# Aligned Knowledge-Rich Contexts from Specialized Comparable Corpora

Firas Hmida, Emmanuel Morin, Béatrice Daille

# Aligned Knowledge-Rich Contexts from Specialized Comparable Corpora

Firas Hmida, Emmanuel Morin, and Béatrice Daille

University of Nantes
LINA, 2 Rue de la Houssinière
BP 92208, 44322 Nantes, France
{firas.hmida,emmanuel.morin,beatrice.daille}@univ-nantes.fr

**Abstract.** During the 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. In this work, we propose a methodology to build a bilingual concordancer providing not parallel context but aligned Rich-Knowledge Contexts (KRCs) from specialized comparable corpora. These contexts share bilingual conserved properties between the source and target language within the comparable corpus. KRCs are intended to assist in verifying usage of the term to be translated and its proposed translation. The assessment of the tool that we propose shows that the obtained KRCs are acceptable in order to help the terminological revision.

**Keywords:** Knowledge-Rich Contexts, collocations, specialized corpus, revision in translation, bilingual concordancer

## 1 Introduction

Specialized translation work requires the preparation and the mobilization of terminological information accessible through glossaries and dedicated management tools. This work needs intrinsic linguistic skills and implements writing and verification skills that can sometimes be performed as a full-time work. As a result, many translators are also terminologists, project managers or reviewers.

According to Gouadec [4], the translation process is a serie of translating activities that can be grouped into three phases: pre-translation, translation and post-translation. The pre-translation phase includes all of the preparations of files and resources needed for the translation. The translation phase is carried out by translators or redactors specially trained to produce notices or guide uses for examples. The post-translation phase mainly includes the reintegration of translated strings in the original format, as well as the control of the integrity of the built document [3]. Beyond the translation of a text by a qualified professional, it is becoming increasingly common to perform a quality control of

the translated text. Everyone recognizes that a translator, whether independent or employed, can not permanently provide faultless translations. The lack of contexts during the translation phase reinforces the need for controls such as revision, which ensures the overall consistency of the produced document.

Revision involves the careful examination of translation and its compliance with quality requirements through specific corrections and improvements. It forms an integral part of the execution process of translation services. Applied on semi-finished translations, revision participates in the implementation of a global strategy of quality assurance. In 2006, the revision has been revealed through the publication of the European standard EN 15038 "Translation Service: requirements for service delivery". This norm implies the obligation to review every translation by a third party translator or reviewer. Despite its common practice among professional translators, revision remains a virtually unexplored field in research.

For its part, the reviewer has many means to ensure as best as possible the quality of the translation to which he is committed. These means are tools and resources such as concordancers and corpora (comparable and parallel). In order to achieve his objectives, the reviewer has two methodological approaches to examine translation: the revision of the single translated document (monolingual revision); or the revision of the translation by comparing it to the original document (bilingual revision). The choice of the revision method is often a compromise between the complexity and the difficulty of the translation, on one hand, and the available resources (time and staff) on the other hand. However, the reviewer does not stick only to the monolingual revision: the original document can be used as backing if in doubt. For example, the translation of the word *cinder* to French in volcanology can be problematic for the translator. Indeed, two acceptable translations are possible: *scorie* and *cendre*. In general language, the translation of *cinder* is *cendre*, which is an old name for *scorie*. However, *cendre* can also correspond to the translation of *ash*. Here, the revision phase (or a self-revision) is required to make a decision. In this case, it is essential for the review to have access to contexts containing typical neighborhoods or providing information on the conceptual relations between the terms in question (*cinder* and *ash* in the source language, and *scorie* and *cendre* in the target language) and the other terms of the domain. These contexts are defined as Knowledge-Rich Contexts (KRCs) [9].

In this work, we focus on assisting reviewers in translation revision task. The translations (terms/proposed translations) were initially proposed by the translator in a production step. Our aim is to suggest KRCs for the reviewer to confirm or disprove the chosen translation. First, we present the concept of the KRC as well as bilingual concordancers which help to validate proposed translations. Then, we detail the methodology to build a bilingual concordancer from specialized comparable corpora.

