genre référentiel	Pronom correspondant
ANG	*2*	*donkey*	neutre	indéterminé	3.SG.N
FR	*1b*	*âne*	masculin	masculin	3.SG.M
		ânesse	féminin	féminin	3.SG.F

(iii) Classe 4 – classe 1b

Ce couple est problématique dans les cas où les composants du GN dans la LS autres que le nom ne sont pas marqués du point de vue du genre, le nom lui-même n'ayant pas de genre inhérent. Dans ce cas, il sera difficile de déterminer le genre référentiel et même grammatical du nom dans la LS et de choisir son correspondant, marqué, dans la LC, du point de vue du genre.

Langue	Classe	Unité lexicale	Genre grammatical	Genre référentiel	Pronom correspondant
ANG	*4*	*dancer*	indéterminé	indéterminé	3.SG.M/F
FR	*1b*	*danseur*	masculin	masculin	3.SG.M
		danseuse	féminin	féminin	3.SG.F

(9) The dancer starts the show.
SYSTRAN: Le danseur commence l'exposition.
REVERSO : Le danseur commence l'exposition (le spectacle).
INTERTRAN: Les danseuse commencer the show.

(10) The dancer starts her show.
SYSTRAN: Le danseur commence son exposition.
REVERSO : Le danseur commence son exposition (spectacle).
INTERTRAN: Les danseuse commencer son être visible.

En laissant de côté la question lexicale de *show,* on remarque qu'aucun des systèmes testés ne prend en compte le fait que *dancer* pourrait être rendu à la fois par *danseur* et par *danseuse* (sauf peut-être d'une certaine façon INTERTRAN, mais il est difficile de dire, au vu de la qualité globale de la traduction, s'il s'agit ou non d'un choix totalement arbitraire). De plus, même lorsque l'information est rendue explicite par l'utilisation du déterminant possessif, le problème n'est pas résolu. L'exemple suivant illustre la même situation :

(11) Or you might be *a musician, an actor, a dancer, a singer* or a sports person who relies on functioning to peak efficiency in order to obtain the best results. (BNC-BM0 198)
SYSTRAN : Ou vous pourriez être *un musicien, un acteur, un danseur, un chanteur* ou une personne de sports qui compte sur fonctionner pour faire une pointe l'efficacité afin d'obtenir les meilleurs résultats.
REVERSO : Ou vous pourriez être *un musicien, un acteur, un danseur, un chanteur* ou une personne sportive qui compte sur le fonctionnement pour atteindre un niveau maximal l'efficacité pour obtenir les meilleurs résultats.

Pour ce qui est de la reprise anaphorique, les problèmes ne surgissent que dans le cas d'une dissymétrie de marquage.

3. Vers un cahier des charges pour un système de TA

L'analyse exhaustive des 13 combinaisons évoquées *supra* montre que de manière générale, les problèmes de traduction des noms nus et les erreurs qui s'ensuivent sont dus au déséquilibre provoqué par la quantité différente d'information sur le genre fournie par les noms dans chaque langue. Plus précisément, c'est le déficit ou le décalage en matière d'information référentielle qui pose problème, l'information grammaticale étant, elle, présente comme trait inhérent dans tous les noms à l'exception de ceux de la classe *4*.

3.1. Diagnostic des erreurs

Les tests effectués sur un corpus constitué de séries de phrases incluant des couples de noms illustrant les 13 combinaisons, traduites par des traducteurs automatiques, montrent que d'une manière générale les erreurs dans la traduction automatique du genre se divisent en deux catégories au sein desquelles plusieurs distinctions sont possibles.

3.1.1. Erreurs lexicales

- Sous-traduction (l'unité lexicale utilisée dans la LC est moins précise que celle utilisée dans la LS) :

(12) Le week-end dernier, j'assistais au mariage de ma *cousine*.
SYSTRAN: Last weekend, I attended the marriage of my *cousin*.

- Sur-traduction (l'unité lexicale utilisée dans la LC est plus précise que celle utilisée dans la LS), comme dans l'exemple (9) :

(9) The *dancer* starts the show.
SYSTRAN: Le *danseur* commence l'exposition.

- Traduction fausse (l'unité lexicale utilisée dans la LC n'est pas le bon équivalent pour celle utilisée dans la LS) comme dans l'exemple (10) :

(10) The dancer starts *her* show.
SYSTRAN: Le *danseur* commence son exposition.

3.1.2. Erreurs syntaxiques

Celles-ci peuvent se rencontrer tant au niveau de l'accord au sein du GN[9] qu'au niveau de l'accord anaphorique.

