Using comparable corpora to characterize knowledge-rich contexts for various kinds of users: preliminary steps

Anne Condamines, Amélie Josselin-Leray, Cécile Fabre, Luce Lefeuvre, Aurélie Picton, Josette Rebeyrolle

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2013) 000–000

Procedia
Social and Behavioral Sciences

www.elsevier.com/locate/procedia

V International Conference on Corpus Linguistics (CILC2013)

Using Comparable Corpora to Characterize Knowledge-Rich Contexts for Various Kinds of Users: Preliminary Steps

Anne Condamines[a], Amélie Josselin-Leray[a], Cécile Fabre[a], Luce Lefeuvre[a], Aurélie Picton[b], Josette Rebeyrolle[a] *

[a]*CLLE-ERSS, UMR 5263, CNRS and Université Toulouse 2 Le Mirail, 5 Allées Antonio Machado, 31058 Toulouse, France*
[b]*TIM, Faculté de Traduction et d'Interprétation, University of Geneva, Rue du Général Dufour 24 1211, Geneva, Switzerland*

Abstract

The paper presents the early stage of the CRISTAL project, an original French project involving linguists, computer researchers and a firm specializing in multilingual text management. What is at stake from a linguistic point of view is a deeper analysis of the notion of Knowledge Rich Context proposed by Meyer (2001). Using comparable corpora, it analyzes how the notion of KRC can vary according to text genre and/or type of users.

Key-words: CAT; comparable corpora; knowledge-rich context; translation; terminology; knowledge engineering.

1. Theoretical background : the notion of KRCs and its applications

1.1. What are KRCs?

The idea that some contexts can prove useful in defining words is far from being new - Aristotle already asserted that there was such a thing as defining contexts. The growing importance of corpus linguistics and information extraction has resulted in a need for a more systematic description of those contexts in order to be able to retrieve them automatically. More recently, the needs have focused on specialized fields, for text-based terminology- or ontology-building. It is within that framework that Ingrid Meyer defined Knowledge Rich Contexts (KRCs) as follows: "By knowledge-rich context, we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis" (2001: 281).

* Corresponding author. Tel.: +33(0)561503608 ; fax: +33(0)561504677
E-mail address: anne.condamines@univ-tlse2.fr

Selection and peer-review under responsibility of CILC2013.

1.2. KRCs & Ontologies

Until now, the notion of KRC has been mainly studied for the building of ontologies (Auger & Barrière, 2008). It is therefore very often linked to the notion of markers of conceptual relations, as illustrated by the following quotation from (Marshman, 2008: 125): "KRCs may be defined as segments of text that convey information that can be useful for concept analysis (e.g. indications of conceptual relations or attributes)". However, our hypothesis in the project is that KRCs can vary (Aussenac-Gilles & Condamines, 2012) depending on the domain, depending on the text genre (Condamines, 2002), depending on the language and, maybe most importantly, depending on the use (translation vs. knowledge engineering for instance).

In knowledge engineering, users are mostly interested in building networks of terms (termino-ontologies) and thus tend to focus on markers of conceptual relations such as meronymy and hyperonymy. Here are two examples of KRCs taken from the comparable corpora we are using for the project and which are described below in section 2.3.1.

(1) Un volcan se compose de trois parties : un réservoir, une cheminée et un édifice visible en surface.

(2) There are three types of lava and lava flows: pillow, pahoehoe, and aa.

In (1) there is a clear relation of meronymy between the French term *volcan* and the French terms *réservoir*, *cheminée* and *édifice* which is signalled by the use of the conceptual markers which are underlined. In (2), the conceptual marker "types of" clearly indicates a relation of hyperonymy between the hyperonyms *lava* and *lava flows*, and the hyponyms *pillow*, *pahoehoe* and *aa*.

From a linguistic point a view, what is at stake is to assess how Term-Relation-Term triplets can be built based on specialized texts, the whole of the triplets making up a network.

1.3. KRCS & Translation

It is a well-known fact that context is a key element in the translation process. What do translators need contextual information for? In translation, users are naturally interested in conceptual relations (Marshman, Gariépy & Harms, 2012) but they also use the contexts for "subject-field understanding, correct term choice and idiomatic expression" (Bowker, 1998: 18). This implies they are not only interested in conceptual markers (or in contexts containing Term-Relation-Term triplets), but also in contexts containing defining elements such as (3), synonyms such as (4) or collocations such as the N+V collocation in (5) (Josselin-Leray, & Roberts, 2007).

