Analyse de la qualité des traductions automatiques du français vers l'anglais, d'Expressions Poly-Lexicales (EPL) à partir d'un corpus parallèle – Quelles sont les erreurs les plus fréquentes par type d'EPL ?

Mots-Clefs : Évaluation de la qualité de traduction, Analyse d'erreurs de traduction, Expressions Poly-Lexicales, Corpus Parallèle.

Depuis plusieurs années, l'évaluation des systèmes de traduction automatique (noté TA ci-dessous) est au cœur de nombreux travaux. Les travaux les plus connus proposent des outils permettant des évaluations automatiques telles que la métrique BLEU (Papineni et al., 2002), qui vise à aider les développeurs afin d'évaluer leurs outils de manière rapide, peu coûteuse, et surtout avec une haute corrélation avec les jugements humains, ou plus proche de nous, la dernière version de la métrique METEOR (Denkowski et al., 2014) ou encore son amélioration par Servan (2016). D'autres études portent sur des critères d'évaluation tels que proposé par NIST (Doddington et al., 2002), le FEMTI (King et al., 2003), l'analyse d'erreurs de traduction suivant des critères linguistiques à partir de la typologie de Vilar (2006). Plus récemment, le projet QT21 (2016) propose des métriques multidimensionnelles de qualité (MQM) ; alors que QuEst++ propose la prédiction de qualité de systèmes (- et al., 2015).
La majorité de ces travaux vise à l'évaluation des systèmes de TA afin de les améliorer. Ils sont tous grandement tournés vers les développeurs afin qu'ils modifient les outils pour augmenter leur qualité de traduction. Lorsqu'il s'agit d'évaluer les outils selon des types d'erreurs à corriger, en terme de mesure de correction à mener pour fournir une meilleure traduction, les critères sont le plus souvent fortement liés au couple de langues utilisés ainsi qu'à la méthode de traduction.

Le travail présenté ici, bien que basé sur l'analyse d'erreurs des sorties de traduction, ne se destine pas à évaluer des systèmes de TA pour viser à leur amélioration, mais plutôt à évaluer la capacité de l'outil à fournir une traduction correcte de certains phénomènes linguistiques, afin d'orienter les utilisateurs de ces traductions vers l'outils le plus adéquat à leur tâche de traduction et à leur compétence. Ainsi, cette étude s'intéresse à l'évaluation de la qualité de traduction des expressions poly-lexicales (notées ci-après EPL) dans le cadre de traductions automatiques destinées à être utilisées par des apprenants de langues. Pour un apprenant, l'utilisation d'outils de TA afin de produire des écrits de bonne qualité dans leur seconde langue vivante, va être fortement liée aux compétences linguistiques qu'ils possèdent dans cette langue. Pour un apprenant débutant (A2 du CECRL), savoir reconnaître une EPL n'est pas évident et repérer qu'une EPL est bien traduite l'est encore moins.
Le couple de langues étudié dans cet article est le français vers l'anglais. Le corpus parallèle utilisé est un document français issu du domaine technique et de sa traduction en anglais, d'environ 12.566 mots. Ce document a été traduit du français vers l'anglais grâce à un outil de traduction maison, basé sur Moses (Koehn et al., 2007). Les EPLs étudiées dans cet article, ont été décrites selon les spécifications de Tutin (2015) qui les classe en neuf types. Parmi ces neufs types, nous nous concentrerons sur cinq d'entre eux à savoir : mots fonctionnels (F), phrasèmes (PH), collocations (C), entités nommées (EN), termes complexes (T).
L'évaluation de la qualité de la traduction de ces EPLs a été réalisée avec l'outil Blast (Stymne, 2011) et la typologie d'erreur de Vilar (2006) contenant 5 grandes catégories d'erreur déclinées en sous-catégorie (Figure 1).
Une description plus générale de ce travail a été présentée à la conférence TC38 en 2016.

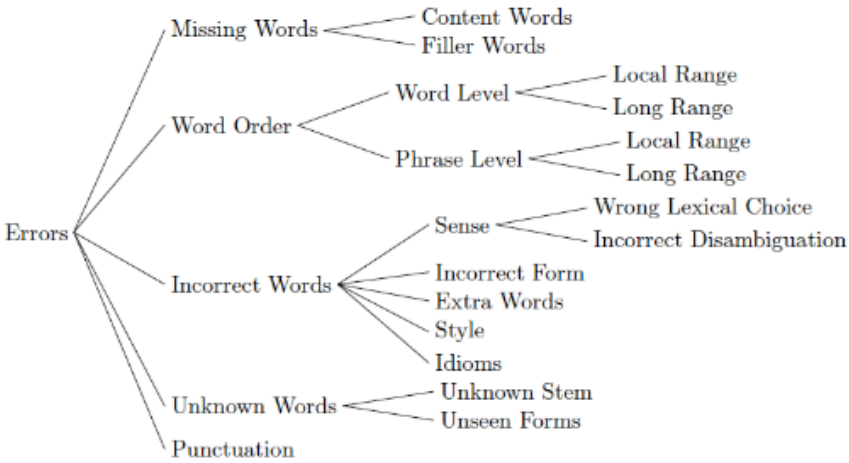

Figure 1: Spécification de la Typologie d'Erreurs de Vilar (2006)

L'étude qui est en train d'être menée à l'heure où ce résumé est écrit, n'est pas organisée par rapport aux types d'erreurs, mais par types d'EPLs. Ainsi, il est possible de repérer si un type d'erreur de traduction est plus souvent associés à un type d'EPL ou non.
La figure 2 montre quels sont les types d'erreur de la typologie de Vilar, en pourcentage, associés à chaque type d'EPL.

