Un corpus d'erreurs de traduction

Guillaume Wisniewski[1,2] Anil Kumar Singh[2] Natalia Segal[3] François Yvon[1,2]
(1) Université Paris Sud 91 403 ORSAY CEDEX
(2) LIMSI–CNRS 91 403 ORSAY CEDEX
(3) Reverso–Softissimo, 5 rue Soyer, 92 500 NEUILLY
{wisniews,anil,yvon}@limsi.fr,nsegal@softissimo.com

RÉSUMÉ
Avec le développement de la post-édition, de plus en plus de corpus contenant des corrections de traductions sont disponibles. Ce travail présente un corpus de corrections d'erreurs de traduction collecté dans le cadre du projet ANR/TRACE et illustre les différents types d'analyses auxquels il peut servir. Nous nous intéresserons notamment à la détection des erreurs fréquentes et à l'analyse de la variabilité des post-éditions.

ABSTRACT
A corpus of post-edited translations

More and more datasets of post-edited translations are being collected. These corpora have many applications, such as failure analysis of SMT systems and the development of quality estimation systems for SMT. This work presents a large corpus of post-edited translations that has been gathered during the ANR/TRACE project. Applications to the detection of frequent errors and to the analysis of the inter-rater agreement of hTER are also reported.

MOTS-CLÉS : Traduction automatique,Analyse d'erreur,Post-édiition.

KEYWORDS: Machine Translation, Failure Analysis,Post-edition.

1 Introduction

La *post-édition* consiste à corriger les sorties d'un système de traduction automatique (TA) afin de produire une traduction de qualité. Cette pratique se développe de plus en plus, aussi bien dans le cadre de traduction professionnelle (Garcia, 2011), que pour l'évaluation des systèmes de TA : quantifier le nombre d'éditions nécessaires pour la post-édition, comme le fait le score hTER (Snover *et al.*, 2006), fournit une indication pertinente de la qualité d'un système de TA.

Le développement de la post-édition suscite le développent et la diffusion de corpus contenant des corrections de traductions (Potet *et al.*, 2012; Callison-Burch *et al.*, 2012). Le travail présenté dans cet article s'inscrit dans cette lignée et décrit la constitution et l'exploitation d'un nouveau corpus de corrections de traductions collecté dans le cadre du projet ANR-TRACE [1]. Le recueil de ce corpus permet de répondre à un des principaux objectifs de TRACE, à savoir le développement de mesures de confiance pour la TA (Zhuang *et al.*, 2012) et la détection de zones difficiles à

1. anr-trace.limsi.fr

traduire. D'autres usages sont également envisageables : il peut, par exemple, être utilisé pour identifier les limites des systèmes de traduction, ou encore pour étudier la cohérence des scores hTER et, de manière plus qualitative, la variabilité des post-éditions.

C'est sur ces derniers points que porte le travail présenté dans cet article : après avoir détaillé les caractéristiques du corpus et la manière dont les données ont été collectées (Section 2), nous discutons à la Section 3 de deux manières de mettre en évidence certaines limites des systèmes de TA. Nous présentons finalement, à la Section 4, une première analyse de la variabilité des post-éditions.

2 Description du corpus

Le corpus TRACE de corrections de traductions comprend 6 693 phrases (soit 109 689 mots) pour la direction français-anglais ; et 5 929 phrases (soit 120 378 mots) pour la direction anglais-français. Ces phrases ont été traduites par deux systèmes de TA : un système commercial à base de règles, SysRule, et un système statistique, NCode (Crego *et al.*, 2011; Le *et al.*, 2012), désigné par SysStat dans la suite du texte. Pour chaque direction de traduction, un traducteur professionnel confirmé [2] a ensuite corrigé une des deux traductions automatiques (choisie aléatoirement) pour produire la référence post-éditée. Conformément à l'usage, les traducteurs traduisaient vers leur langue maternelle. Le corpus TRACE contient, en outre, pour chaque direction de traduction, 1 000 phrases qui ont été corrigées par deux traducteurs différents. Ces corpus sont librement téléchargeables sur le site du projet TRACE. [3]

