KRCTool : un concordancier bilingue pour l'aide à la révision

Firas Hmida, Emmanuel Morin, Béatrice Daille, Emmanuel Planas

KRCTool : un concordancier bilingue pour l'aide à la révision

Firas Hmida Emmanuel Morin Béatrice Daille Emmanuel Planas
LS2N - UMR CNRS 6004, Université de Nantes, France
{firas.hmida, emmanuel.morin, beatrice.daille,
emmanuel.planas}@univ-nantes.fr

RÉSUMÉ

En traduction spécialisée, une phase de révision est nécessaire afin de valider les traductions initialement attestées par le traducteur. Cette phase, qui veille à la cohérence du document produit, nécessite la mobilisation d'informations terminologiques accessibles à travers des glossaires et des outils de gestion dédiés. Nous proposons un prototype de concordancier bilingue qui permet de saisir un terme et sa traduction, et fournit des Contextes Riches en Connaissances (CRC) alignés en corpus comparables spécialisés. Les évaluations manuelles et expérimentales menées avec des réviseurs montrent que les CRC bilingues proposés peuvent être perçus comme utiles en complément d'autres ressources d'aide à la traduction, malgré la difficulté de l'exercice.

ABSTRACT

KRCTool : A Bilingual KRC Concordancer for Assisted Revision

In specialized translation process, a revision phase is necessary to validate the initial translation proposed by the translator. This phase, which ensures the consistency of the document produced, requires the preparation of terminological information accessible through glossaries and dedicated management tools. We propose a first prototype of bilingual concordancer that takes as input a term and its translation, and provides not parallel but aligned Knowledge-Rich Contexts (KRC) from specialized comparable corpora. Both the manual evaluation and a real experiment with student revisers show that our concordancer can assist revisers as a complement to their habitual ressources, despite the difficulty of the task.

MOTS-CLÉS : Contextes riches en connaissances, révision en traduction, concordancier bilingue.

KEYWORDS: Knowledge-rich contexts, revision in translation, bilingual concordancer.

Résumé long

Qu'il soit indépendant ou salarié, le traducteur ne peut fournir en permanence des traductions irréprochables. Le manque de contextes lors de la phase de traduction renforce la nécessité de contrôles tels que la révision, qui permet d'assurer la cohérence globale du document produit.

De manière officielle, la révision a été révélée grâce aux normes German DIN 2345, European EN 15038 et ISO 17000 prévoyant l'obligation de faire réviser toute traduction. Selon Robert (2012, p. 95), l'exercice de révision comporte deux axes principaux : la révision bilingue dans laquelle le réviseur compare le texte original avec le texte traduit ; et la révision monolingue où seul le texte traduit est révisé. Ces deux révisions peuvent être effectuées par le traducteur lui-même, afin d'améliorer les traductions qu'il produit, ou comme le recommandent les normes européennes, par un autre traducteur aussi appelé réviseur. Dans cet article nous nous focalisons sur la révision bilingue dans laquelle le réviseur doit confirmer ou infirmer la traduction attestée par le traducteur (Delisle *et al.*, 1999, p. 71).

Dans un contexte de révision, la terminologie est un facteur important. Considérons la traduction du terme anglais $blob$ dans la phrase suivante : « *When the basalt magma first breaks out at the surface, the dissolved gases bubble off vigorously enough to carry* ***blobs*** *of magma into the air with them. The* ***blobs*** *may rise up 2,000 feet or more.* ». Ici, la traduction du terme $blob$ en français n'est pas évidente. Bien qu'en langue générale la traduction commune de $blob$ soit $goutte$, la traduction la plus appropriée est $projection$. Dans ce cas, l'accès à des contextes contenant des voisinages typiques ou renseignant sur les relations entre les termes en question ($blob$ en anglais, ainsi que $goutte$ et $projection$ en français) et les autres termes du domaine, est essentiel pour le réviseur. Ces contextes sont définis comme étant des contextes riches en connaissances (CRC) (Meyer, 2001).

La notion de CRC a été introduite par Meyer (2001) pour désigner les contextes qui illustrent des relations entre les termes d'un domaine spécialisé. Ces relations sont le plus souvent explicitées par des unités lexico-syntaxiques définies comme des patrons de connaissances. Par exemple, « *L'Olympus ci-contre est le volcan géant du système solaire* » est un CRC illustrant le terme $Olympus$. Dans ce CRC, *est le* est un patron de connaissances explicitant une relation hiérarchique entre *Olympus* et *volcan* qui sont deux termes du domaine de la vulcanologie. Même si les CRC ont été introduits dans un cadre terminologique, cette notion fait écho à d'autres types de contextes dans des domaines différents, notamment les « exemples » de Kilgarriff *et al.* (2008). Il s'agit de contextes identifiés, grâce à des collocations à partir d'un corpus monolingue. Ici, nous retenons la définition de Sinclair *et al.* (1970) qui considèrent la collocation comme la co-occurrence régulière de deux items (base, collocatif) dans un contexte donné.

