A Bilingual KRC Concordancer for Assisted Translation Revision based on Specialized Comparable Corpora

Firas Hmida, Emmanuel Morin, Béatrice Daille, Emmanuel Planas

A Bilingual KRC Concordancer for Assisted Translation Revision based on Specialized Comparable Corpora

Firas Hmida, Emmanuel Morin, Béatrice Daille, and Emmanuel Planas

LINA-UMR CNRS 6241
University of Nantes, France
firstname.lastname@univ-nantes.fr

Abstract. Terminology is used all through the process of specialized translation. Indeed, many translators confirm that an error on terminology has a major impact on their work. Thus, a revision phase is necessary to validate the initial translation proposed by the translator. This paper deals with the assisted terminological revision in specialized translation from English to French. We propose a new generation of bilingual concordancers that takes as input a term and its translation, and provides not parallel but aligned Knowledge-Rich Contexts from specialized comparable corpora. Both the manual evaluation and a real experiment with student revisers show that our concordancer actually assists revisers despite the difficulty of the task.

Keywords: bilingual concordancer, Knowledge-Rich Contexts, specialized comparable corpus, collocations, revision, human translation

1 Introduction

In a survey conducted in Morin-Hernandez (2009, p. 143), 90% of the French translation professionals respond that an error on terminology has a major impact on translation work. Terminology is indeed crucial all through the translation process. Gouadec (2002) identifies three main steps in the translation process of specialized texts (specialized translation): pre-translation, translation and post-translation. The translation phase is itself divided into two classical sub-tasks conducted by translators: a translation task and a revision task.

Robert (2012, p. 95) identifies two main types of revision: the bilingual revision where the reviser carefully compares the original text (the source text written in the source language) and its translation (the target text written in the target language); and the monolingual revision where the translation is only revised in the target text. Both revisions can be conducted by the translator himself in a quest of a better production; or by a different translator called the reviser. The translation industry standards (German DIN 2345, European EN 15038, ISO 17100) imply the obligation for professional translators to review every translation by a third party translator or reviser.

In this paper, we are specifically concentrating on the bilingual revision where the reviser has to check different aspects of the first specialized translation draft (Delisle et al., 1999, p. 71). Thus, terminology comes as an important factor. To concretely illustrate the point, let us consider the translation of the term *blob* in the following text: *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.*

Here, the translation of the term *blob* into French, in the field of volcanology, is not obvious. While in the general language the common translation of *blob* is *goutte* (drop, a scrap of something), a more suited translation is *projection* (spatter, splash). In this case, it is essential for the reviser to get access to textual contexts containing typical neighborhoods or providing useful information about the links between the terms involved in this translation (*blob* in the source language, and the translator's choice in the target language, either *projection* or *goutte*) and the other terms and expressions of the field. These contexts are defined as Knowledge-Rich Contexts (KRCs) (Meyer, 2001). In the Cristal[1] project, we concluded with a list of attested KRCs that we automatically extract from prepared comparable corpora[2].

In this work, we aim at assisting revisers in the bilingual revision task by providing them with KRCs that will help them confirm or disapprove the translation that was already proposed. We will more precisely provide revisers with both source (EN) and target (FR) KRCs extracted from specialized comparable corpora, in a new generation of bilingual concordancer that we call KRCTool. We will show that this tool actually helps revisers in the framework of specialized revision.

2 Framework

We define at first the KRC concept, then we present the issue of classical bilingual concordancers in a revision framework.

2.1 Knowledge-Rich Contexts

Meyer (2001) introduces the notion of Knowledge-Rich Context to describe contexts that contain terms and relations between them in a specialized domain. These relations are usually expressed with lexical and syntactic patterns (Morin, 1999). For example, *An impact crater is caused by two celestial bodies impacting each other* is a KRC of the term *impact crater*, in which *is caused by* is a pattern reflecting a causality relation between *impact crater* and *celestial bodies*. All of these terms are from the domain of volcanology. KRCs have historically been introduced in the framework of terminology and knowledge extraction purposes.

[1] `http://www.agence-nationale-recherche.fr/?Projet=ANR-12-CORD-0020`

[2] Corpora that contain multilingual documents that are not translations of but share characteristics such as period and theme (Bowker and Pearson, 2002).