Ces données proviennent, pour moitié, de demandes de traduction d'utilisateurs « grand public » collectées sur le portail de traduction en ligne de Softissimo (3 434 phrases en français et 2 541 en anglais) ; l'autre moitié est issue d'extrait d'un site journalistique en ligne (2 268 phrases en français), de différents corpus utilisés dans les compagnes d'évaluation de traduction WMT (Callison-Burch *et al.*, 2012) (991 phrases en français et 864 en anglais) et IWLST (Cettolo *et al.*, 2012) (1 524 phrases en anglais) ainsi que d'une campagne d'évaluation de modules de désambiguïsation sémantique (Lefever et Hoste, 2010) (1 000 phrases en anglais). Les exemples de ce dernier sous-corpus sont accompagnés d'informations complémentaires, telles que des traductions de référence ou des annotations sémantiques, qui ont été collectées par les organisateurs de ces différentes campagnes d'évaluation.

Des consignes de correction précises (diffusées avec le corpus) ont été fournies aux traducteurs afin d'assurer que celles-ci soient *minimales* : l'objectif est d'obtenir des traductions jugées correctes (aussi bien au niveau du sens que de la langue) tout en restant le plus proche possible de la traduction automatique. Afin de garantir leur qualité, des échantillons des corrections ont été validées par un expert et, au besoin, des modifications ont été demandées aux traducteurs pour assurer le respect des consignes. Par ailleurs, les traductions corrigées ont été utilisées pour évaluer automatiquement la qualité des systèmes de TA. Comme le montre le Tableau 1, les principales métriques ont des valeurs bien plus élevées que celles généralement observées, montrant clairement que les références produites sont effectivement plus proches des sorties des systèmes que les références utilisées dans les campagnes d'évaluation. Ainsi, lorsque SysStat est évalué par rapport aux références fournies pour la campagne d'évaluation WMT 2012, son score

2. Au total, 10 traducteurs différents (5 pour chaque direction de traduction) ont été sollicités
3. `anr-trace.limsi.fr`

	SysStat	SysRule
BLEU↑	57,0	47,6
hTER↓	29,1	36,8
Météor↑	40,6	33,8

Table 1 – Évaluation des systèmes de TA quand les hypothèses post-éditées sont prises comme références. Les scores suivis de ↑ (resp. ↓) sont d'autant meilleurs qu'ils sont grands (resp. petits).

TER est de 56,3 (contre 36,8 ici). Notons également que, comme cela a déjà observé par ailleurs, les métriques automatiques défavorisent fortement le système à base de règles.

3 Analyse des limites des systèmes de TA

Nous montrons dans cette section comment la comparaison des hypothèses de traduction avec leur post-édition permet d'identifier certaines limites des systèmes de TA. Pour des raisons de place, seuls les résultats obtenus pour les traductions de l'anglais vers le français sont présentés.

3.1 Erreurs fréquentes

Le calcul de la distance d'édition entre les hypothèses de traduction et leur post-édition permet de déterminer automatiquement les corrections à effectuer pour rendre « acceptables » les traductions automatiques. L'étude des éditions les plus fréquentes permet de caractériser certaines limites des systèmes de TA actuels.

Une première observation porte sur le type des éditions fréquentes : il s'agit essentiellement de substitutions (Tableau 2), même si le système à base de règles a tendance à produire des traductions trop longues. Une part non négligeable des substitutions (près de 9 %) correspond à la modification de la terminaison d'un mot (par exemple, « penserai » est corrigé en « penserais », « spéciales » en « spécial », ...). Il est toutefois difficile d'évaluer si ces modifications sont des corrections isolées (par exemple, pour corriger une erreur d'accord) ou si bien elles découlent d'autres corrections (accord d'un adjectif suite à la substitution du mot avec lequel il s'accorde).

