

Traduction automatique ancrée dans l'analyse linguistique

Jessie Pinkham et Martine Smets
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
jessiep@microsoft.com
martines@microsoft.com

Résumé – Abstract

Nous présentons dans cet article le système de traduction français-anglais MSR-MT développé à Microsoft dans le groupe de recherche sur le traitement du langage (NLP). Ce système est basé sur des analyseurs sophistiqués qui produisent des formes logiques, dans la langue source et la langue cible. Ces formes logiques sont alignées pour produire la base de données du transfert, qui contient les correspondances entre langue source et langue cible, utilisées lors de la traduction. Nous présentons différents stages du développement de notre système, commencé en novembre 2000. Nous montrons que les performances d'octobre 2001 de notre système sont meilleures que celles du système commercial Systran, pour le domaine technique, et décrivons le travail linguistique qui nous a permis d'arriver à cette performance. Nous présentons enfin les résultats préliminaires sur un corpus plus général, les débats parlementaires du corpus du Hansard. Quoique nos résultats ne soient pas aussi concluants que pour le domaine technique, nous sommes convaincues que la résolution des problèmes d'analyse que nous avons identifiés nous permettra d'améliorer notre performance.

In this paper, we present an overview of the MSR-MT translation system for French-English, developed at Microsoft Research in the NLP group. Our system is based on rule-based analysis which produces logical form representations of the source language and the target language. These are aligned to produce mappings then stored in an example-base. We examine the development of the system since November 2000, and show that by October 2001, we had exceeded the quality of the commercial system Systran in the technical domain. We describe the linguistic work that allowed our system to improve. We present preliminary results on the Hansard parliamentary corpus. While these results are less impressive currently, we are convinced that changes to linguistic analysis will allow the system performance to improve.

Keywords – Mots Clés

Traduction automatique, français-anglais, base d'exemples.
Automatic translation, French-English, example-base.

1 Présentation du système de traduction

Les systèmes de traduction automatiques peuvent être basés sur des dictionnaires bilingues (Gerber 1997, voir Boitet 2001 sur l'importance des dictionnaires), et peuvent utiliser soit des règles de transfert, des bases de données d'exemples de traduction (Kaji 92, Meyers 2000, Watanabe et al. 2000), une langue intermédiaire pivot ou *interlangue* (Poyen et Vauquois 1959, Vauquois 1968, Nyberg 1991, Dorr 1993) ou des modèles statistiques (Brown et al. 1993, Al-Onaizan et al. 1999).

Le transfert dans notre système est basé sur les exemples; le module de transfert est composé d'une base de données d'exemples de traduction appris durant l'alignement de textes bilingues pendant la phase d'apprentissage. L'alignement est effectué entre les formes logiques des langues source et cible, qui sont le résultat de l'analyse des textes d'apprentissage. C'est donc dire l'importance que revêt l'analyse linguistique dans notre système. Le système français-anglais comprend les modules suivants:

- un analyseur du français, à base de règles
- un dictionnaire monolingue français
- une base de données de transfert, composée de textes alignés
- un dictionnaire bilingue français-anglais automatiquement appris à partir des textes de la base de données du transfert
- un générateur de l'anglais, à base de règles
- un dictionnaire monolingue anglais.

Certains de ces modules existaient avant que nous ne commençons notre travail en traduction automatique : l'analyseur et les dictionnaires monolingues ont été utilisés pour les correcteurs de grammaire, qui font partie des logiciels Word. Le générateur de l'anglais est utilisé pour la traduction automatique d'autres langues vers l'anglais, par exemple l'espagnol, le japonais et le chinois.

Nous n'utilisons plus de dictionnaire bilingue général traditionnel construit à la main. Nous avons en effet remarqué (Pinkham et Smets 2002a) qu'un dictionnaire bilingue général n'améliore pas la traduction, dans le type de système que nous avons, et peut même nuire à la qualité de la traduction. Nous avons donc décidé de n'utiliser qu'un dictionnaire bilingue appris automatiquement sur les corpus d'apprentissage.¹

Dans le reste de cette section, nous présentons avec plus de détails l'analyseur et le module de transfert.