We consider that this notion refers also to other types of contexts, like the "examples" of Kilgarriff et al. (2008). These examples are contexts identified thanks to collocations extracted from a general monolingual corpus. A Collocation is a regular co-occurrence of two items (base, collocate) within a specified field (Sinclair et al., 1970). A good command of collocations is an essential component of the proficiency of any language or specific discourse. Indeed, it is more correct to say *to prescribe medication* than *to write medication* in medical domain, or *to gush lava* instead of *to push lava* in volcanology. In these examples, *medication* and *lava* are the bases. Based on collocations, Kilgarriff's examples are undoubtedly considered as rich of knowledge since they illustrate typical neighborhoods in contexts. This knowledge is well appreciated by revisers. Planas et al. (2014) already showed that KRCs, based on collocations or relations between terms, can be useful to illustrate terms in specialized domain. Thus, we focus here on KRCs containing collocations of the source term or its proposed translation.

2.2 Bilingual Concordancers

Bilingual concordancers are resources more and more used to assist translators in terminological translation tasks. They often rely on parallel corpora These tools allow translators to enter one term and, if this term occurs in the bilingual parallel corpus, to look how it was dealt with across the different contexts the tool returns. Perhaps one of the more popular concordancer among translators is the online service Linguee[3], actually built from aligned parallel corpora.

In bilingual revision, the reviser who uses these kind of tools, that take only one term as input, has to enter the source or target terms independently. The link between source term (resp. the target term) and the term used as translation comes from the fact that contexts sentences are aligned in the parallel corpora. Despite their general usefulness, the main problem of classic concordancers is the scarcity of parallel corpora, especially in specialized domain. Furthermore, contexts proposed by Linguee are generally quite broad and lack specific knowledge that could be found in specialized corpora. A special use of SketchEngine[4], the "bilingual word-sketch", allows the input of the couple (source term, target term) and provides a series of available collocations from which some context can be retrieved. These use large corpora (parallel and comparable) in general domain, and different alignment schemata where the compositional term alignment is used (Baisa et al., 2014). The multilingual sentence alignment from comparable corpora drew much research attention. Rauf and Schwenk (2011) shows that parallel sentences are quite scarce in comparable corpora, especially in specialized domain.

In this paper, we rely on the comparability of comparable corpora collected from specialized texts. We build a bilingual concordancer called KRCTool that provides not parallel but aligned KRCs to help in revising a pair (source term/ proposed translation) given as input.

[3] `http://www.linguee.fr/`

[4] `https://www.sketchengine.co.uk/bilingual-word-sketch/`

3 Method

In the comparable corpora, parallel or lexically similar contexts are rare. It would be even more restricted to align KRCs on the base of their lexicon. Consequently, our aim is to determine bilingual "properties" which enable the reviser to build transition bridges between source and target contexts. We then allocate to each source KRC an equivalent target KRC based on these properties. We propose a methodology first based on **extraction of KRCs**: for each (source term/proposed translation), we extract the collocations of the source term and its proposed translation and then retain the sentences that contain the automatically translated collocates. These sentences are considered as KRCs. And after based on **alignment of KRCs**: the bilingual sentences resulting from the previous step will be filtered and aligned.

3.1 KRC Extraction

Mammino (1995) approached the issue of specialized terms and their use, that are faced by translators without in-depth knowledge of the terminology. In this case, a translation that does not respect the standard collocations of the domain may be negatively perceived by revisers (Musacchio and Palumbo, 2008). Revisers frequently look for approximations of the source collocation, in the target language. If the literal translation is correct, it would be unwise to try at all costs to avoid it, because it may allow referential and pragmatic equivalences (Newmark, 1988, p. 68-96). Here, our purpose is not to translate collocations, but to provide relatively close collocations, that can help revisers check if the proposed translation is in its typical context.

First, we implement the z-score to automatically extract collocations according to their syntactic structures: (T, Adj), (T, N) and (T, V), with T the single term we want to illustrate. Then, we align collocations, pairing collocates belonging to the same grammatical category. Even if the overlap between collocations and multi-word terms is a well-known problem in collocation extraction, here, we do not distinguish between these two phenomena that may share co-occurrence and syntactic criteria.

3.2 KRC Alignment

The obtained KRCs at this stage are aligned only on the basis of collocations, which often prove to be insufficient. Therefore, we will refine the KRC alignment using other anchor points in addition to collocations. Our goal now is to filter and align them:

1. **filtering criteria:**
   - **context length:** short sentences could not contain more knowledge than the collocation. Conversely, it is very difficult to consult sentences that are very long, also they may illustrate irrelevant information for the revision. As Kilgarriff et al. (2008) we retain only sentences containing between 10 and 20 full words.