- Sous-traduction (l'anaphorique dans la LC perd l'information de genre présente dans la LS, sans être pour autant erroné) :

(13) He says *the victim* was walking to *her* home just a few minutes away along a very busy street (BNC- K1U 3873)

[9] Les erreurs de traduction du genre des déterminants au sein des GN sont très rares si les analyseurs syntaxiques sont bons, dans la mesure où il ne s'agit pas à proprement parler de traduction mais d'accord dans la LC, accord qui est habituellement réalisé correctement.

SYSTRAN: Il dit que *la victime* marchait à *sa* maison juste quelques minutes loin le long très d'une rue passante.

- Sur-traduction (l'anaphorique dans la LC ajoute une information de genre qui était absente dans la LS) :

(14) The two young journalists decided to leave but then *they* changed their minds.
SYSTRAN: Les deux jeunes journalistes ont décidé de partir mais d'autre part *ils* ont changé d'avis.

- Traduction fausse :

(15) I saw a tit and this one was bigger than the other.
REVERSO : J'ai vu une mésange et *celui-ci* était plus grand que l'autre.

3.2. Typologie des problèmes et solutions

Les problèmes théoriques de traduction du genre pour le français et l'anglais qui donnent naissance à ces différents types d'erreurs sont à traiter aux mêmes niveaux, en envisageant pour chacun d'entre eux les solutions à mettre en œuvre. Je traiterai ici les problèmes que soulèvent la traduction des noms nus, l'accord intraphrastique et l'anaphore interphrastique.

3.2.1. Traduction des noms nus

Deux types de problèmes peuvent surgir à ce niveau.

(i) Problème de type 1 (entre les couples *1b-2*, *1b-3*, *1b-4*, *4-3*) : l'information concernant le sexe du référent est présente dans la LS mais absente dans la LC, d'où perte de cette information. Deux solutions sont envisageables.

Tout d'abord, on peut adopter une solution *lexicale* pour les noms qui dans la LC appartiennent aux classes *2* et *3*, puisque même si les noms en question possèdent le même genre grammatical que le genre grammatical et référentiel du nom dans la LS, cette information reste arbitraire[10] et n'apporte pas d'éclaircissements sur le sexe du référent. Cette solution consiste à indiquer par des noms comme *homme-femme, mâle-femelle*, ou par des pronoms différenciés du point de vue du genre/sexe, le sexe du référent en question, spécifié dans la LS. C'est ainsi que *hérissonne*, nom appartenant à la classe *1b*, sera traduit en anglais par *female hedgehog*, l'utilisation de *female* étant obligatoire à cause de l'appartenance de *hedgehog* à la classe *2* en anglais.

La deuxième solution que l'on peut adopter, pour les noms de la classe *4* cette fois-ci, est une solution *syntaxique*, réalisable par l'indication du genre sur les déterminants : les noms devant obligatoirement porter deux informations identiques (genre grammatical et genre référentiel), les déterminants spécifiant le genre grammatical du nom en question spécifieront forcément son genre référentiel. Les cas de ce type étant marginaux pour le couple de langues traité, je donnerai un exemple roumain-français : *ziarist* (« journaliste », féminin, classe *1b*) – *une*

[10] Comme c'est le cas pour *une mésange,* par exemple, l'accord au féminin et le genre inhérent féminin ne disant rien sur le sexe de l'oiseau en question.

journaliste (classe *4*, l'élément qui porte le genre, et apporte la précision référentielle nécessaire, étant dans ce cas le déterminant).

(ii) Problème de type 2 (entre les couples *2-1b*, *3-1b* et *4-1b*) : l'absence de l'information concernant le sexe du référent dans la LS rend impossible le choix du nom (qui doit obligatoirement spécifier cette information) dans la LC.

La solution consiste ici à rechercher des *indices* sur le genre référentiel dans la LS. L'information nécessaire à la traduction du nom peut se trouver, lorsqu'elle est présente, soit sur les déterminants[11], ce qui est possible pour les noms de la classe *4*, soit en d'autres endroits de la phrase (indices lexicaux généralement), si les déterminants n'indiquent pas le genre, ou s'ils n'indiquent que le genre grammatical et ne donnent aucune information sur le sexe du référent.

(16) The lawyer grabbed *her* briefcase and left without a word.
L'avocate saisit sa serviette et s'en alla sans dire un mot.

Si dans le premier cas on peut espérer trouver facilement une solution, cela s'avère nettement plus difficile dans le deuxième, l'information en question étant souvent absente. Sans cette information, la solution est compliquée, mais non impossible pour autant, puisque dans la plupart des langues du monde, il existe toujours une forme utilisée « par défaut », une forme indiquant un emploi générique, qui dans les langues traitées ici est le masculin. Il suffit donc, pour résoudre le problème, en cas d'échec de toutes les autres solutions, d'utiliser cette forme générique qui permet d'évoquer ainsi les deux sexes.