¹ Le dictionnaire bilingue appris comprend 30,000 paires de mots dans le domaine technique. Nous utilisons un dictionnaire bilingue traditionnel de 500 mots pour les prépositions, conjonctions, et pronoms.

1.1 Analyseur à base de règles

L'analyseur comprend trois modules différents : la grammaire syntaxique, le rattachement, et les règles de la forme logique.

La grammaire est une grammaire syntagmatique augmentée (augmented phrase structure grammar). Les règles sont binaires (Jensen 1993), et expriment des contraintes et conditions sur leur application. Par exemple, la règle ci-dessous introduit le complément d'un verbe : la contrainte exprimée sur le verbe à gauche de la flèche spécifie qu'il doit être transitif et l'action déclarée sur le syntagme verbal (SV) à droite de la flèche définit le syntagme nominal (SN) comme l'objet du verbe. Cette règle est bien sûr très simplifiée, mais illustre le type de règle de la grammaire.

SV[Transitif] SN → SV [Objet = SN]

Il est nécessaire de préciser deux points importants : nous n'essayons pas de capturer une théorie linguistique précise en développant la grammaire. Notre point de départ est empirique. Nous travaillons avec des corpus dont nous essayons de couvrir les constructions. Si nous pouvons adopter l'analyse d'un phénomène linguistique qui ait un fondement théorique et qui apporte une explication cohérente, nous le faisons volontiers. Mais notre but n'est pas d'essayer de démontrer la validité d'une approche théorique. De même, nous n'excluons pas l'adoption de différentes approches théoriques dans notre grammaire, pour autant que ces approches rendent compte des faits linguistiques observés.

Un deuxième point important à souligner est que nous n'essayons pas de contraindre la grammaire à n'analyser que des phrases grammaticales (Heidorn 2000). C'est une grammaire d'analyse, et non de génération, et nous la voulons robuste. Les contraintes que nous ajoutons aux règles ont pour but de bloquer les mauvaises analyses des phrases rencontrées dans les corpus, non pas d'éliminer les analyses de phrases potentiellement émises mais non grammaticales. En d'autres mots, nous ne nous préoccupons pas de surgénération.

Nous voulons une grammaire flexible, et qui puisse rendre compte des faits linguistiques rencontrés dans des corpus variés. Notre grammaire est d'ailleurs à large couverture et très robuste. Même si l'analyse de la phrase complète échoue, nous obtenons quand même une analyse partielle des syntagmes qui ont été analysés avec succès.

Lors du développement de la grammaire, nous nous assurons que nous n'introduisons pas de problèmes pour des constructions traitées précédemment en maintenant des suites de régression contenant plusieurs milliers de phrases (plus de 6000 phrases² de corpus général, et plus de 1000 phrases d'un corpus technique de manuels d'ordinateur). Nous venons de créer un corpus supplémentaire de plus de 18000 phrases que nous allons utiliser à cette fin.

La grammaire produit ce que nous appelons le sketch : un arbre syntaxique, associé à un agrégat d'attributs et de valeurs de ces attributs, qui définissent les constituants de l'arbre. A ce niveau, nous n'avons pas résolu le problème de l'attachement de certains constituants, tels que les post-modificateurs (groupes prépositionnels, adjectifs, propositions relatives). En

² L'analyse des 6000 phrases de notre corpus général prend moins de deux minutes.

effet, pour pouvoir avec certitude déterminer l'endroit où s'attachent ces constituants, on a souvent besoin de pouvoir considérer l'arbre syntaxique dans sa totalité (donc, lorsque l'analyse en arbre est terminée). La stratégie adoptée est donc d'attacher ce type de constituant le plus bas et le plus à droite possible dans l'arbre. Et c'est lors du stage suivant de l'analyse, lors du rattachement, que le site de l'attachement de ces post-modificateurs est déterminé.