Une étude statistique des éditions montre que la plupart des modifications (près de 70 %) sont uniques, ce qui rend difficile l'identification de motifs d'erreurs . Les erreurs les plus fréquentes portent presqu'exclusivement sur des mots outils (Table 3) et, comme précédemment, il est difficile de savoir si ces révisions sont dues à des erreurs de la TA, ou bien découlent d'autres corrections. Le filtrage des mots outils permet de faire apparaître certains motifs d'erreurs récurrents. Ainsi, sur les 5 929 traductions du corpus, la traduction de « *order* » par « ordre » a été corrigée 23 fois en « commande » et « maison » 10 fois en « chez ... ». Une centaine de motifs de ce type ont été extraits, même si tous ne sont pas aussi facilement interprétables.

opération	SYSSTAT	SYSRULE
déplacement	2 861	3 473
substitution	10 065	10 991
suppression	3 572	7 371
insertion	2 502	2 263

TABLE 2 – Nombre d'opérations nécessaires pour corriger les sorties des deux systèmes de TA

Substitution		Insertion		Suppression	
148	les → des	380	de	799	de
93	des → les	233	la	335	à
60	la → le	204	le	329	la
57	du → le	204	a	278	le
55	des → de	184	à	277	que
53	du → de	141	dans	256	les
51	de → des	131	que	242	en
46	de → pour	99	en	215	et
43	cela → il	97	un	212	des
42	une → un	96	des	167	pour

TABLE 3 – Corrections les plus fréquentes

3.2 Différences entre les traductions automatiques et leur post-édition

Une autre analyse, inspirée des travaux en estimation de confiance pour la traduction (Kulesza et Shieber, 2004), permet d'avoir une vision plus globale des différences entre hypothèses de traduction et traductions post-éditées. Cette analyse repose sur l'apprentissage d'un classifieur capable de distinguer ces deux types de traductions et l'étude des caractéristiques utiles pour faire cette distinction. Le même principe peut être utilisé pour caractériser les différences entre les références obtenues en post-éditant des hypothèses de traduction et les références « libres » utilisées dans les campagnes d'évaluation de la traduction.

Dans les expériences de cette section, chaque traduction est représentée par un ensemble de 336 caractéristiques utilisées dans un système d'estimation de confiance pour la TA (Wisniewski *et al.*, 2013). Ces caractéristiques se répartissent en quatre grandes catégories :
– des mesures de la qualité de l' « association » entre la source et l'hypothèse de traduction, telles des caractéristiques dérivées des modèles d'alignement ;
– des mesures de la fluidité et de la grammaticalité de l'hypothèse de traduction ainsi que de la phrase source, telles des caractéristiques dérivées des modèles de langue ;
– des caractéristiques de surfaces telles le nombre de mots hors vocabulaire, de signes de ponctuation, ... ;
– des caractéristiques syntaxiques simples comme le nombre de noms, de mots outils, ...
Une liste complète des caractéristiques utilisées est donnée dans (Wisniewski *et al.*, 2013).

Pour mener cette analyse, nous avons utilisé comme classifieur une forêt aléatoire (Breiman, 2001), une méthode d'apprentissage ensembliste qui repose sur la combinaison des prédictions de plusieurs arbres de décision. Les forêts aléatoires ont montré leur efficacité dans de nombreuses tâches ; elles sont connues pour être particulièrement robustes au sur-apprentissage et pour permettre la modélisation d'interactions complexes entre les caractéristiques. En plus de la construction d'un classifieur, l'algorithme d'apprentissage permet d'estimer l'*importance* de chaque caractéristique (Breiman, 2001) qui quantifie directement son *pouvoir discriminant* : plus cette importance est élevée, plus la caractéristique est utile à la prédiction de l'étiquette.

Nous avons utilisé, dans nos expériences, l'implémentation des forêts aléatoires fournies par la bibliothèque `scikit-learn` (Pedregosa *et al.*, 2011). Les paramètres de la forêt aléatoire sont appris sur 2/3 des données ; le dernier tiers des données étant utilisé pour évaluer les

performances du classifieur. L'ensemble des hyper-paramètres sont choisis par validation croisée.

La première tâche considérée a pour objectif de distinguer les traductions produites par un système de TA de leur post-édition : elle nécessite donc de distinguer automatiquement une bonne traduction d'une mauvaise. C'est une tâche difficile, ces deux traductions étant par construction proches l'une de l'autre. Il n'est donc pas surprenant que la précision du classifieur ne soit que de 63 % en apprentissage et de 59 % en test. Les performances de la seconde tâche, visant à distinguer les références obtenues par post-édition des références « libres » sont sensiblement meilleures : la précision en apprentissage est de 71 % et de 67 % en test.

