Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol

Emmanuel Planas, Aurélie Picton, Amélie Josselin-Leray

Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol

Emmanuel Planas[1], Aurélie Picton[2], Amélie Josselin-Leray[3]

[1]LINA, Université de Nantes & IPLV, Université Catholique de l'Ouest

[2]TIM, Faculty of Translation and Interpreting, University of Geneva

[3]CLLE-ERSS, Université Toulouse 2 & CNRS (UMR589)

emmanuel.planas@univ-nantes.fr, aurelie.picton@unige.ch, josselin@univ-tlse2.fr

Abstract: This paper outlines the design of and some methodological conclusions drawn from a pilot study conducted among trainee translators to measure the use and usefulness of Knowledge-Rich Contexts (KRCs) in the translation process. After discussing the issue of context and KRCs in translation, it reviews the literature on previous observation protocols and tools designed for the study of the translation process. It then presents the customized software designed for the experiment to record the translator's activity. It describes the details of the pilot study, and, finally, some preliminary results and methodological changes planned for the subsequent final experiment(s).

Keywords. Knowledge-Rich Contexts; Terminology; Empirical studies; Translation Process; CAT tools; Logging

1 Introduction

Knowledge-Rich Contexts (KRCs), defined by Meyer [1] as "context[s] indicating at least one item of domain knowledge that could be useful for conceptual analysis", are a well-known notion in terminology and knowledge extraction. Although the existing studies about KRCs originally focused mainly on text-based terminology or ontology-building [2][3], more recently, several papers (e.g. [4]) have shed light on the importance of such contexts for translators: having access to usage information for a given term or to semantic and conceptual relationships between terms –be it in the source language or in the target language– is essential for translators. The (semi-) automatic extraction of Knowledge-Rich Contexts thus seems very relevant.

This is what the CRISTAL project[1] aims at doing, by retrieving KRCs from bilingual comparable corpora and integrating them into CAT tools. However, tailoring this kind of tools to suit the translator's needs implies refining what underlies the notion of KRC. In this paper, we are thus testing a protocol that precisely aims at providing insights into various elements to better understand what a good KRC for translators is. In the medium term, the experiments to come should allow us to develop a typology of the most useful KRCs for a translator, to gather details on their required extension and structure, to get some information about the stages of the translation process in which KRCs are most needed and about the way they are used in relation to other resources (such as dictionaries or term banks). For the time being, we focus on the protocol itself. To meet this ultimate objective, the protocol relies on a combination of different technologies to record the translator's use of KRCs in an environment thought to be as "ecological" as possible [5]. The first part of this paper (section 2) focuses on the issue of context in translation; section 3 then provides a short review of existing methods to observe the translation process. This review provides the basis of a new interface we designed to better identify the use of resources by translators: Argos (section 4). In section 5, we present the pilot study led at the University of Geneva to validate our protocol. The preliminary results are provided in section 6.

2 KRCs and Translation

Even though it is generally agreed that context is an essential component of translation, the definition of that fuzzy notion remains somewhat unclear, maybe due to the fact that it is used in many fields, e.g. philosophy, psychology, and linguistics. Following Melby & Foster [6],we define the context of a lexical unit as the text that surrounds it, i.e. the units that precede/follow it, at sentence level or on a larger scale.

A number of shortcomings regarding context can be identified in the tools that translators generally have at their disposal: dictionaries, term banks, and CAT tools. As underlined by Varantola [7] and Bowker [4], since dictionaries try to provide general information that can be applied to a wide array of situations, they usually provide "context-free descriptions of word-use", i.e. prototypical information, which is of limited use to translators who need context-specific information. Moreover, when provided, the context-related data is usually presented in a very condensed version, while translators "also need information relating to longer stretches of text than a single lexical item" [7]. Paradoxically enough, despite the advances in terminology research about context, and in particular KRCs, Bowker [4] notes that what translators usually find in term banks are "terms presented out of context, or in

[1] CRISTAL ("Contextes RIches en connaissanceS pour la trAduction terminoLogique") is an original French project involving linguists, computer researchers and a firm specializing in multilingual text management. The CRISTAL project is a three-year project funded by the French National Agency for Research (ANR; ANR- 12-CORD-0020).

only one single context" (which is usually provided only for the "best" term), while what they need is actually "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". Barrière [8] shows that very simple IR techniques on the biggest corpora available provide better terminological support than the biggest term banks available. Finally, terms automatically provided by term databases in CAT tools are not shown in context, but in a small window providing a translation proposal, a comment and some non-linguistic data like the date or author. However, in tools such as Transit[2] and Multitrans[3], translators can intentionally search in translation memory databases (parallel texts) for some concordance-like contexts for a given term.