## 2 Framework

We start by defining the KRCs, then we present bilingual concordancers that are operated in a revision framework.

### 2.1 Knowledge-Rich Contexts and Lexicographic Examples

The notion of Knowledge-Rich Context (KRC) was introduced by Meyer [9] to describe contexts that contain terms and relations between them in a specialized domain. These relations are often expressed with lexical and syntactic elements named knowledge patterns. For example, *Cinders are glassy particles that contain vesicles* is a KRC of the term *cinder*. In this KRC *are* is a knowledge pattern reflecting a hierarchical relation between *cinder* and *particle* which are terms from the domain of volcanology. Although the undeniable interest of knowledge patterns to identify KRCs, one of the major difficulties is the fact that their identification is costly in terms of time and stuff [8]. Moreover, there is no library of knowledge patterns that would manifest the cumulative aspect of the research in this field. Each new study, should repeat the synthesis of the existing studies to identify lists of knowledge patterns.

Even if KRCs have been introduced in terminological perspectives, this notion refers also to other types of contexts in different fields, especially the "examples" of Kilgarriff & *al*. [5] These examples are contexts initially identified thanks to collocations extracted from a general monolingual corpus. Then, they are filtered with ranking criteria in order to propose the top 20. A good command of collocations is an essential component of the proficiency of any language or specific discourse. This explains the importance allocated to this notion for language generation. Collocations are "transparent" lexical associations, in terms of comprehension, that a non-native speaker must learn to control. For example, *to prescribe medication* in medical domain, or *to gush lava* in volcanology. A collocation is a pair (base, collocate). In the previous examples, *medication* and *lava* are the bases. The notion of collocations has various definitions according to the research framework in which it is employed. Here, we use the definition of Sinclair [17] that refers to statistical collocation:

> "*[...] the occurrence of two items in a context within a specified environment. Significant collocation is a regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict.*" [17, p. 150]

### 2.2 Bilingual Concordancer

Bilingual concordancers are terminological resources mainly used to assist translators in terminological translation tasks, often using parallel corpora. These resources enable translators to consult and use texts aligned in advance.

In a translation task, the usefulness of bilingual concordancers consists of providing the translator with all occurrences of a term to translate, called *head*, within a defined window in the source language. For each source segment, an equivalent target one is also shown, since it may contain the translation of the desired term (*i.e* the head). For example, for the term *crater*, the proposed source segment (*i.e* concordance) is "*... point of a large mine, a* ***crater*** *will be formed...*". This source concordance is aligned with "*... l'explosion d'une grosse mine, un cratère se formera...*". Here, the translator has to identify the "good" translation of the term *crater* within the proposed target concordance in order to choose the best translation.

In bilingual revision, also known as comparative revision, the reviewer performs back and forth checks between source and target resources in an iterative process, to verify if the translation is valid in the target language. Here, he estimates what will be the transition bridges to pass from one language to the other by consulting source and target concordances. Resources are consulted in a bilingual way. Despite their general usefulness, the main problem of classic concordancers is the scarcity of parallel corpora, especially in specialized domain.

The context of this study is the bilingual revision of translations. We rely on the comparability of comparable corpora in order to review a translation produced by the translator. We build a bilingual concordancer providing for a pair (source term/proposed translation) "aligned" KRCs from comparable corpora. These KRCs will be bilingually exploited to confirm the proposed translation. Although the interest of collocations and knowledge patterns in a revision perspective, the state of the art [10] shows a low recall of knowledge patterns, in favor of their precision. Therefore, it would be more restricted to align KRCs obtained by knowledge patterns. We focus on KRCs provided by collocations.

We propose a methodology to build a bilingual concordancer based on two steps:

1. alignment of collocations: we use the alignment of collocations rather than sentences (actually, it is unlikely to find sentences that match translation in specialized comparable corpora). Here, collocations are treated as the only anchor point ensuring both the identification of bilingual KRCs, and the transition from the source to the target context.
2. alignment of KRCs: retained collocations in the previous step provide KRCs. For a given translation (source term/proposed translation) we associate to each source KRC an equivalent target KRC. The process that we follow is to first filter contexts in a monolingual way, in order to align sources with target contexts.