3.2.2. Accord intraphrastique

A ce niveau, le caractère arbitraire (purement grammatical, sans fondement référentiel) ou motivé (déterminé sur des bases référentielles ou autrement dit par le sexe) du genre et de l'accord joue un rôle important dans la traduction.

(i) Si *dans les deux langues le genre est motivé* (genre grammatical = genre référentiel), les déterminants et autres porteurs d'accord auront la même valeur genre ; il n'y aura donc aucun problème d'accord, et une traduction littérale peut être envisagée pour des couples comme *1b-4* et *4-1b*.

(ii) La situation *genre arbitraire (LS) / genre motivé (LC)* génère des problèmes de type 2. Pour les couples *2-1b* et *3-1b*, cela se manifeste par l'absence d'information de type genre référentiel en LS permettant le choix du bon équivalent en LC, ce qui se répercute sur le choix des déterminants ; pour le couple *3-4* l'absence d'information de type genre référentiel, une fois trouvé le nom équivalent en LC, rendra difficile voire impossible le choix des déterminants, si ceux-ci marquent le genre.

(iii) Dans la situation *genre motivé (LS) / genre arbitraire (LC)* (*1b-2*, *1b-3*, *4-3*), aucun problème n'est identifié en principe pour l'accord intraphrastique puisqu'il n'y a pas de choix possible, le genre en LC étant imposé grammaticalement. Un

[11] Indices essentiellement morphologiques, dans la mesure où la forme de ces éléments indique leur genre, mais aussi syntaxiques dans la mesure où il s'agit d'éléments autres que le nom qui permettent de calculer le genre de celui-ci.

éventuel problème peut tout de même surgir dans le cas où les deux genres sont différents, ce qui rend impossible la traduction littérale et indispensable l'analyse syntaxique puisque le genre à attribuer aux éléments porteurs d'accord est à chercher dans la LC.

(iv) Si enfin *dans les deux langues le genre est arbitraire*, lorsque le genre est le même, l'on peut envisager sans difficulté une traduction littérale du GN, sans analyse syntaxique particulière ; lorsqu'il est différent, on se trouve dans le même cas que dans la situation décrite en (iii).

On voit surgir de la différence de genre dans les deux langues évoquée dans les cas (iii) et (iv) ci-dessus ce qui pourrait apparaître comme un troisième type de problème. Je ne lui accorderai pas d'importance particulière ici, dans la mesure où il peut être écarté par l'utilisation d'un bon analyseur syntaxique, qui fera le lien entre le nom et ses déterminants dans la LC et attribuera aux déterminants le genre du nom qui est tout à fait stable dans la LC, puisque arbitraire.

3.2.3. Anaphore interphrastique

Au niveau anaphorique, aucun problème nouveau n'apparaît. Il convient toutefois de préciser que le lien entre l'antécédent et l'expression anaphorique doit être établi dans la LS pour déterminer de quel couple il est question. La situation peut se résumer comme suit :

- pour les cas *motivé (LS) / motivé*[12] *(LC)* (*1b-3*, *1b-4*, *3-1b*, *3-3*, *3-4*, *4-1b*, *4-3*, *4-4*), le genre reste stable ;

- pour les cas *arbitraire (LS) / motivé (LC)*, le couple *2-1b* ne présente pas de problème si le nom *1b* a été préalablement déterminé ;

- pour les cas *motivé (LS) / arbitraire (LC)* (*1b-2*), nous avons affaire à un problème de type 1, c'est-à-dire à une perte de l'information du genre référentiel qu'il faudra récupérer d'une manière ou d'une autre si cela ne s'est pas effectué à un autre niveau (traduction du nom) ;

- pour les cas *arbitraire (LS) / arbitraire (LC)* (*1a-1a* et *2-2*), le genre de l'anaphorique peut changer en fonction du genre de l'antécédent, qu'il faut chercher dans la LC (tout comme dans le cas de l'accord intraphrastique), ce qui nécessite un bon analyseur syntaxique.

Pour résumer, un cahier des charges pour la traduction du genre doit inclure quatre étapes principales : la détermination de la classe de comportement à laquelle appartiennent les noms dans la LS et dans la LC, l'identification du type de problème de traduction en fonction du couple de classes de comportement, l'identification des solutions possibles et la traduction proprement dite. La démarche à suivre, couple par couple, pour l'implémentation du modèle, est récapitulée dans le tableau suivant :

[12] Dans ce cas, les termes « motivé-arbitraire » se réfèrent bien sûr à l'accord.