Les 8 caractéristiques les plus discriminantes et leur importance sont représentées Figure 1. Pour les deux tâches, seules quelques caractéristiques sont discriminantes et celles-ci sont presque uniquement dérivées des scores de modèles de langue. Les modèles de langue neuronaux (Le *et al.*, 2011) (caractéristiques comportant SOUL dans leur nom), appliqués aussi bien à la source qu'à la traduction, jouent un rôle prédominant, surtout pour la distinction entre les hypothèses de traduction et leur post-édition. Ces caractéristiques sont complétées par des modèles de langue « classiques » appris aussi bien sur les étiquettes morpho-syntaxiques (POSLMLOGPROB correspond à la log-probabilité d'une séquences d'étiquettes morpho-syntaxiques) que sur les mots (BIGRAMSFREQQUARTILE1 décrit le pourcentage de bi-grams dont la fréquence est dans le premier quartile). Dans tous les cas, les valeurs des caractéristiques sont plus faibles pour les traductions automatiques que pour les hypothèses post-éditées qui ont elles-mêmes des valeur plus faibles que celles observées dans les références libres. Cette observation indique soit que l'espace de recherche des systèmes de TA n'est pas assez riche puisque le système de TA n'est pas capable de générer des hypothèses suffisamment « fluides », soit le modèle de langue n'a pas un poids suffisant dans la fonction de score qui permet au système de TA d'évaluer la qualité des hypothèses. Des expériences supplémentaires sont toutefois nécessaires pour déterminer laquelle de ces deux hypothèses est correcte.

Parmi les autres caractéristiques importantes, on peut noter la présence de descripteurs de surface simples décrivant les longueurs des phrases (SENLENGTH), le nombre de signes de ponctuation (NUMPUNC) ou la longueur moyenne des tokens (AVGTOKENLENGTH). Finalement, la caractéristique la plus importante pour distinguer les références post-éditées des références libres est fondée sur la probabilité d'alignement de la traduction avec la source, telle qu'estimée par un modèle IBM 1, et quantifie le nombre moyen de mots dont la probabilité d'alignement est plus grande que 0,02.

4 Évaluation de la variabilité des post-éditions

Une autre application du corpus TRACE est l'étude de l'accord inter-annotateur de la post-édition, puisque, pour chaque direction de traduction, 1 000 traductions ont été corrigées deux fois indépendamment. À notre connaissance, c'est la première fois que deux annotateurs différents corrigent les mêmes phrases, permettant une comparaison des post-éditions et une estimation de l'accord inter-annoteur du score hTER. Pour des raisons de place, nous décrirons uniquement les résultats obtenus sur le corpus de traductions de l'anglais vers le français. Les résultats pour la direction français vers anglais sont similaires.

De manière quantitative, il est possible de mesurer la similarité entre les post-éditions effectuées par les différents correcteurs en mesurant la corrélation entre les scores hTER obtenus lorsque

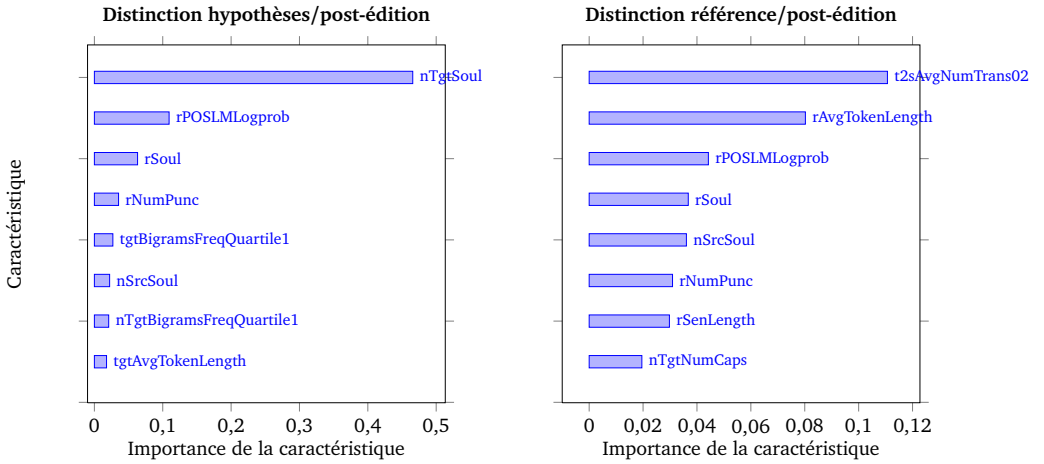

FIGURE 1 – Importance des caractéristiques les plus discriminantes pour les deux tâches considérées. Les caractéristiques dont le nom commence par un N sont normalisées par la longueur de la phrase ; celles dont le nom commence par un R sont constituées par le rapport entre les valeurs de la caractéristique calculée sur la phrase source et sur la traduction.