## 3 Alignment of Collocations

In this section, we describe our assumptions of the collocation alignment and we discuss their implementation. The aligned collocations will provide KRCs in both source and target languages.

### 3.1 Bilingual Collocations

Our aim is to determine invariant "properties" between the two languages, which enable the reviewer to build transition bridges between source and target contexts. Thus, based on these contexts, the translator can check the translation in question. Concretely, in the translation process, the literal translation is considered as an initial bridge between the original (*i.e* source) and the final (*i.e* target) text. This intermediate step allows for the disambiguation of text segments which are linguistically or cognitively complex. Primarily, specialized terms and their use are the major problem faced by the translator who does not have in-depth knowledge of the terminology. This problem has also been tackled by [7]:

> "*Very often, use is normalized. We employ specific terms for very specific aspect and phenomena, so that the use becomes almost the rule. [...] Opposing the use often means introducing errors.*"[1] [7, p. 7-9]

Indeed, the available terminological tools are not able to indicate how to use all of the terms in their contexts. In this case, the translation that does not respect the standard collocations of the domain may be negatively perceived by the experts [12]. Since we already had the pair (source term/proposed translation), it is necessary to check the validity of the translation based on the typical neighborhoods of the terms in question, namely their collocations. Especially, collocations are considered as an "index" of linguistic richness marking the contexts in which they appear. At first, we use the alignment of collocates rather than the alignment of sentences. Here, we consider the collocates as an anchor point preserved in an interlingual way for a given pair (source term/produced translation).

**Assumption 1:** *"the collocates of a term and its translation are preserved between the source and the target languages"*

### 3.2 Collocations and Literal Translations

Translators used to look for approximations of the source collocation, in the target language. The literal translation is by far the most frequent practice, since it is the first intuition of the translator. 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 [13, p. 68-96]. The literal translation represents the standard practice in specialized translation. Usually, the more technical the text is, the more literal the translation will be [14, p. 64].

In general language, the alignment of collocations may be problematic because of their particular semantics. Indeed, they can not be translated literally only on the basis of the lexical units that compose them. For example, the word-by-word translation of the collocation *fierce battle* (EN) gives *bataille féroce* (FR)

[1] "*Molto spesso l'uso standardizzato: si impiegano ben precisi termini per ben specifici aspetti o fenomeni, al punto che l'uso diviene quasi la regola. [...] opporsi all'uso significa spesso introdurre errori.*"

instead of *bataille acharnée* which is the valid translation. Here, our purpose is not to translate collocations but to provide relatively close collocations in source and target languages. These pairs of collocations are supposed to be acceptable for human revision. They help reviewers ensure that the proposed translation is "obvious" in its typical context. Even if the literally translated collocation does not correspond to the best translation, we assume that it will allow the reviewer to be closer to the meaning and use of the desired term based on different contexts. Thus, thanks to the aligned collocations, the reviewer would be able to ensure the consistency of the translation produced by the translator.

**Assumption 2:** *"the literal translation of a source collocate is a target collocate which allows reviewer to be closer to the mean and use of the target term in question"*

### 3.3 Alignment of Collocates

Many studies, such as [16], were focused on the automatic identification of equivalent collocations from comparable corpora. These works show that, in practice, the two words composing a source collocation are themselves a context. The latter facilitates the extraction of the similar associations within a comparable corpora, by filtering similar terms in the corpus on the basis of their grammatical classes. Here, we rely on part-of-speech (POS) tagging of the studied corpora in order to identify, within a target language, the equivalent associations that are close to the source collocation. We assume that the POS of a collocation in the source language is identical in the target language.