- **pronouns:** Kilgarriff et al. (2008) penalize contexts that contain pronominal anaphora, since it may refer to text unities in previous sentences. We assume that pronouns inside contexts are less problematic because they can refer to unities in the same sentence. We eliminate only contexts starting with a pronoun.
- **affirmative contexts:** Kilgarriff et al. (2008) prefer affirmative sentences rather than interrogative ones. We also retain this criterion to filter out interrogative contexts.
- **context complexity:** this criterion was also addressed by Didakowski et al. (2012) to measure the readability of the sentence. We follow the same strategy using a dependency parser to filter complex contexts. In our case, we use the sum of the scores of all possible parse trees for a given sentence to measure the complexity: the more complex the context is, the greater is the sum of all its possible trees.

2. **alignment criteria:**
   - **number of cognates:** we consider cognates as two words starting with the same 4 characters as Léon (2008). They represent transition bridges easily detected by the reader, in pairs of source and target contexts. Contexts sharing at least one cognate, will be aligned.
   - **number of translated simple terms:** despite their scarcity in the corpus, sentences containing translated terms are exceptionally operational for the reviewer. The single word terms of the studied corpus were extracted by a dedicated terminological tool. Contexts containing at least one simple term and its translation will be aligned.

4 Manual Evaluation

To evaluate the quality of the aligned KRCs, we manually prepared reference KRCs for each studied term and its translation. In this section, we present the used corpora, the reference data and the experiments.

4.1 Corpora and Bilingual Dictionary

This evaluation was carried out on specialized comparable corpus built by Josselin-Leray (2005) and obtained through a thematic research from newspapers and magazines in the field of volcanology. This corpus is composed of English and French scientific documents containing roughly 400,000 words per language. They have been cleaned and standardized through TermSuite[5] that also extracts terminology. For the automatic alignment of collocations, we used ELRA[6], a bilingual dictionary of general language (EN-FR) containing 145,542 entries. It also contains the POSs of entries.

[5] `https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite`
[6] `http://catalog.elra.info/product_info.php?products_id=666`

4.2 Evaluation Data

Bilingual Aligned KRCs were manually prepared for 15 pairs of single-word terms essential for the volcanology domain. Here are some examples: *basalt/basalte*, *cinder/scorie*, *volcan/volcan*, *eruption/éruption*... The multi-word terms have been excluded for the reason that the identification of complex terms collocations can be treated as a separate issue that we do not regard in this work. The process that we followed to prepare the reference KRCs was:

1. For each pair of terms, we manually identify the source and target collocations in which collocates are translations. Then, we extract contexts that contain these collocations. Here, experts were solicited to check the manual translation.
2. We checked manually if contexts provided for each collocation were valid. A context is valid only if the collocation in question is valid within it.

4.3 Experimentation

We applied our method on the 15 pair of terms and we evaluate the bilingual KRCs aligned with and without filters. The aligned KRC pairs were manually validated if at least one of the following conditions is valid:

1. the alignment criteria are also valid within a window of 7 words (approximately) containing the term in question or its proposed translation. For example:
   - pair of translation: *lava, lave*
   - aligned collocations: *(lava, basaltic)* and *(lave, basaltique)*
   - source KRC : *Shield cones are broad, slightly domed volcanoes built primarily of fluid,* ***basaltic lava****.*
   - target KRC: *Volcan bouclier, volcan de forme ovale, très aplati, dû à l'accumulation de coulées de* ***lave basaltique*** *fluide.*

   Here, the concentration of the alignment criteria within a window of words that can be easily consulted, help to validate the pair of the aligned KRCs.
2. the "global topics" of the two KRCs are similar. The alignment criteria, which are mainly lexical, could be non relevant towards the reviewer. In this case, if the topics of the contexts in question are similar, they can be considered as a bridge transition between the contexts. In the following example, KRCs have been validated thanks to the similarity of the subjects that they treat:
   - pair of translation: *cinder, scorie*
   - aligned collocations: *(cinder, incandescent)* and *(scorie, incandescent)*
   - source KRC: *Strombolian eruptions are named for Stromboli volcano off the west coast of Italy, where a typical eruption consist of the rhythmic ejection of* ***incandescent cinder****, lapilli, and bombs to heights of a few tens or hundreds of feet meters.*
   - target KRC: *Le dynamisme strombolien s'exprime par des explosions rythmiques qui projettent des blocs et des* ***scories incandescents****.*