(3) Tephra is any material ejected explosively from a volcano (ash, lapilli, cinder, and spatter).

(4) Composite cones, or stratovolcanoes, are large, nearly symmetrical structures built of interbedded lavas and pyroclastic deposits.

(5) Lorsque le magma s'épanche en surface il forme une roche volcanique.

Bowker (2011: 214) underlines that "translators really want usage information – such as contexts or collocations – which will help them to produce a target text that reads well". The problem is that the various resources at their disposal do not sufficiently meet their needs. One of the main sources translators rely on is term records found in term banks such as *Termium* or *Le Grand Dictionnaire Terminologique*. Even though translators make up the largest user of term banks, those term records also serve a wide range of other users. For various reasons that are carefully listed by Bowker (2011: 214-215), the information found on those records is rather limited and usually consists in definitions and terms presented out of context, or in only a single context. Bowker (2011: 215) thus suggests: "it would be more helpful for translators to have access not simply to term records that provide a single 'best' term with a solitary context, but rather to information that would allow them to see all possible terms in a range of contexts and thus find the solution that works best in the target text at hand". She insists on the fact that looking at a wide range of contexts should not be considered as a waste of time, and that this has been made easier thanks to corpus-analysis tools that present information in an easy-to-read format.

Another tool most translators use nowadays is CAT (Computer-Aided Translation) tool suites which usually include translation memory systems and automatic term extractors, such as SDL Trados Studio 2011 or else DéjàVu. Translation memory systems are a database of aligned source and target texts, that can be considered as a type of parallel corpus (Bowker, 2011: 218). Even though these have proved very useful to translators since they

allow them to work faster by reusing segments of texts that have been previously translated, using translated material as a resource may have drawbacks since this type of text is more likely to contain "conceptual or terminological errors, awkward syntactic constructions, or non-idiomatic expressions" (Bowker, 2011: 221). While comparable corpora have been considered very useful in the translation process (Zanettin: 1998, Josselin-Leray: 2005), to our knowledge to this day no CAT tool relies on such corpora.

Moreover, Bowker (2011: 215) states that while the corpus-based approach has had an impact on the terminological process in the field of terminology, and while corpora and tools for corpus processing have been in widespread use in the translation profession for approximately fifteen years, "the impact on the product (i.e. term records in a term bank) has been negligible". Even if there have been a few attempts at improving the context-related contents of term banks (e.g. the model template devised by Pearson (1998: 2000) or the DicoInfo compiled at the University of Montreal), the notion of KRC and the potential usefulness of comparable corpora have not been seriously taken into account so far. The CRISTAL project attempts to bridge the gap.

2. The CRISTAL Project

2.1. General background

The CRISTAL project is a three-year project funded by the French National Agency for Research (ANR; ANR-12-CORD-0020). CRISTAL is an acronym that stands for "Knowledge-rich contexts for terminological translation" (Contextes Riches en Connaissances pour la Traduction Terminologique in French). Four different partners are involved: a Computing research team at the University of Nantes, France (LINA), a linguistics research team at the University of Toulouse-Le-Mirail, France (CLLE-ERSS), the Translation Technologies team from the Faculty of Interpreting and Translation at the University of Geneva in Switzerland, and a firm specializing in multilingual text management (Lingua et Machina). The project started in November 2012 and is thus only at its very beginning.

The aim is to retrieve the contexts from bilingual corpora that are the most relevant for translation and to provide them to users through the CAT tool developed by Lingua & Machina, the Libellex Platform. In the current state of the platform, the term to be translated and its candidate translations are enriched with basic contextual information, i.e. part-of-speech, a list of the first three cooccurrents ranked by an association score, and a passage – or context – selected according to a metric based on the association scores between the cooccurrents and the length of the sentence. In other words, a "rich context" so far is a context in which significant cooccurrents of the term appear.

2.2. The linguistic issues

From a linguistic point of view, the interest of this project is threefold.

First, while KRCs have been the focus of various studies which take into account variation depending on text genre (Condamines, 2002), (Marshman & al., 2008), the idea of how relevant KRCs might be depending on use has never been investigated as such so far.

The second major element in this project is the use of comparable corpora as a resource for translation and for building and enhancing termino-ontologies. As far as translation is concerned, we are trying to show how using a comparable corpus differs from using an aligned corpus and whether it proves more efficient. For the building of termino-ontologies, we want to find out whether a comparable corpus simply doubles the available textual evidence or if there might be some other benefits to it.