**Assumption 3:** *"The POS of a collocation source and its equivalent in target language are identical"*

### 3.4 Discussion

First, we implement the z-score [1] 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. (*cf.* assumption 3). Concretely, some (T, Adj) and (T, N) collocation patterns are also multi-word term patterns used for terminology extraction [15]. Some collocations are also multi-word terms as noticed by Sinclair [17]. The overlap between collocations and multi-word terms is well-known problem in collocation extraction: *toxic gas* is a multiword term of Adj N pattern, but also a collocation with *gas* the base and *toxic* the collocate. Both phenomena share co-occurrence and syntactic criteria. The intersection between collocations and complex term sets regroups lexical associations sharing co-occurrence and syntax criteria. The assumptions of our collocation alignment method are based, in some ways, on the compositional translation of collocations, in specialized comparable corpora. In our case, the term and its proposed translation are considered as already aligned pivot. Morin & Daille [11] highlighted this criterion of compositionality for aligning multi-word terms.

## 4 KRC Alignment

After having identified source and target KRCs based on the translated collocations pairs, our goal now is to filter these KRCs, then to align the retained ones.

### 4.1 Monolingual Filtering of Contexts

The invalid contexts gathered by collocations will negatively affect the KRCs alignment. We implement criteria that we qualify as negative, in order to eliminate the less interesting contexts. We retain only "normalized" KRCs according to the following criteria applied in a monolingual way:

1. **context length:** we postulate that short sentences do not contain enough knowledge other than the collocation in question. Conversely, it is very difficult to consult those that are very long, also they may illustrate irrelevant information for the revision. As Kilgarriff & *al.* [5] we retain only sentences containing between 10 and 20 full words.
2. **pronouns:** Kilgarriff & *al.* [5] penalize contexts that contain pronominal anaphora, since it causes ambiguity. Especially, pronouns at the beginning of contexts may refer to text unities in previous sentences. Pronouns inside contexts are less problematic because they can refer to unities in the same sentence. For example, in the French context *"Desmarest commence ses recherches en volcanologie en 1763, en Auvergne, où* ***il*** *étudie les colonnes de basalte :* ***il*** *est le premier à reconnaître leur origine volcanique*" the pronoun *il* is not problematic because it is referring to *Demarest*. Here, we eliminate only contexts starting with a pronoun.
3. **affirmative contexts:** Kilgarriff & *al.* [5] prefer affirmative sentence rather than interrogative. We also retain this criterion to filter interrogative contexts.
4. **context complexity:** this criterion, which provides information on the readability of the sentence, was also addressed by Didakowski & *al.* [2]. We follow the same strategy using a dependencies 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.

### 4.2 Alignment of Contexts

The obtained KRCs at this stage are aligned only on the basis of the pair (source term/proposed translation) and its aligned collocations. Therefore, we suggest KRCs in an operational way: aligned according to other anchor points in addition to collocations. In the comparable corpora, parallel or lexically similar sentences are rare. It would be even more restricted to align KRCs on the base of their lexicon. Consequently, we propose to use similarity criteria that allocate to each source KRC an equivalent target KRC. These criteria represent "static" transition bridges from one language to the other:

1. **number of cognates**[2]**:** cognates represent transition bridges easily detected by the reader, in pairs of source and target contexts. Contexts sharing at least one cognate, will be aligned.
2. **number of translated simple terms:** although their scarcity in the corpus, sentences containing translated terms are exceptionally operational for the reviewer. The simple 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.

## 5 Evaluation

We manually analyzed the available corpus to prepare the reference KRCs for each studied term. We present in this section the used corpora, the baseline data and the experiments.

### 5.1 Corpora and Bilingual Dictionary

The evaluation of our method was carried on a specialized comparable corpus in the field of volcanology. This corpus is composed of English and French scientific documents containing about 400,000 words per language. These documents are obtained through a thematic research from newspapers and magazines such as *Le Monde*, *Sciences et avenir*, *Sciences et Vie*... They have been cleaned and standardized through the platform TermSuite[3] that also extracts the terminology.

The ELRA[4] dictionary that we used for the automatic alignment of collocations is a bilingual dictionary of general language English-French, containing 145,542 entries. It also contains the part-of-speech tags of entries.