Le module de rattachement est composé d'un nombre limité de règles qui rattachent des constituants au site estimé correct. Ces règles sont heuristiques et se basent sur des données morphologiques, syntaxiques et sémantiques. Dans la phrase suivante, par exemple, la relative introduite par *que* est attachée dans le sketch à *commandes* (Image 1). Une règle de rattachement la rattache à *fichier*, utilisant l'information que *commandes* fait partie d'un syntagme prépositionnel introduit par *de* et n'est pas précédé d'un article (Image 2). Le syntagme nominal exprime donc, probablement, un type d'objet, et la relative devrait être attachée à la tête du syntagme nominal, non au syntagme prépositionnel. Dans cette même phrase, le syntagme prépositionnel introduit par *avant* est rattaché au verbe. Pour tester nos règles de rattachement, nous utilisons un corpus de régressions de 18.000 phrases.

Vous insérez la commande pause avant une section du fichier de commandes que vous ne voulez pas traiter.

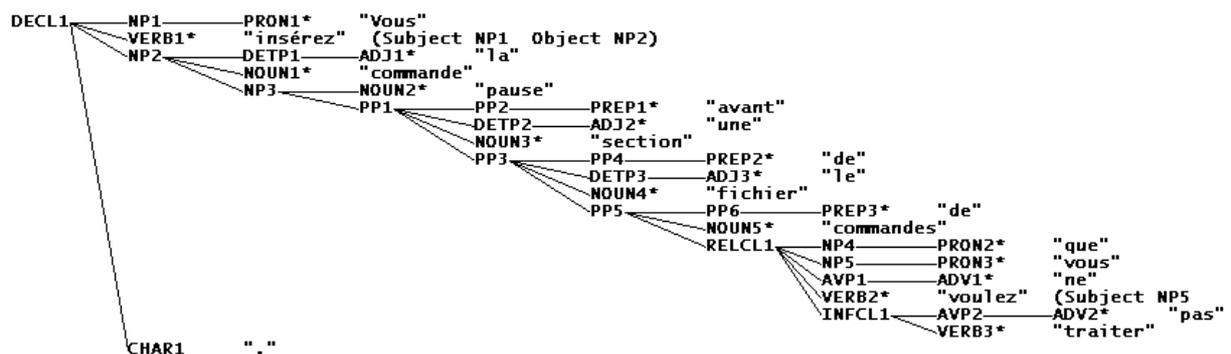


Image 1

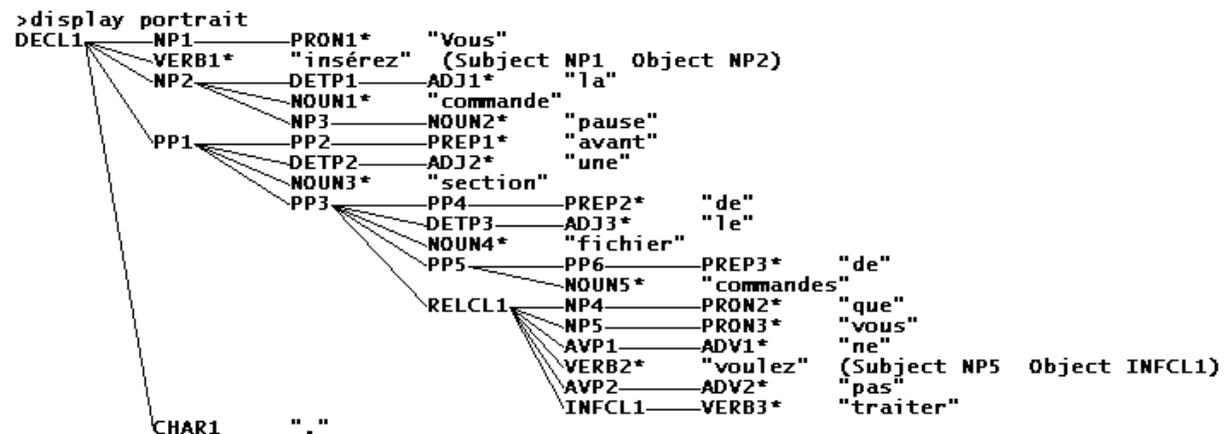


Image 2

La forme logique représente l'analyse profonde de la phrase (les fonctions syntaxiques, par exemple, telles que le sujet profond), et certaines informations sémantiques (par exemple, la signification de certaines prépositions : temporelles, de lieu, etc.). La forme logique est un graphe orienté étiqueté non ordonné, qui doit déterminer les fonctions syntaxiques dans des constructions à problèmes, telles que le contrôleur dans les infinitives et les rôles syntaxiques dans les constructions avec dépendances à longue distance (par exemple, l'objet dans la phrase *le fichier de commandes que vous voulez traiter*). La forme logique relie les formes actives et passives en leur donnant la même représentation ; elle exprime également la quantification, et les anaphores sont résolues dans la forme logique, mais seulement au niveau de la phrase.