ces corrections sont utilisées comme référence. Cette corrélation est faible : le coefficient de Pearson entre les deux notes n'est que de 0,642 et le τ de Kendall de 0,476. L'interprétation est que si les traductions étaient ordonnées suivant leur score hTER, deux traductions quelconques ne seraient dans le même ordre pour les deux références qu'une fois sur deux. De manière globale, les post-éditions produites ne sont identiques que dans 12 % des cas[4]. La distance d'édition normalisée moyenne entre les deux post-éditions est de 24 % : il faut donc, pour passer d'une post-édition à l'autre, changer en moyenne un mot sur quatre. Bien qu'elles ne soient pas directement comparables, puisque dans l'un des cas le score (h)TER n'est pas calculé par rapport à une référence « adaptée », cette valeur est à peine plus petite que celle observée lors de l'évaluation des sorties de SYSSTAT. Ce résultat illustre les limites de l'évaluation de la TA par des score (h)TER. Les opérations les plus fréquentes dans cette transformation sont les substitutions de mots (57 % des modifications) suivi des suppressions et des insertions de mots (16 % dans les deux cas) ; les déplacements de mots n'interviennent que dans 11 % des cas.

Plus qualitativement, le Tableau 4 reprend des exemples des corrections les plus différentes ainsi que des phrases sources et des traductions automatiques. Ces exemples illustrent la variété des différences entre les post-éditions qui peuvent être dues à :

– une sensibilité différente aux traductions littérales : dans de nombreux cas, un correcteur accepte une traduction parfaitement compréhensible et juste d'un point de vue grammatical, même si elle n'aurait jamais été « produite » par un locuteur natif, alors que le second préfère la reformuler (4e exemple) ;
– une reformulation non nécessaire de la traduction automatique (le second correcteur qui corrige « cette réglementation » en « le présent règlement » dans le 1er exemple)
– une utilisation de paraphrases ou de synonymes sans raisons apparentes (« ultramodernes »

4. La comparaison entre les deux corrections ne tient compte ni de la ponctuation, ni de la casse.

1.	source	Each year, the Member States shall send the Commission a report on the evaluation of the execution and effectiveness of this regulation.
	trad. autom.	Chaque année, les États membres transmettent à la Commission un rapport sur l'évaluation de l'exécution et l'efficacité de cette réglementation.
	correction n° 1	Chaque année, les États membres **transmettent** à la Commission un rapport **sur l'évaluation de** l'exécution et l'efficacité **de cette réglementation.**
	correction n° 2	Chaque année, les États membres **communiquent** à la Commission un rapport **d'évaluation concernant** l'exécution et l'efficacité **du présent règlement.**
2.	source	I'm thinking this must be an ancient print date, right.
	trad. autom.	Je retiens ce doit être une date imprimée antique.
	correction n° 1	Je pense qu'il s'agit une **ancienne édition, c'est évident.**
	correction n° 2	Je pense que ça doit être une **ancienne date d'impression, n'est-ce pas.**
3.	source	So let's take a tour of this state-of-the-art clean coal facility.
	trad. autom.	Donc prenons un tour de cet état de l'art nettoient la facilité de charbon.
	correction n° 1	Alors allons voir **ces installations ultramodernes** de charbon propre.
	correction n° 2	Donc faisons une visite de **cette installation** de charbon propre **à la pointe de la technologie.**
4.	source	Dear Valued Customer, please follow the steps below to have a troubleshooting.
	trad. autom.	Cher valorisées à la clientèle, veuillez suivre les étapes ci-dessous pour avoir un dépannage.
	correction n° 1	Cher client estimé, veuillez suivre les étapes ci-dessous **pour avoir un dépannage**.
	correction n° 2	Très cher client, veuillez suivre les étapes ci-dessous **pour être dépanné**.