Table 1. Evaluation of aligned KRCs: with and without filters

Corpora	# terms	# aligned terms	# pairs of aligned coll.	# contexts	# pairs of aligned KRCs	P. valid pairs of aligned KRCs
without filters						
Vulcano EN 15 Vulcano FR 14	10	23	677 665	309	43,04%	
with filters						
Vulcano EN 15 Vulcano FR 14	10	16	241 296	157	61%	

4.4 Results

The analysis of table 1 shows that the aligned collocations are productive: each collocation pair produces on average 28 contexts without filter, and 15 with filter, for each language. We note that even if the application of filters deteriorate the number of aligned KRCs, it significantly improves the precision of the alignment criteria since it moves from 43% to 61%. We could not provide bilingual contexts for five pairs of terms. Some of these pairs have a too small number of extracted collocations or only one syntactic structure. For the others, the alignment method act as a filter and eliminates contexts in both languages.

5 Experiment with Revisers

After having studied the quality of bilingual KRCs, we perform real experiment with student revisers using the KRCTool.

5.1 Experimental Data

We had conducted an experiment in a previous framework where 11 second year Master students translated the same text from English to French. For our current experiment, we used the same English text, and selected one of the student translations for the revision task. We retain one of the most perfectible ones. We identified three terms in the source text; and changed the translation of these items with more common translations in the target text.We then expected the revisers to correct these "lazy" translations by terms more specific to the domain of volcanology, with the use of the KRCTool. Table 2 contains the source and the translation terms that we changed, with acceptable translations.

Table 2. Source and changed terms

source term	modified translation	correct translations
cinder	débris	scorie, cendre
vesicle	poche	vacuole, vésicule
blob	boule	paquet, projection

Here is a detailed view of the reasoning we expected. When the couple *blob* and *goutte* is searched in KRCTool, only one target KRC containing *goutte* is shown. Nevertheless, this KRC shows a use of *goutte* which is restricted to an in-vitro experiment, that does not fit with *blob* here. A good reviser should here disapprove *goutte* and search for an alternative solution.

Instead, if *blob* and *projection* are searched in the KRCTool, as suggested by the available translation, instances of *projection de lave* are displayed along with *blob of magma*. This provides a more acceptable translation for *blob*.

5.2 Protocol

In order to test whether the KRCTool would help the revisers or not, we designed the following protocol. We had two groups A and B of first year students from a Master in Professional Translation. We divided each group into two sub-groups and asked each sub-group to work on a different part of the text, as sums-up table 3. This was done to prevent and smoothen any specificity of these text parts that may influence the revision task. In a first phase, students A had to revise the translation text with their usual resources like Linguee, Le Grand Dictionnaire Terminologique or CRISCO (synonyms): the objective was to correct as best as possible the text so as to get a good translation. In a second phase, the same students A had to correct the translated text only using KRCTool. Students B did the same task, but started in Phase 1 with the use of the KRCTool first. In Phase 2, they made use of their usual resources.

Table 3. Group repartition

Group A	Text 1	Text 2	Time (min)
Phase 1:common res.	Aa	Ab	20
Phase2:KRCTool	Ab	Aa	20
Group B	Text 1	Text 2	
Phase 1:KRCTool	Ba	Bb	20
Phase2:common res.	Bb	Bb	20

5.3 Results

Based on table 4, KRCTool proved to be useful for correcting the translation of the three terms. For each term, a revised translation was provided by 1 to 4 students (out of 14) with the use of KRCTool. All revised translation were correct. In an post survey, students declared that the KRCTool provided them with specific and specialized contexts that they did not find in their usual resources. We see that group B provided more corrections using the KRCTool that group A. We believe this is because group B started in Phase 1 by using the KRCTool. Whereas Group A first used common resources in Phase 1, and then used the KRCTool only in Phase 2: hence, most of the terminology searches for group

Table 4. Revision results. (Nb: number of performed revision; x: performed revision; possible translations provided by KRCTool for *cinder*: *scorie*, *cendre*, *débris*; for *blob*: *projection*, *paquet* and *boule*; for *vesicle*: *vésicule*, *vacuole* and *poche*; for *bubble off*: *partent*; and for *spewed out*: *sort*).