L'image 3 représente la forme logique de la phrase analysée dans les images 1 et 2. La relation exprimée par le syntagme prépositionnel est représenté par l'attribut *TmeAt* ; la relative par l'attribut *Attrib*, qui introduit également les adjectifs modificateurs. Le sujet profond de *vouloir* est indiqué, de même que les compléments de *traiter*.

La forme logique est le but de l'analyse syntaxique, et également la base de notre composant de transfert: en effet, l'alignement est basé sur les formes logiques des segments alignés dans les deux langues.

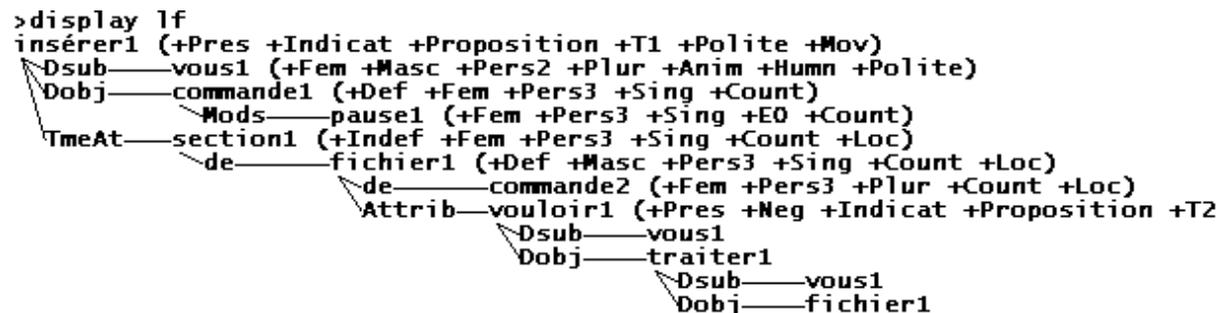


Image 3

1.2 Le composant de transfert

A la base du composant de transfert sont les formes logiques de la langue source et de la langue cible, ce qui souligne l'importance de l'analyse linguistique dans notre système. Les formes logiques résultant de l'analyse des corpus bilingues sont alignées, phrase par phrase, et des correspondances de traduction entre mots isolés, structures ancrées dans des mots spécifiques, ou syntagmes sont déduites de ces alignements. L'algorithme d'alignement procède en deux phases. Durant la première phase, des correspondances lexicales sont établies entre nœuds individuels des graphes de la langue source et de la langue cible, sur la base d'un dictionnaire appris automatiquement. Dans la seconde phase, les nœuds sont alignés sur la base de l'information structurale en plus de l'information lexicale. L'alignement se fait à partir des nœuds dont la correspondance lexicale est la plus étroite (meilleur d'abord, 'best—first'). Une série de règles d'alignement basées sur l'information linguistique sont utilisées dans cette deuxième phase. Plus de détails et des exemples sont donnés dans (Menezes et Richardson 2001).