Third, we wish to extend and compare the testing methods used in translation studies by combining feedback-based and eye-tracking-based experimentations. Eye-tracking has not been often used with specialized texts and the few existing studies rely mostly on experimentations led by computer scientists (Jensen, 1998).

Finally, the project should allow us to assess the interpretabiblity and the relevance of the results thus obtained.

In summary, we are trying to establish how what KRCs are used for – for building termino-ontologies or as a help in translation – can help define what they really are. Besides the "variety of use" aspect, we are also trying to assess how helpful a comparable corpus can be for such a task.

2.3. Methodology

The methodology we are trying to devise is based on two essential elements: the comparable corpora and the implementation of approaches used in other studies in order to see how relevant they might be to reach our goal.

2.3.1. The use of comparable corpora

Let us remind here that by comparable corpora, we mean the same as (Altenberg & Granger 2002: 8): "Comparable corpora consist of original texts in each language, matched as far as possible in terms of text type, subject matter and communicative function". The two languages involved in our corpora are French and English. The following table sums up the main features of the corpora:

	VOLCANOLOGY	BREAST CANCER
POPULAR SCIENCE TEXTS	400,000 words/language	200,000/language
	1980-2002	2002-2008
SPECIALIZED TEXTS	Under construction	200,000/language
	Under construction	2001-2008

Table 1. Presentation of the corpora

2.3.2. Two different approaches

Two different approaches will be implemented: the first one will allow us to identify new relation markers in order to build termino-ontologies; the second one will enable us to spot and analyse which KRCs prove to be of greatest use to translators.

2.3.2.1. The recursive method for identifying new relation markers

The so-called recursive method, is well-known method in text-based Knowledge Engineering (Hearst, 1992) (Morin & Jacquemin, 1999) and which was already used by (Condamines & Rebeyrolle 2001). In this method, pairs of terms linked by a clearly identified conceptual relation are first searched in texts in order to identify new patterns; then these patterns are researched in order to spot new pairs of terms and so on.

This method relies on the idea that some relation markers are not necessarily known beforehand, either because they have not been described yet in the literature, or because they are specific to the field under study. But the very fact that there is a relation between two terms can be used as a starting point for identifying new markers.

For example, in the medical field, a "generic" marker such as [N1 entraîner N2] in French can allow us to immediately sport a cause relation between *anti-hypertenseur* and *vertiges* in (6)

(6) La prise d'un anti-hypertenseur a entraîné des vertiges chez Monsieur Y

If we then search the same pair *antihypertenseur/vertiges* in a medical corpus, we can identify another marker which is specific to the medical field: ([sous N1, N3 (développer, avoir) N2], as in (7)

(7) Sous anti-hypertenseur, Monsieur X s'est mis à avoir des vertiges

The comparable corpus will reveal whether the markers are equivalent from one language to another or if they often differ.

2.3.2.2. Experiment(s) with translators

Our second approach consists in carrying out one or several experiments with translators, in order to identify which contexts (containing conceptual relation markers, or not) translators find most useful when faced with several contexts aiming at helping them with their translation.

Several testing methods have been so far for empirical research in translation studies: think-aloud protocols, direct observation of translators, post-translation feedback…(see for instance Bowker 1998 & Künzli 2001). More recently, experiments based on eye-tracking systems named UAD (User Activity Data) have been led. They "consist of the translator's recorded keystroke and eye-movement behavior, which makes it possible to replay a translation session and to then register the subjects' comments on their own behavior during a retrospective interview" (Carl & al., 2008, 21); see also (Gopferich, Jakobsen, & Mees, 2008).

We are planning to rely partly on the methodology devised by (Bowker, 1998) who asked translator trainees to translate a semi-specialized text into their native language. The translators were divided into two groups, some using monolingual corpora and some using only conventional translation resources (i.e. no corpus). The quality of the translations thus obtained was then compared, and the translators were asked to provide comments on the process. In our case, several groups will probably be tested: one group with only conventional resources, one group which will be provided with several KRCs extracted beforehand from our comparable corpora and which will have to rate their usefulness, one with free access to our comparable corpora whose behaviour will be monitored thanks to UAD and screen-recording. We are thus considering combining several methods in our own experiment(s), which will all be based on comparable corpora and might involve expert translators as well.

3. Conclusion

There are many challenges to address in the near future: *(i)* adapt the recursive method to comparable corpora and, at the same time, refine the list of markers, *(ii)* complete our methodology for the experiment with translators, *(iii)* carry out the experiment, *(iv)* analyze and compare the results of the two methodological approaches.