Dans ce travail, nous nous situons dans un cadre de révision dans lequel la traduction est produite en amont. Nous souhaitons fournir au réviseur des CRC bilingues qui permettront de confirmer ou infirmer la traduction choisie. Nous proposons un premier prototype d'outils d'aide à la révision proposant des CRC bilingues en corpus comparables spécialisés : un concordancier bilingue.

Les concordanciers bilingues sont des ressources de plus en plus utilisées pour aider à la traduction terminologique. Ils se basent principalement sur les corpus parallèles dans lesquels les phrases sources sont préalablement alignées avec des phrases cibles. Ces ressources permettent à l'utilisateur de saisir un terme pour consulter ses usages dans les différents contextes retournés par l'outil (voir p. ex Linguee [1]).

1. `www.linguee.com`

Le réviseur qui utilise les concordanciers bilingues doit saisir le terme et sa traduction de façon indépendantes. Le lien entre le terme saisi et sa traduction s'établit à travers les contextes alignés dans le corpus parallèle. Bien que l'intérêt de ces ressources soit indéniable, une limite tient au fait que les corpus parallèles sont rares, notamment en domaines de spécialité. En outre, les contextes proposés par ces outils, sont la plupart du temps généraux et manquent de connaissances spécifiques que l'on peut trouver dans des corpus spécialisés. Pour cela, les corpus comparables, qui sont constitués de textes authentiques ont commencé, ces dernières années, à être exploités comme des ressources d'aide à la traduction (Bowker & Pearson, 2002). Ils sont définis par Bowker & Pearson (2002) comme étant des corpus multilingues qui ne sont pas des traductions à proprement parlé, mais qui partagent certaines caractéristiques telles que la période et le thème. Dans la continuité des recherches de Kilgarriff, SketchEngine [2] permet de saisir un terme source et un terme cible pour fournir des contextes extraits grâce à des collocations. Cet outil repose sur l'exploitation de grands corpus bilingues (comparables et parallèles) de langue générale, pour l'alignement de collocations (Baisa *et al.*, 2014), sans toutefois proposer des contextes sources et sans traiter des corpus spécialisés.

Nous exploitons la comparabilité des corpus comparables pour proposer des CRC bilingues en vue d'aider à la révision en traduction terminologique. Nous proposons un concordancier bilingue fournissant des CRC alignés mais non parallèles : KRCTool.

Bien que les connaissances apportées par les collocations et les patrons de connaissances soient intéressantes en aide à la traduction, l'état de l'art montre un faible rappel des patrons de connaissances au bénéfice de la précision (Morin, 1999), contrairement au collocations qui sont plus productives mais moins précises (Hmida *et al.*, 2015). Il serait alors trop restreint d'aligner les CRC obtenus par les patrons de connaissances. Nous privilégions ceux obtenus par les collocations alignées.

L'idée consiste à proposer tout d'abord pour le couple (terme source/terme cible), des collocations alignées automatiquement à l'aide d'un dictionnaire bilingue de langue générale.

1. **Extraction de CRC monolingues :** nous avons tout d'abord utilisé le z-score (Berry-Rogghe, 1973) pour extraire les collocations où la base est le terme (source ou cible) et le collocatif a pour catégorie syntaxique : adjectif, nom ou verbe. Les phrases contenant les collocations des termes en question sont considérées comme des CRC.
2. **Alignement de CRC :** ces collocations sont ensuite alignées à l'aide d'un dictionnaire bilingue sur la base du collocatif (puisque nous disposons de l'association terme source/terme cible).

   Par exemple, pour la traduction (*lava, lave*), deux collocatifs peuvent être alignés : *basaltic* et *basaltique*. Les deux collocations *basaltic lava* et *lave basaltique* permettent de récolter des CRC sources et cibles non alignés.

Après avoir identifié des CRC sources et cibles en s'appuyant sur des couples de collocations traduites, notre objectif consiste à associer à chaque CRC source des CRC cibles équivalents. Pour cela, nous filtrons tout d'abord les CRC, ensuite nous alignons ceux qui sont retenus.

2. www.sketchengine.co.uk/bilingual-word-sketch/

3. **Filtrage monolingue de CRC :**

   a) longueur du contexte : nous postulons comme Kilgarriff *et al.* (2008) que les phrases courtes ne contiennent pas assez de connaissances autres que la collocation en question. D'autre part, celles qui sont très longues sont difficiles à consulter et risquent d'illustrer des informations inutiles pour la révision. Seulement les phrases contenant entre 10 et 20 mots sont retenues.

   b) présence de pronoms : dans leur travaux, Kilgarriff *et al.* (2008) pénalisent les phrases contenant des anaphores pronominales, puisqu'elles présentent un facteur d'ambiguïté. En particulier, les pronoms en début de phrase posent un problème de référence alors que ceux qui apparaissent au milieu sont moins problématiques car ils peuvent référer à des noms de la même phrase. Ici, nous choisissons d'éliminer les contextes commençant par un pronom.

   c) phrases affirmatives : Kilgarriff *et al.* (2008) considèrent les phrases interrogatives comme non intéressantes et privilégient les phrases affirmatives. Nous retenons aussi ce critère.

   d) complexité du contexte : ce critère qui renseigne sur la lisibilité de la phrase a été également pris en compte par Didakowski *et al.* (2012). Nous suivons la même stratégie en utilisant un analyseur syntaxique pour éliminer les phrases complexes.