Le dictionnaire bilingue utilisé dans la première phase d'alignement est appris automatiquement d'après un algorithme décrit dans (Moore 2001), à partir des corpus

bilingues d'apprentissage. Les corpus bilingues utilisés pour l'alignement comportent un nombre de phrases important : plus de 200.000 pour le corpus technique, et 500.000 pour le corpus du Hansard. Le corpus du Hansard est beaucoup plus général que notre corpus technique, ce qui explique la différence de taille. Les correspondances extraites de cet alignement sont représentées comme une base de données d'exemples pour la traduction. Durant la traduction, cette base de données est consultée pour la traduction de termes isolés ou de syntagmes entiers.

2 Evaluations du système de traduction

2.1 Méthode d'évaluation

La méthode dont nous nous servons a été décrite dans plusieurs articles récents (Richardson et al. 2001, Pinkham et al. 2001). Nous traduisons un ensemble de phrases non vues avec notre système de traduction, et avec le système de comparaison, en l'occurrence le système français-anglais de Systran. Nous associons à ces paires de traduction la phrase traduite par un traducteur humain, donc supposée correcte. Ce trio de phrases est présenté à plusieurs évaluateurs (entre 5 et 7), et les traductions automatiques sont présentées au hasard de façon à ce que les évaluateurs n'en sachent pas la provenance. La traduction humaine, ou traduction de référence est clairement indiquée. Les évaluateurs comparent les traductions automatiques, et notent celle qui est la plus proche de la traduction humaine comme étant la meilleure. Lors du calcul des scores, une préférence pour notre système se traduit par un score de 1 et une préférence pour Systran par un score de -1. Le score 0 indique que les systèmes sont équivalents. Nous utilisons de 250 à 400 phrases dans chaque évaluation.

2.2 Traduction pour le domaine technique (informatique)

Les progrès de notre système sont illustrés dans la Tableau 1, où nous indiquons tous les scores de novembre 2000 à janvier 2002. Le score de novembre 2000, par exemple, indique que la performance de notre système était sensiblement inférieure à celle de Systran (-0,50 sur une échelle de -1 à 1. L'expression +/- 0,1 indique l'intervalle de confiance).

Date de l'évaluation	Description du système	Score vs. Systran	Nombre de phrases
Novembre 2000	Système de base	-0,50 +/- 0,1	308
Février 2001	Dictionnaire amélioré	-0,18 +/- 0,1	250
Mars 2001	Lexique bilingue appris du domaine	-0,14 +/- 0,11	250
Octobre 2001	Améliorations linguistiques	+0,20 +/- 0,1	250
Janvier 2002	Dictionnaires appris automatiquement	+0,28 +/- 0,11	400

Tableau 1 : Progrès du système technique

En novembre 2000, les modules du système existaient, mais n'avaient pas été appliqués au français-anglais. Nous avons créé une base de donnée ('exemple-base') contenant des exemples de transfert à partir de 200 000 paires de phrases techniques bilingues alignées. Une première version très rudimentaire d'un dictionnaire bilingue a été employée. Les résultats par rapport à Systran étaient assez mauvais (-0,5). En février 2001, nous avons remplacé la première version du dictionnaire bilingue par une version de meilleure qualité, et les résultats sont devenus meilleurs car les alignements se sont améliorés. En mars, nous avons ajouté un lexique de paires de mots appris directement des textes d'apprentissage, ce qui a apporté une amélioration à peine significative. Ces résultats sont expliqués en détail dans l'article de Pinkham et Corston-Oliver (2001a). Pour examiner la contribution de chaque module du système, nous avons décidé de ne pas faire de modifications linguistiques pour l'analyse du français, et de ne pas faire non plus de changements à la génération de l'anglais motivés par les problèmes du français.