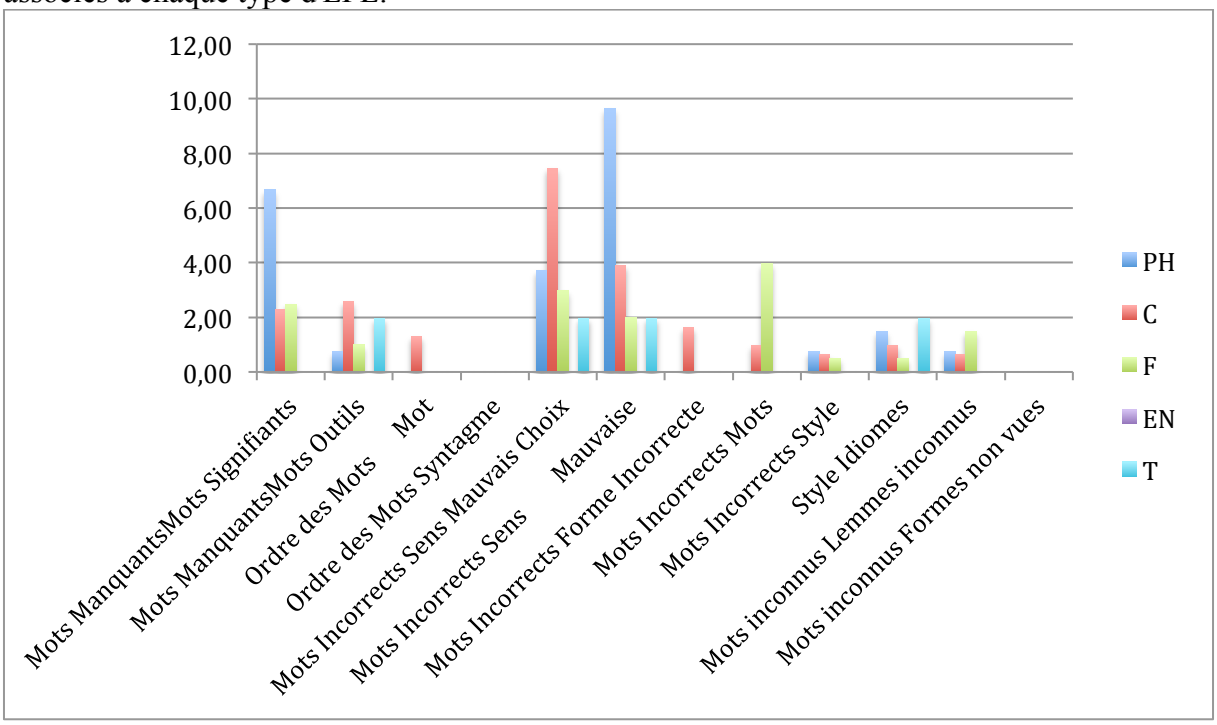

Figure 2 : Type d'erreurs associées à chaque type d'EPL (%)
Tout d'abord aucune EPL n'est associée aux types d'erreur d'ordre des mots au niveau du syntagme, ni des mots inconnus car leur forme n'est pas connue.
Force est de constater que les Phrasèmes (PH) connaissent le plus d'erreur du type mots incorrects car il y eu une mauvaise désambiguïsation du sens (9,7%), et en second, du type mot signifiants manquants (6,7%). Ensuite, le type Collocations (C), quand mal traduit, ont été annotées à 7,47% comme mots incorrects à cause d'un mauvais choix lexical.
Enfin, ce sont les EPLs mots Fonctionnels (F) qui ont été annotées avec le plus de type d'erreur mots supplémentaires (soit presque 4%).
Par respect du format, des exemples concrets ne peuvent être donnés ici.

Michael Denkowski and Alon Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language",Proceedings of the EACL 2014 Workshop on Statistical Machine Translation, 2014

Kishore Papineni, Salim Roukos, Todd Ward and Wei- Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA. July 2002. pp. 311-318.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co- occurrence Statistics. In Proceedings of 2nd Human Language Technologies Conference (HLT-02). San Diego, CA. pp. 128-132.

Margaret King, Andrei Popescu-Belis and Eduard Hovy. 2003. FEMTI: Creating and Using a Frame- work for MT Evaluation. In Proceedings of MT Summit IX, New Orleans, LA. Sept. 2003. pp. 224- 231.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi,Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions , pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

QT21: Quality Translation 21. Available at: http://www.qt21.eu/. Accessed: September 25, 2016.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier (2016). Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources?. COLING 2016, Dec 2016, Osaka, Japan. 26th International Conference on Computational Linguistics (COLING 2016), 2016.

Specia L., Paetzold G. H. et Scarton C. (2015): Multi-level Translation Quality Prediction with QuEst++. Actes de ACL-IJCNLP 2015 System Demonstrations, Beijing, China, 115-120.

Sara Stymne, 2011. Blast: A tool for error analysis of ma- chine translation output. In Proc. of the ACL 2011 System Demonstrations, pages 56–61, Portland, OR, USA, Jun. ACL.

Agnès Tutin, Emmanuelle Esperança-Rodier, Manuel Iborra, Justine Reverdy, 2015, Annotation of multiword expressions in French. Malaga, Espagne, Actes de la conférence Europhras2015, Juin 2015.

David Vilar, Jia Xu, Luis Fernando D'Haro et al., 2006. Error analysis of statistical machine translation output. In : Proceedings of LREC. 2006. p. 697-702.