### 5.2 Evaluation Data

The evaluation data is composed of 29 single-word terms central of the volcanology domain. The multi-word terms have been excluded for two reasons: on one hand, the identification of complex terms collocations can be treated as a separate issue that we do not regard in this work. On the other hand, the alignment of the collocations in which the term is complex could make the result interpretation more ambiguous. Terms have been separated into sources and targets, thus corresponding to translation pairs. Reference KRCs were given to every term of the translation pairs, in source and target languages. Finally, we obtain a list of 29 terms that we analysed in parallel: *basalt*/*basalte*, *cinder*/*scorie*, *crater*/*cratère*, *cone*/*cône*, *debris*/*débris*, *dome*/*dôme*, *fountain*/*fontaine*, *gas*/*gaz*, *lava*/*lave*,

[2] According to Léon [6], two cognates are two words starting with the same 4 characters.

[3] `https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite`

[4] `http://catalog.elra.info/product_info.php?products_id=666`

*magma/magma*, *scoria/scorie*, *tephra/téphra*, *volcan/volcan*, *vesicle/vésicule*, *eruption/éruption*.

The process that we followed to annotate KRCs of each pair (source term/ proposed translation) was:

1. Identify for each source term (and its translation) collocations that the collocate can be manually aligned (*i.e* without dictionary) with the collocate of the target term. For example, for the term *lava*, we found the collocation (*lava*, *gush*) in which the collocate *gush* is aligned with *jaillir* (FR) that forms a collocation with *lave* (FR). We obtain two collocations aligned through the collocate. Here, experts were solicited to check the manual translation of collocates. A monolingual concordancer[5] and a bilingual dictionary[6] were also used for the same aim.
2. Chek if contexts provided for each collocation are valid. A context is valid only if the collocation in question is valid within it.

**Table 1.** Manual validation of aligned collocations

| Corpora | # terms | # aligned terms | # pairs of aligned coll. |
|---|---|---|---|
| Vulcano EN (source) | 15 | 14 | 91 |
| Vulcano FR (target) | 14 | 14 | 85 |

Table 1 and table 2 present the results of the manual collocation alignment, as well as the results of the manual KRCs annotation respectively, for the reference data. *# terms* contains the number of terms on which our analysis are carried (15 EN and 14 FR). *# aligned terms* is the number of terms having at least one collocate manually aligned with one collocate of the corresponding produced translation. For example, we found that for the pair (*lava* $(EN)$/*lave* $(FR)$) the source collocation *(lava, pasty)* could be aligned with the target collocation *(lave, pâteux)*. Therefore, we count 1 source term aligned with 1 target term. *# pairs of aligned coll.* is the total number of collocation pairs that are manually aligned. For example, ($fountain$, *spout*) and ($fountain$, *spurt*) may be aligned with the same collocation ($fontaine$, $jaillir$). We count 2 EN collocation pairs aligned with only 1 FR pair.

**Table 2.** Manual validation of contexts aligned through collocations

| Corpora | # pairs aligned coll. | of # contexts | # valid contexts (KRCs) | # invalid contexts |
|---|---|---|---|---|
| Vulcano EN | 91 | 692 | 550 (79,74%) | 142 (20,52%) |
| Vulcano FR | 85 | 764 | 571 (74,73%) | 193 (25,26%) |

[5] `http://www4.caes.hku.hk/vocabulary/concordancer.htm`
[6] `http://www.wordreference.com/`

The second column in table 2 is borrowed from table 1. *# contexts* is the number of contexts provided by the pairs of aligned collocations: 91 EN collocations provided 692 EN contexts. *# valid contexts* correspond to KRCs that have been manually validated. A context is valid only if the collocation in question is valid within it. For example, the pair ($lava(EN)$, $lava(FR)$) produced by the translator. Two collocates *gush* (in *lava gush*) and *jaillir* (in *lave jaillit*) have been aligned, thus giving the contexts (a), (b), (c) and (d) (*cf.* table 3). The last one (d) has not been validated because the collocate *jaillir* depends of the term *fontaines de laves* instead of *lave*.