Nous avons attaqué les problèmes d'analyse du français à partir d'août 2001. Les améliorations linguistiques nous ont permis de dépasser la qualité de Systran en octobre 2001, avec un score de +0,20, qui montre que les évaluateurs ont trouvé notre système meilleur que Systran dans ce domaine technique. Bien sûr, nous savons que seul notre système avait été entraîné dans ce domaine. L'importance du résultat néanmoins, est que nous avons pu, en moins d'un an, avec un personnel très réduit (2 personnes pour le français), améliorer la performance de notre système de façon importante.

Les systèmes évalués jusqu'en octobre comprenaient un dictionnaire bilingue traditionnel. Nous avons été amenées alors à éliminer ce dictionnaire pour nous servir uniquement de dictionnaires de traductions appris des corpus eux-mêmes (Pinkham et Smets 2002a, Pinkham et Smets 2002b). L'évaluation de janvier donne les résultats d'un système qui ne comprend pas de dictionnaire bilingue traditionnel. Ce résultat de +0,28 représente notre meilleur résultat jusqu'à présent. Le meilleur score pour MSR-MT est en espagnol-anglais (+0,48), et nous sommes persuadées qu'avec davantage de travail linguistique, nous atteindrions des résultats semblables.

2.2.1 Exemples de modifications linguistiques dans le corpus technique

Dans cette section, nous discutons du travail linguistique qui a été nécessaire pour améliorer la qualité de la traduction dans le domaine technique. Nous avons remarqué qu'un problème lié au corpus technique est l'utilisation très fréquente de groupes nominaux avec fonction d'apposition. Ces groupes nominaux sont le plus souvent des noms de commandes ou de menus. Un exemple illustratif est le syntagme nominal *la commande pause*. Ce syntagme pourrait être analysé en syntagme verbal (le sujet *la commande* suivi du verbe *pauser*) ou en syntagme nominal, avec *pause* apposition de *commande*. Nous avons dû modifier la grammaire pour obtenir l'analyse correcte, tout en nous assurant que nous n'adaptions pas la grammaire à ce domaine spécifique mais préservions les analyses correctes de notre corpus de régression. D'autres exemples d'apposition incluent des syntagmes du type *la boîte de dialogue Ouvrir un fichier*. Dans cet exemple, la difficulté vient du fait que l'apposition est en fait un syntagme verbal, mais qui dans ce cas doit être analysé comme syntagme nominal. La majuscule est un indice que ce syntagme est un nom de menu, mais il n'est pas toujours aisé de déterminer où se termine le nom de menu (1). La phrase (2) présente un problème similaire, la présence d'un syntagme verbal qui doit être interprété comme syntagme nominal. Ces deux phrases sont maintenant correctement analysées par notre système.

(1) Si Dr. Watson ne peut pas utiliser le chemin spécifié pour créer un fichier de vidage sur incident, la boîte de dialogue **Ouvrir un fichier** s'affiche pour vous permettre de spécifier un nouveau chemin

(2) Si vous souhaitez toujours utiliser le rappel, cliquez sur **Toujours me rappeler au(x) ou aux numéro(s) ci-dessous**, et sélectionnez le modem ou le périphérique à rappeler.

Nous avons dû également résoudre des problèmes d'analyse plus généraux afin d'améliorer notre performance, tels que désambiguïser *des*, qui peut être une préposition ou un déterminant, désambiguïser entre participe passé et participe passif, etc. Enfin, nous nous sommes également penchés sur la traduction d'idiomes tels que *mettre à jour*, *mettre à niveau*, *prendre en charge*, etc. (Menezes et al. 2002)

Au niveau du rattachement, de nombreux syntagmes ont la structure SN de N SN (apposition) SA/SV, pour lesquels le syntagme adjectival (SA) ou verbal doit être rattaché à la tête du premier groupe nominal. Par exemple, dans la phrase (3), la participiale *défini dans...* est rattachée à *protocole*. La règle de rattachement se base sur le fait que la structure SN de N sans déterminant indique un type d'objet (identifié par l'apposition) et qu'un modificateur suivant l'apposition est en général attaché à la tête du premier SN.