The CRISTAL project is clearly still in its infancy but seems very promising. Linguistically speaking, the most innovative aspect lies in the refined characterization of KRCs. First, it will take into account how the needs for KRCs might vary depending on the type of user (translators vs. knowledge engineers). Second, it will study how using comparable corpora might improve the quality of KRCs proposed to users. Finally, it will assess how useful eye-tracking might be in terminology studies.

Altenberg, B. & Granger, S. (2002). Recent Trends in Cross-Linguistic Lexical Studies. In B. Altenberg & S. Granger (Eds.), *Lexis in Contrast, Corpus-Based Approaches* (pp. 3-48). Amsterdam, Philadelphie: John Benjamins.

Auger, A., & Barrière, C. (2008). Pattern based approaches to semantic relation extraction: a state-of-the-art. *Terminology*, 14:1, 1-19.

Aussenac-Gilles, N., & Condamines, A., (2012). Variation and semantic relation interpretation: Linguistic and processing issues ». In L. Aguado de Cea & al. (Eds.): *Proceedings of the 10th Terminology and Knowledge Engineering Conference* (TKE 2012), (pp.106-122). 19-22 June 2012, Madrid, Espagne.

Bowker, L. (1998). Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta : journal des traducteurs / Meta: Translators' Journal*, 43-4, 631-651.

Bowker, L. (2011). Off the record and on the fly: Examining the impact of corpora on terminographic practice in the context of translation. In A. Kruger, K. Wallmach & J. Munday (Eds.) : *Corpus-based Translation Studies: Research and Applications.* London/New York: Continuum. 211-236.

Carl, M., Arnt L. J., & K.T., Jensen. (2008). Modelling human translator behaviour with User-Activity Data. In *[Proceedings of the] 12th EAMT conference, 22-23 September 2008, Hamburg*, 21-26.

Condamines, A. (2002). Corpus Analysis and Conceptual Relation Patterns. *Terminology*, 8:1, 141-162.

Condamines, A. & Rebeyrolle, J. (2001). Searching for and Identifying Conceptual Relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB) : method and results (pp.127-48). In D. Bourigault, M.-C. L'homme & C. Jacquemin, (Eds.), *Recent Advances in Computational Terminology*. Amsterdam, Philadelphia: John Benjamins.

Gopferich, S., Jakobsen, A. L, & Mees I., M. (Eds.) (2008). *Looking at Eyes: Eye-Tracking Studies of reading & Translation Processing*. Copenhagen: Copenhagen Studies in Languages, 36.

Hearst, M.A. (1992). Automatic Acquisition of Hyponyms From Large Text Corpora. 14th International Conference on Computational Linguistics, Nantes, France. 539-545.

Josselin-Leray, A. (2005). Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues. Etude d'un domaine de spécialité : volcanologie. PhD thesis, Université Lyon II Lumière, France.

Josselin-Leray, A. & Roberts, R.P. (2007). La définition des termes dans les dictionnaires généraux unilingues : analyse de quelques exemples du domaine de la volcanologie à la lumière d'un corpus de vulgarisation. In L'Homme, M. C. & Vandaele, S. (Eds.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes* (pp. 141-188). Ottawa: Presses de l'Université d'Ottawa.

Künzli, A. (2001). Experts versus novices : l'utilisation de sources d'information pendant le processus de traduction. *Meta*, 43 :3, 507-523.

Jensen, C. (2008). Assessing eye-tracking accuracy in translation studies. Copenhagen : Copenhagen Studies in Language, 36, 157-174.

Marshman, E., J., Gariépy & C., Harms (2012). Helping translators manage terminological relations: Storing and using occurrences of terminological relations and lexical relation markers. *JoSTrans: Journal of Specialised Translation* 18, 30-56.

Marshman, E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts. *Terminology* 14: 1, 124-151.

Marshman, E., L'Homme, M.-C. & Surtees, V., (2008). Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, 3, 141-17.

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

Morin, E. & Christian J. (1999). Projecting Corpus-Based Semantic Links on a Thesaurus. *In Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL)*.University of Maryland, USA. 389–396.

Pearson, J. (1998). *Terms in Context*. Amsterdam, Philadelphia: John Benjamins.

Zanettin, F. (1998). Bilingual Comparable Corpora and Training of Translators. *Meta*, 43-4, 616-630.