Classes LS-LC	*Types de problèmes*	*Solutions possibles*
1a-1a	éventuellement problème de type 3	effectuer une bonne analyse syntaxique dans la LS et utiliser le genre grammatical du nom dans la LC afin de réaliser l'accord correctement
1b-1b	aucun problème	
1b-2	problème de type 1	solution lexicale
1b-3	problème de type 1	solution lexicale
1b-4	problème de type 1	solution grammaticale
2-1b	problème de type 2	*solution 1* = chercher l'information genre référentiel dans la LS *solution 2* = si la solution 1 est impossible, utiliser la forme générique par défaut
2-2	éventuellement problème de type 3	effectuer une bonne analyse syntaxique dans la LS et utiliser le genre grammatical du nom dans la LC afin de réaliser l'accord correctement
3-1b	problème de type 2 au niveau intra-phrastique - aucun problème au niveau anaphorique	*solution 1* = chercher l'information genre référentiel dans la LS *solution 2* = si la solution 1 est impossible, utiliser la forme générique par défaut
3-3	aucun problème	
3-4	problème de type 2 au niveau intra-phrastique - aucun problème au niveau anaphorique	*solution 1* = chercher l'information genre référentiel dans la LS *solution 2* = si la solution 1 est impossible, utiliser la forme par défaut
4-1b	problème de type 2	rechercher l'indication genre référentiel sur les porteurs d'accord en LS
4-3	problème de type 1 au niveau intra-phrastique - aucun problème au niveau anaphorique	solution lexicale
4-4	aucun problème	

3.3. Développements possibles du modèle

La démarche d'identification des problèmes de traduction liés au genre peut être complexifiée, afin d'améliorer davantage les performances du système, en prenant en compte un nouveau facteur, à savoir le *caractère marqué – non marqué* du genre des éléments de la phrase autres que le nom (déterminants, anaphoriques et autres porteurs d'accord comme par exemple les participes passés). On qualifie ici de *non marquée* toute forme à laquelle il est impossible d'attribuer une valeur genre, quelle qu'elle soit (comme *the* en anglais, par exemple), par opposition à une forme *marquée*, à laquelle on peut attribuer une valeur genre.

Cette acception du marquage est étroitement liée à la question des *indices morphosyntaxiques* (Cristinoi 2007, 66), dans la mesure où ceux-ci permettent d'identifier d'emblée certaines formes linguistiques comme étant des formes

marquées (*mariée* – féminin*)* et de calculer la valeur des formes *non marquées* (indices syntaxiques). L'exemple (17) illustre les difficultés qui peuvent surgir lorsque l'on a affaire à des formes non marquées du point de vue du genre.

(17) *Les jeunes* journalistes sont *arrivées* hier soir. *Elles* étaient *ravies*.

SYSTRAN : The young journalists arrived yesterday evening. They were delighted.

Les et *jeunes* étant des formes non-marquées du point de vue du genre, tout comme *journalistes*, la seule indication de genre que l'on trouve est présente plus loin dans la phrase, sur le participe passé, puis sur l'anaphorique, *elles*, ce qui veut dire que si en LC une précision au niveau du genre était nécessaire pour choisir le bon équivalent, l'information serait à chercher plus loin qu'au niveau des déterminants, sur un autre porteur d'accord possible (c'est cette démarche qu'il convient d'indiquer à un logiciel de TA). Si dans la LC cette information n'est pas nécessaire pour choisir un nom ou un autre, ni pour faire l'accord (comme c'est le cas en anglais), le système n'aura aucune difficulté à traduire le couple de phrases, mais il perdra tout de même l'indication de genre puisque le genre n'est marqué dans ce cas ni sur les déterminants, ni sur les anaphoriques.

La dissymétrie du marquage du genre entraîne donc des problèmes de traduction supplémentaires qu'il est important de prédire (voir Cristinoi 2007, 220). Précisons simplement ici que si les problèmes de genre sont résolus au niveau du nom, il n'est pas indispensable de les résoudre à chaque niveau de la phrase ou du texte, et que la présence d'un bon analyseur syntaxique écarte la plupart des cas problématiques. Si les noms possèdent un genre grammatical inhérent et stable et que ce genre est utilisé pour l'accord, les problèmes n'ont pas lieu de se poser. Pour les noms de type 4 en revanche, l'absence de marquage du genre sur les porteurs d'accord dans la LS et l'obligation de le marquer dans la LC peuvent faire obstacle à la traduction, dans la mesure où ces marqueurs sont les seuls indices nous permettant de calculer le genre du nom. Si le genre n'est marqué sur les porteurs d'accord dans aucune des deux langues, alors les problèmes sont les mêmes que pour les noms nus. L'intérêt de compléter ainsi le modèle présenté est précisément d'anticiper ce genre de situations et d'indiquer la marche à suivre, ce qui enrichira le cahier des charges et améliorera les performances du logiciel de TA.