**Table 3.** Example of reference KRCs to review (*lava*, *lave*)

| Terms | Collocates | Contexts | evaluation |
|---|---|---|---|
| Lava (EN) | gush | (a) ***Lava*** *began* ***gushing*** *out of mayon's crater before dawn yesterday, accompanied by loud rumblings.* | valid |
| Lava (EN) | gush | (b) *The weight of* ***lava*** *above caused* ***lava*** *to* ***gush*** *freely from the vents, and for a few hellish hours rock flowing almost like water engulfed villagers, their livestock, and wild elephants on the slopes.* | valid |
| Lave (FR) | jaillit | (c) *Profitant de la cassure, la* ***lave jaillit****.* | valid |
| Lave (FR) | jaillit | (d) *De grandioses fontaines de* ***laves*** *ont* ***jailli*** *au travers des fissures qui se sont formées sur la face nord du dolomieu, le cône central de ce volcan culminant à près de 2 500 mètres.* | invalid |

### 5.3 Experimentation

We applied the collocation alignment as well as the KRC alignment on the 15 pair of terms. We notify that the evaluation is carried out on the qualitative and quantitative aspects of the aligned contexts obtained through the KRC alignment method.

The process of our experiments is to first evaluate the contexts provided by the automatically aligned collocations, this being a monolingual evaluation of the contexts, and to align the already retained KRCs. These pairs of source and target KRCs will be manually validated, thus corresponding to a bilingual evaluation.

In this study, we not present the monolingual evaluation of the contexts. On one hand, the collocation alignment only enables us to get pairs of collocations providing source and target contexts, but not aligned. On the other hand, the context evaluation, in this level, does not question the collocation alignment. Indeed, it does not inquire if a pair of aligned collocations provides KRCs in source and target language at the same time. Table 4 itemizes the four possibilities. We focus in particular on the case of *collocations pair 4*. In the other cases, the alignment of the contexts would be less relevant for revision.

**Table 4.** Alignment of collocations

| Type of source contexts | Collocations pairs | Type of target contexts |
|---|---|---|
| non KRC | Collocations pair 1 | non KRC |
| non KRC | Collocations pair 2 | KRC |
| KRC | Collocations pair 3 | non KRC |
| **KRC** | **Collocations pair 4** | **KRC** |

Concerning the bilingual evaluation of the aligned contexts, two experiments were performed: alignment of the contexts with and without filters. Here, our purpose is not to evaluate or to validate the proposed filtering criteria (cf. section 4.1), but rather to study their impact on the context's quality. 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.*

   The pair to be revised is *lava, lave*, the collocates which are *basaltic, basaltique* have been at first automatically extracted, then, translated using the bilingual dictionary ERLA. The cognates identified in these KRCs are *volcanoes* and *volcan*. The terms that were extracted in advance with TermSuite and translated using the same dictionary, are *shield, fluid, bouclier* and *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 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 5.** 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<br>Vulcano FR 14 | 10 | 23 | | 677<br>665 | 309 | 43,04% |
| **with filters** | | | | | | |
| Vulcano EN 15<br>Vulcano FR 14 | 10 | 16 | | 241<br>296 | 157 | 61% |

### 5.4 Results

Table 5 illustrate analysis of the aligned KRCs with and without filters. *# aligned terms* is the number of the pairs (term/proposed translation) having aligned source and target KRCs. In table 5, 10 translations among the 15 obtain at least one pair of aligned KRCs. *# pairs of aligned coll.* is the number of collocation pairs that the collocate is translated with a bilingual dictionary. The column *# contexts* is the number of the contexts provided by the aligned collocations. *# pairs of aligned KRCs* contains the number of the pairs (source KRC/target KRC) for all the aligned terms (*# aligned terms*).

In table 5, the analysis of the column *# contexts* comparing to *# pairs of aligned coll*, 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 can deduce that the application of filters deteriorate the productivity of the aligned collocations, nevertheless, it remains acceptable.