4. **Alignement des CRC retenus :** en sus des collocations, nous alignons les CRC en exploitant d'autres points d'ancrage qui représentent des ponts de transition d'une langue à l'autre :

   a) nombre de cognats : les cognats représentent des ponts de transition aisément repérés par le lecteur dans les couples de contextes (source et cible). Les contextes partageant au moins un cognat seront alignés.

   b) nombre de termes simples traduits : bien qu'elles soient rares dans le corpus, les phrases contenant des termes traduits sont exceptionnellement utiles pour le réviseur. Les termes simples des corpus étudiés ont été préalablement extraits par un outil d'extraction terminologique. Les contextes contenant au moins un terme et sa traduction seront alignés.

   Cette étape d'alignement permettra de proposer pour la traduction (*lava, lave*), les deux contextes *« Shield cones are broad, slightly domed volcanoes built primarily of fluid,* ***basaltic lava*** *»*, et *« Volcan bouclier, volcan de forme ovale, très aplati, dû à l accumulation de coulées de* ***lave basaltique*** *fluide »*

Afin de vérifier si KRCTool peut réellement aider les réviseurs, nous avons élaboré le protocole suivant. Nous avons deux groupes A et B d'étudiants en Master 1 de traduction. Nous avons partagé chaque groupe en deux sous-groupes et avons demandé à chacun d'entre eux de travailler sur une partie différente d'un texte préalablement traduit par des apprentis traducteurs. Cela vise à empêcher et affiner toute spécificité de ces parties de texte qui pourrait influencer l'exercice de révision. Dans un premier temps, les étudiants A révisent le texte traduit en utilisant leurs ressources habituelles : l'objectif est de corriger le mieux possible le texte afin d'obtenir une traduction correcte. Dans un second temps, les mêmes étudiants A doivent

corriger le texte traduit par le seul moyen de KRCTool. Les étudiants B effectuent le même exercice, mais en utilisant tout d'abord KRCTool dans la première phase. En phase 2, ils utilisent leurs ressources habituelles. Les premières évaluations manuelles et expériences menées auprès de réviseurs ont montré que même si leur apport n'est pas manifeste, les CRC bilingues proposés sont une piste prometteuse dans un exercice d'aide à la révision, malgré la difficulté de celui-ci.

Références

BAISA V., JAKUBÍČEK M., KILGARRIFF A., KOVÁŘ V. & RYCHLÝ P. (2014). Bilingual word sketches : the translate button. In A. ABEL, C. VETTORI & N. RALLI, Eds., *Proceedings of the 16th EURALEX International Congress*, p. 505–513, Bolzano, Italy : EURAC research.

BERRY-ROGGHE G. (1973). The computation of collocations and their relevance in lexical studies. In A. AITKEN, R. BAILEY & N. HAMILTON-SMITH, Eds., *The Computer and Literary Studies*, p. 103–112. Edinburgh : Edinburgh University Press.

BOWKER L. & PEARSON J. (2002). *Working with specialized language : a practical guide to using corpora.* Routledge.

DELISLE J., LEE-JAHNKE H. & CORMIER M. C. (1999). *Terminologie de la Traduction : Translation Terminology. Terminología de la Traducción. Terminologie der Übersetzung*, volume 1. John Benjamins Publishing.

DIDAKOWSKI J., LEMNITZER L. & GEYKEN A. (2012). Automatic example sentence extraction for a contemporary German dictionary. In *Proceedings of the 15th EURALEX International Congress*, p. 343–349, Oslo, Norway.

HMIDA F., MORIN E. & DAILLE B. (2015). Extraction de Contextes Riches en Connaissances en corpus spécialisés. In *Actes de la 22^e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 425–431, Caen, France : Association pour le Traitement Automatique des Langues.

KILGARRIFF A., RYCHLÝ P., HUSÁK M., RUNDELL M. & MCADAM K. (2008). GDEX : Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, p. 425–432, Barcelona, Spain.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 279–302. John Benjamins Publishing Company.

MORIN E. (1999). Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues*, **40**(1), 143–166.

ROBERT I. S. (2012). *La révision en traduction : les procédures de révision et leur impact sur le produit et le processus de révision*. PhD thesis, University of Antwerp.

SINCLAIR J. M., JONES S. & DALEY R. (1970). *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08.* Department of English.