Term	Nb	Aa1	Aa2	Ab6	Ab7	Ab8	Ba1	Ba2	Ba3	Ba4	Bb6	Bb7	Bb8	Bb9	Bb10
With Common resources															
cinder	6	-	-	-	-	x	-	x	x	-	-	x	x	x	-
blobs	3	-	-	-	-	-	-	-	x	x	-	-	-	x	-
vesicles	4	x	x	-	-	-	-	x	x	-	-	-	-	-	-
Total (T1)		1	1	-	-	1	-	2	3	1	-	1	1	2	-
With KRCTool															
cinder	2	-	-	-	-	x	-	x	-	-	-	-	-	-	-
blobs	4	-	-	-	-	-	-	-	-	x	-	x	x	x	-
vesicles	1	-	-	-	-	-	-	-	-	-	-	x	-	-	-
Total (T2)		-	-	-	-	1	-	1	-	1	-	2	1	1	-
T2 $\geq$ T1 $\geq$1		**-**	**-**	**-**	**-**	**x**	**-**	**-**	**-**	**x**	**-**	**x**	**x**	**-**	**-**
1 < T1 < T2		**x**	**x**	**-**	**-**	**-**	**-**	**x**	**x**	**-**	**-**	**-**	**-**	**x**	**-**

A were processed in Phase 1 using common resources: there was less searches left for KRCTool. Two students (Ba1 and Bb7) provided more corrections with the KRCTool than with other common resources. Table 4 also shows that using the KRCTool, four students among the 13 ones which carried out corrections have successfully accomplished the same revision as with common resources, or better. However, five students performed a better revision based on common tools. We have to admit that these students were only first year Master and did not have previous knowledge of this specialized domain to correct all the terms as a professional reviser would. In average, students provided more corrections with common resources that provide more output. Debutant students tend to be seduced by the quantity rather than the quality of the resources.

6 Conclusion

This paper proposes KRCTool as an example of a new generation of bilingual concordancers that takes as input a source and a target term and provides aligned KRCs from specialized comparable corpora, for an assisted revision purpose. KRCTool is based on a methodology that uses collocations, cognates and the translation of simple terms as anchor points for the identification and the alignment of KRCs in specialized comparable corpora. The manual evaluation shows that the KRCs we obtain are quite acceptable for a manual revision. The experiment performed with revisers confirms indeed that KRCs proposed by the KRCTool actually assist revisers in a translation revision task. The study we carried out deals with qualitative aspects of the obtained KRCs that we wanted to completely control. That is why our experiments relied on few terms. Further experiment should be driven for confirming our findings.

Bibliography

Baisa, V., M. Jakubek, A. Kilgarriff, V. Kov, and P. Rychl (2014). Bilingual word sketches: the translate button. In *EURALEX*, Bolzano, Italy, pp. 505–513.

Bowker, L. and J. Pearson (2002). *Working with specialized language: a practical guide to using corpora.* Routledge.

Delisle, J., H. Lee-Jahnke, and M. C. Cormier (1999). *Terminologie de la Traduction: Translation Terminology.* John Benjamins Publishing.

Didakowski, J., L. Lemnitzer, and A. Geyken (2012). Automatic example sentence extraction for a contemporary German dictionary. In *EURALEX*, Oslo, Norway, pp. 343–349.

Gouadec, D. (2002). *Profession: traducteur.* La Maison du dictionnaire.

Josselin-Leray, A. (2005). *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues: étude d'un domaine de spécialité: volcanologie.* Ph. D. thesis, Université de Lyon 2.

Kilgarriff, A., P. Rychlý, M. Husák, M. Rundell, and K. McAdam (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *EURALEX*, Barcelona, Spain, pp. 425–432.

Léon, S. (2008). *Acquisition automatique de traductions d'unités lexicales complexes à partir du Web.* Ph. D. thesis, Université de Provence.

Mammino, L. (1995). Il linguaggio e la scienza. *Torino: Societá Editrice Internazionale*.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, pp. 279–302. John Benjamins Publishing Company.

Morin, E. (1999). Using lexico-syntactic patterns to extract semantic relations between terms from technical corpus. In *TKE*, Innsbruck, pp. 268–278.

Morin-Hernandez, K. (2009). *Revision as a key function of translation quality management in a professional context.* Ph. D. thesis, Université Rennes 2.

Musacchio, M. T. and G. Palumbo (2008). Shades of Grey: A Corpus-driven Analysis of LSP Phraseology for Translation Purposes. In *Corpora for University Language Teachers*, pp. 69–79. Bern: Peter Lang.

Newmark, P. (1988). *A Textbook of Translation.* Prentice-Hall International.

Planas, E., A. Picton, and A. Josselin-Leray (2014). Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol. In *TKE*, Berlin, Germany, pp. 188–228.

Rauf, S. A. and H. Schwenk (2011). Parallel sentence generation from comparable corpora for improved smt. *Machine translation 25*(4), 341–375.

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

Sinclair, J. M., S. Jones, and R. Daley (1970). *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08.* Department of English.