Conclusion

L'utilisation de la notion de classes de comportement, sur laquelle s'appuie l'analyse linguistique *a priori* des problèmes de traduction du genre entre le français et l'anglais, permet ainsi de prédire, tant au niveau des unités lexicales isolées qu'aux niveaux intra- et interphrastiques, les problèmes que pose à la traduction la structure différente des deux langues et de trouver en amont (en théorie) une solution automatisable à ces problèmes.

Même s'il nécessite une analyse linguistique assez complexe, le modèle présente l'avantage de fixer un cahier des charges précis pour la conception / amélioration d'un logiciel de TA. La démarche proposée permet de calculer, dès le niveau lexical, les problèmes de traduction liés au genre : il suffit en effet d'indiquer dans l'entrée lexicale de chaque nom la classe à laquelle il appartient et de prendre

en compte également le marquage du genre dans les langues traitées. Le modèle peut être étendu à d'autres langues, les mêmes analyses étant réutilisables pour de nouveaux couples de langues et pourrait aussi s'appliquer à d'autres catégories que le genre.

Pour aller encore plus loin dans la préparation du terrain pour la traduction automatique, il est également intéressant de dresser un inventaire exhaustif de tous les porteurs d'accord, de toutes leurs formes et de leurs caractéristiques en termes de marquage du genre dans les langues étudiées, ce qui doit permettre de calculer d'emblée les problèmes potentiels pour la traduction d'un nom accompagné d'un certain déterminant par exemple, dans la mesure où les déterminants peuvent constituer des indices syntaxiques du genre du nom dans le cas où il serait non marqué. Il suffit alors, en utilisant cet inventaire, de vérifier si un porteur d'accord (ou un anaphorique) rattaché à un nom est marqué ou non du point de vue du genre afin de décider à quel cas général on a affaire et d'identifier à la fois le problème et la solution.

Antonia Cristinoi-Bursuc
Université d'Orléans, UFR LLSH, EA 3850 LLL-Laboratoire Ligérien de Linguistique
10 rue de Tours, BP 46527, 45065 Orléans Cedex 02
<acristinoi@hotmail.com>

Références

Boitet, C. (1999) : A research perspective on how to democratize machine translation and translation aids aiming at high quality final output. *Proceedings of Machine Translation Summit VII: MT in the Great Translation Era*. Singapore, 13-17 Sept. 1999, 125-133.

Bowker, L. (2002) : *Computer-Aided Translation Technology*. Ottawa, Ottawa University Press.

Corbett, G. (1991) : *Gender*. Cambridge, CUP.

Craig, C. (ed.) (1986) : *Noun Classes and Categorization*. Amsterdam, Benjamins.

Cristinoi, A. (2005) : Aspects typologiques des problèmes de traduction. In *Comunicare profesionala si traductologie*, Actes du colloque *International Conference on Professional Communication and Translation Studies*. Timi□oara, Editura Politehnica, 191-199.

Cristinoi, A. (2007) : *Analyse contrastive des indices morphosyntaxiques nominaux de genre et de nombre en vue d'une […] traduction automatique. Applications sur le français, l'anglais et le roumain*. Thèse de doctorat, Université d'Orléans.

Hockett, C. (1958) : *A Course in Modern Linguistics*. New York, Macmillan.

Hutchins, J. (2001) : Machine Translation and human translation: in competition or in complementation? *International Journal of Translation* 13, 1-2, 5-20.

Hutchins, J. (2006) : Computer-based translation in Europe and North America and its future prospects. *JAPIO 2006 Yearbook*. Tokyo, Japan Patent Information Organization, 170-174.

Quah, C.K. (2006) : *Translation and Technology*. Basingstoke, Palgrave MacMillan.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik (1985) : *A Comprehensive Grammar of the English Language*. London, Longman.

Riegel, M., J.-C. Pellat, R. Rioul (1994) : *Grammaire méthodique du français*. Paris, PUF.

Senft, G. (ed.) (2000) : *Systems of Nominal Classification*. Cambridge, CUP.
Unterbeck, B. & *al*. (2000) : *Gender in Grammar and Cognition*. Berlin, Mouton de Gruyter.