We note that even if the application of filters deteriorate the number of aligned KRCs, it significantly improve the precision of the alignment criteria since it moves from 43% to 61%. The number of the proposed pairs (*# aligned termes*) is conserved. We could not provide bilingual contexts for five pairs of terms. Some of these pairs have a too small number of extracted collocations or one syntactic structure, such as *vesicle/vésicule* that appears only twice in the studied corpus. For the others, *basalte/basalte*, *volcan/volcan* and *fountain/fontaine*, the alignment method act as a filter and eliminates contexts in both languages.

## 6 Conclusion

This work implements the first concordance programme that takes as input comparable corpora in specialised languages. We studied the problem of aiding the revision of a pair (term to translate/proposed translation) in order to ensure the consistency and the reliability of documents produced by the translators. We propose pairs of aligned knowledge-rich contexts from specialized comparable corpus. First, these contexts have been extracted in a monolingual way. Then, they have been aligned based on collocations, cognates as well as simple terms,

being anchor points that ensure the identification and the alignment of KRCs in specialized comparable corpora. The performed experiments show that the obtained contexts are acceptable for a manual revision. The study that we carried out deals with the qualitative aspects of the obtained KRCs. That is why our experiments relied on few terms.

## References


1. Berry-Rogghe, G.: The computation of collocations and their relevance in lexical studies. In: Aitken, A., Bailey, R., Hamilton-Smith, N. (eds.) The Computer and Literary Studies, pp. 103–112. Edinburgh University Press, Edinburgh (1973)
2. Didakowski, J., Lemnitzer, L., Geyken, A.: Automatic example sentence extraction for a contemporary German dictionary. In: Proceedings of the 15th EURALEX International Congress. pp. 343–349. Oslo, Norway (2012)
3. Gondoin, D.: Localisation de sites web: contraintes et enjeux. In: Lavault-Olléon, E. (ed.) Traduction spécialisée: pratiques, théories, formations, pp. 179–188. Peter Lang, Bern (2007)
4. Gouadec, D.: Profession: traducteur. La Maison du dictionnaire (2002)
5. Kilgarriff, A., Rychlý, P., Husák, M., Rundell, M., McAdam, K.: GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the 13th EURALEX International Congress. pp. 425–432. Barcelona, Spain (2008)
6. Léon, S.: Acquisition automatique de traductions d'unités lexicales complexes à partir du Web. Ph.D. thesis, Université de Provence, Aix-Marseille I (2008)
7. Mammino, L.: Il linguaggio e la scienza. Torino: Societá Editrice Internazionale (1995)
8. Marshman, E., L'Homme, M.C., Surtees, V.: Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. Corpora 3(2), 141–172 (2008)
9. Meyer, I.: Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In: Bourigault, D., Jacquemin, C., L'Homme, M.C. (eds.) Recent Advances in Computational Terminology, pp. 279–302 (2001)
10. Morin, E.: Des patrons lexico-syntaxiques pour aider au dpouillement terminologique. Traitement Automatique des Langues 40(1), 143166 (1999)
11. Morin, E., Daille, B.: Compositionality and lexical alignment of multi-word terms. Language Resources and Evaluation 44(1), 79–95 (2009)
12. Musacchio, M.T., Palumbo, G.: Shades of Grey: A Corpus-driven Analysis of LSP Phraseology for Translation Purposes. In: Taylor, C., Ackerley, K., Castello, E. (eds.) Corpora for University Language Teachers, pp. 69–79. Peter Lang, Bern (2008)
13. Newmark, P.: A Textbook of Translation. Prentice-Hall International (1988)
14. Permentiers, J., Troiano, F., , Springael, E.: Traduzione, adattamento ED Editing multilingue. TCG edition Bruxel (1996)
15. Roche, M.: Intégration de la construction de la terminologie de domaines spécialisés dans un processus global de fouille de textes. Ph.D. thesis, Université de Paris 11 (2004)
16. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. WaCky! Working papers on the Web as Corpus. Gedit pp. 63–98 (2006)
17. Sinclair, J.M., Jones, S., Daley, R.: English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08. (1970)