(3) Les serveurs d'accès distant et leurs clients utilisent *le protocole de configuration PPP IPXCP* (IPX ///Configuration Protocol) *défini dans RFC 1552*, « The PPP Internetwork Packet Exchange Control Protocol (IPXCP) » pour configurer la connexion d'accès distant pour IPX.

En améliorant l'analyse de ce type de phrases, nous facilitons l'alignement des textes bilingues qui est nécessaire à la création de la base d'exemples de traduction.

2.3 Traduction pour le corpus du Hansard

Le corpus du Hansard est composé de débats parlementaires canadiens. Nous avons construit une base de données d'exemples ('exemple-base') avec 500 000 phrases alignées. Ces phrases ont été prises au hasard dans les 1,3 million de phrases disponibles sur le site web de Ulrich Germann (<http://www.isi.edu/natural-language/download/hansard/index.html>).

Les résultats de l'évaluation sont présentés dans le Tableau 2.

Date de l'évaluation	Description du système	Score vs. Systran	Nombre de phrases
Janvier 2002	Système de base sans changements	-0,16 +/- .12	250

Tableau 2: Traduction de texte Hansard

Le score de -0,16 montre que notre système général est pour l'instant moins bon que Systran. Pour ce test, nous avons incorporé plusieurs dictionnaires spécialisés de Systran (politique, économique, etc.) et nous considérons que le Hansard est un domaine général où l'entraînement est moins avantageux. Mais si nous comparons ce score au score technique, nous pouvons noter que nous étions au même point en mars de 2001, et que nous avons réussi

à améliorer le système en six mois. Nous pensons donc que nous pourrions améliorer la traduction de la même façon. Les plus gros problèmes que nous voyons sont des problèmes d'analyse.

2.3.1 Exemples de problèmes linguistiques dans le corpus Hansard

Texte original	Référence	MSR_MT	Systran	Classes de problème
Monsieur le Président, ma question supplémentaire, qui s'adresse encore une fois au ministre, est très simple.	Mr. Speaker, my supplementary question again to the minister is very simple.	My supplementary question, which is once again to the minister, is very simple Mr. Speaker.	Mr. President, my additional question, which is addressed to the minister once again, is very simple	Vocatif
C'est la seule arme qu'il leur reste.	It is the only tool we have left.	It is the only weapon that it remains for them.	It is the only weapon that it remains for them.	Verbe impersonnel
Le véritable problème, ce sont les récidivistes.	The real problem is the repeat offenders.	The real problem, it is the recidivists.	The true problem, they are the recidivists.	Complément disloqué à gauche
Quand la ministre entend-elle agir?	When does the minister intend to act ?	When does the minister intend for it to act?	When the minister intend does to act?	Sujet double avec pronom dans les interrogatives

Tableau 3 : Classes d'erreurs à corriger pour le domaine général

Les exemples présentés dans le Tableau 3 illustrent certaines constructions où nous savons que nous pouvons faire des progrès en améliorant l'analyse du français et de l'anglais si nécessaire. Ces constructions sont typiques du style parlé, ce qui explique que nous n'ayons pas rencontré beaucoup de cas jusqu'à présent dans nos fichiers de régression.

Le vocatif est très courant dans le corpus du Hansard. Lorsque le vocatif n'est pas reconnu comme tel par la grammaire, il risque de devenir l'objet direct ou indirect de la phrase par erreur. Dans la traduction MSR-MT (ligne 1 du tableau), nous remarquons qu'il n'a pas été reconnu correctement, et n'est pas placé en tête de phrase. Un autre exemple à corriger dans notre représentation est la dislocation du nom à gauche (ligne 3). En corrigeant la forme logique, nous pourrions facilement arriver à la traduction modèle de la référence, et donc dépasser la qualité de la traduction de Systran. Les autres erreurs, elles aussi typiques du langage parlé, pourront être corrigées dans notre système d'analyse.