TABLE 4 – Exemple de différences de post-éditions.

versus « à la pointe de la technologie » dans le 3e exemple) ;
– une ambiguïté liée au manque de contexte en source (« cette installation » *versus* « ces installations » dans le 3e exemple).

Remarquons que les corrections sont différentes aussi bien quand la traduction automatique est *plutôt* bonne (1er exemple) que quand elle est complètement fausse (2e et 3e exemples).

Ces observations mettent en évidence les limites inhérentes à l'évaluation des systèmes de TA par un score comme hTER : dans la mesure où la post-édition semble aussi subjective que la traduction elle-même, les scores hTER seront aussi variables et difficiles à interpréter que les autres métriques automatique utilisées pour évaluer la TA.

5 Conclusion

Nous avons présenté, dans ce travail, un grand corpus de corrections de traductions et illustré différents types d'analyse que celui-ci rend possible. Bien qu'ils ne soient que préliminaires, les résultats présentés sont déjà riches en enseignements : ils montrent notamment les limites de la métrique hTER et illustrent une manière d'identifier les erreurs fréquentes en traduction. D'autres exploitations sont possibles, notamment en exploitant les annotations complémentaires qui sont disponibles pour diverses sous-parties du corpus TRACE. Nos travaux futurs ont pour objectif d'approfondir ces observations et d'arriver à les intégrer dans les systèmes de TA afin d'améliorer la qualité des hypothèses produites. Une autre piste de recherche consiste à comparer les erreurs faites par les systèmes de TA aux erreurs faites par les humains en utilisant, par exemple, des corpus contenant des corrections de traduction (Abekawa *et al.*, 2010).

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche au travers du projet ANR/CONTINT-2010/TRACE.

Références

Abekawa, T., Utiyama, M., Sumita, E. et Kageura, K. (2010). Community-based construction of draft and final translation corpus through a translation hosting site minna no hon'yaku (mnh). *In Proc. of LREC*. ELRA.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R. et Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. *In Proc. of WMT*, pages 10–51, Montréal, Canada. ACL.

Cettolo, M., Girardi, C. et Federico, M. (2012). Wit^3 : Web inventory of transcribed and translated talks. *In Proc. of EAMT*, pages 261–268, Trento, Italy.

Crego, J. M., Yvon, F. et no, J. B. M. (2011). N-code : an open-source Bilingual N-gram SMT Toolkit. *Prague Bulletin of Mathematical Linguistics*, 96:49–58.

Garcia, I. (2011). Translating by post-editing : is it the way forward ? *Machine Translation*, 25:217–237.

Kulesza, A. et Shieber, S. M. (2004). A learning approach to improving sentence-level mt evaluation. *In Proc. of TMI*.

Le, H.-S., Lavergne, T., Allauzen, A., Apidianaki, M., Gong, L., Max, A., Sokolov, A., Wisniewski, G. et Yvon, F. (2012). LIMSI @ WMT12. *In Proc. of WMT*, pages 330–337, Montréal, Canada. ACL.

Le, H. S., Oparin, I., Allauzen, A., Gauvain, J.-L. et Yvon, F. (2011). Structured Output Layer Neural Network Language Model. *In Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 5524–5527, Prague, Czech Republic.

Lefever, E. et Hoste, V. (2010). Semeval-2010 task 3 : Cross-lingual word sense disambiguation. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden. ACL.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. et Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python . *JMLR*, 12:2825–2830.

Potet, M., Esperança-Rodier, E., Besacier, L. et Blanchon, H. (2012). Collection of a large database of French-English SMT output corrections. *In Proc. of LREC*, Istanbul, Turkey. ELRA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *In Proc. of AMTA*, pages 223–231.

Wisniewski, G., Singh, A. K. et Yvon, F. (2013). Quality estimation for machine translation : Some lessons learned. *Machine Translation*, page accepté pour publication.

Zhuang, Y., Wisniewski, G. et Yvon, F. (2012). Non-linear models for confidence estimation. *In Proc. of WMT*, pages 157–162, Montréal, Canada. ACL.