3. Conclusion

La traduction de MSR-MT dépasse la qualité de la traduction de Systran dans le domaine technique, grâce en partie au travail linguistique sur les analyses du français. Nos résultats dans un domaine plus général (Hansard) montrent que la traduction est assez proche de la qualité de Systran, bien que nous n'ayons pour l'instant fait aucun travail spécifique au domaine. Nous avons identifié des problèmes linguistiques auxquels nous devons nous adresser pour améliorer la qualité de notre traduction automatique dans ce nouveau domaine.

Bibliographie

- Al-Onaizan Y., Curin J., Jahr M., Knight K., Lafferty J., Melamed D., Och F-J, Purdy D., Smith N. A., Yarowsky D. (1999), *Statistical Machine Translation: Final Report*, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD.
- Boitet C. (2001), Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes, *Actes de l'Atelier sur la Traduction automatique et applications en grandeur réelle, TALN 2001*.
- Brown P., Della Pietra S., Della Pietra V., Mercer R. (1993), The mathematics of statistical machine translation, *Computational Linguistics*, 19, 263-312.
- Dorr B. (1993), Interlingual Machine Translation: a Parameterized Approach, *Artificial Intelligence* 63(1&2), 429-492.
- Dorr Bonnie, Johnson P, Benoit J (1999), A Survey of Current Paradigms in Translation, *Advances in Computers*, Vol. 49, M. Zekowitz (Ed), Academic Press, London, 1-68.
- Kaji H., Kida Y., Morimoto Y. (1992), Learning Translation Templates from Bilingual Text, *Actes de COLING 1992*, 672-678.
- Heidorn G. (2000), Intelligent Writing Assistance, *Handbook of Natural Language Processing*, Robert Dale, Hermann Moisl et Harold Somers, eds., 181-207.
- Hutchins W. J. (2001), Machine Translation over fifty years, *Histoire Epistémologie Langage* vol. 23 (1), 7-31.
- Jensen K. (1993), The PLNLP English grammar, *Natural Language Processing: The PLNLP approach* Jensen K., Heidorn G. et Richardson S. eds. Boston: Kluwer Academic.
- Meyers A., Kosaka M., Grishman R. (2000), Chart-based transfer rule application in machine translation, *Actes de COLING 1998*, 843-847.
- Menezes A., Richardson S. (2001), A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora, *Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference*, 39-46.
- Nyberg E., Mitamura T. (1991), An Efficient Interlingua Translation System for Multilingual Document Production, *Actes de MT Summit III*, 55-61.
- Pinkham J. and Corston-Oliver M. (2001a), Adding Domain Specificity to an MT system, *Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference*, 103-110.
- Pinkham J., Corston-Oliver M., Smets M., Pettenaro M. (2001b), Rapid assembly of a large-scale French-English MT system, *Actes de MT Summit VIII*, 277-282.
- Pinkham J., Smets M. (2002a), Machine Translation without a Bilingual Dictionary, *Actes de TMI*, 146-156.
- Pinkham J., Smets M. (2002b), Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair, A paraître dans *Actes de COLING 2002*.
- Poyen J., Vauquois B. (1959), A propos d'un langage universel, *IFIP Congress 1959*, 132-137.
- Richardson S., Dolan W., Menezes A., Corston-Oliver M. (2001), Overcoming the Customisation Bottleneck Using Example-Based MT, *Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference*, 9-16.
- Menezes A., Pentheroudakis J., Smets M. (2002), Translation of verbal idioms, A paraître dans *Proceedings of International Workshop on Computational Approaches to Collocations*.
- Moore R.C. (2001), Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words, *Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference*, 79-86
- Vauquois B. (1968), A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation, *IFIP Congress 68*, 254-260.
- Watanabe H., Kurohashi S. and Aramaki E. (2000), Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation, *Actes de COLING 2000*, 906-912.