What makes a Knowledge-Rich Context in the field of translation? According to Bowker [9], the notion of KRC can be widely understood as "any context that contains useful information" for the translation process. In Bowker [9][4], she draws a list of those items of information that can prove useful for the translator which can be summed up as follows: (*i*) information about usage; this of course includes collocations, in particular which general-language words collocate with terms, (*ii*) information about the frequency of use of a particular word or term, (*iii*) information about lexical and conceptual relations (such as synonymy, meronymy, hyperonymy etc.), (*iv*) pragmatic information about style, register and genre –something which was already underlined by Varantola [7] back in 1998, (*v*) information about usages to avoid. This list can seem really extensive, and Bowker [4] even adds that "translators might not even know what they need: they are seeking inspiration, associations, similar examples, parallel situations that can be adapted." She concludes by saying that "it is often a case of I don't know what I'm looking for, but I'll recognize it when I see it". According to her, the information needed by translators could/should be provided through corpus-based "word-clouds" (with frequency data), "collocate clouds" and a large number of corpus-based contexts that could be presented as KWIC concordances [9][4]. While Barrière [10] proposes a tool to help terminologists collect corpora and build KRCs semi-automatically, a hands-on experiment to test the use and usefulness of pre-selected KRCs thus seems very welcome. That is what we propose, through the observation of translators and the recording (log) of their actions while they translate.

3 Observation of Translators in Action: a Brief State of the Art

In order to gather information about the KRCs translators resort to when translating, it seemed necessary to first examine previous observation protocols and tools designed for the study of the translation process.

---

[2] http://star-group.net/ENU/group-transit-nxt/transit.html (last consulted 02.28.14)

[3] http://multicorpora.com/products-services/other-available-products/ (last consulted 02.28.14)

Before technology was available for recording translators' activity on the fly, researchers favored methods where the translator would express his processes either orally or on paper. Göpferich & Jääskeläinen [11] and Ehrensberger-Dow & Massay [12] identify: (*i*) Think Aloud Protocols (TAPs) where the translator, while translating, comments aloud his choices, which are recorded on a tape recorder; (*ii*) dialogue protocols where the decisions about translation are taken through a dialogue between peer translators, which is also tape-recorded; (*iii*) retrospective interviews, where the translation is explained just after being done (for short memory matters); (*iv*) integrated problem and decision reporting (IPDR), where translators write down and explain points they think critical; (*v*) questionnaires, interviews, and diaries. It was common –and still is– to ask about the translators' background and translation habits through (*vi*) pre-questionnaires or (*vii*) interviews [13][14]. These methods have often been used to evaluate the differences in the use of translation resources between experimented and trainee translators [15][13][5]. Varantola [7], Künzli [13], Desilets et al. [16] noted down (*viii*) which resource was used during translation. Bowker [17] and Delpech [18] used the (*ix*) separation into two or more groups to study the influence of the use of specific resources on translation quality.

Yet, some of these methods were proved [19][12] to be invasive enough to disturb the translator's natural translation process. That is why, following the idea of "ecological validity" introduced by Ehrensberger & Massey [5], we prefer to use a method fostering the respect of the translator's natural environment, such as key-logging software. Since the late 2000s, key logging software (such as Inputlog [20] or Translog [21]) have allowed recording the translator's textual production without intrusion. The QRedit interface of the MNH-TT plateform [22] gives an alternative for logging a collaborative translation on the web. For studying the use of translation resources, key logging software are fully useful only if there is a means to set a link between the text that is typed and the resources used by the translator. This is the case when the screen activity as a whole is recorded at the same time. Pieces of software like Camtasia[4] or BB Flash Back[5] can help in that respect. In addition to these, eye-trackers (ET) became precise enough around 2010 to map which word was looked at, and at what time, by the translator's eyes (e.g. [23]). Even though ET could give us more precise information about a translator's decision, at this stage of our research, we are more interested in textual information that we can post-process automatically.

4 Argos: a New Interface for Translation Process Observation

In order to observe and record the translators' use of KRCs, we looked at the existing software. Logged interfaces like Translog or InputLog show the participants

[4] http://www.techsmith.com/camtasia (last consulted: 28 February 2014)

[5] http://www.bbsoftware.co.uk (last consulted: 28 February 2014)

the source text in a source window and records all the edition changes the translator inputs in a target window. But such tools do not provide any logged interface for the translation resources used by the translator. CAT tools like OmegaT [6], or TradosStudio[7], do provide a complete interface for the use of translation resources, but the editing activity is not recorded in a log.

Using a combination of such a CAT tool with existing logging software like Camtasia Studio or BB Flashback would provide us with a recorded video of the screen, a log of the typed text, and possibly the changes of software window. But it would not record which resource was used and which proved useful. This led us to the conclusion that we had to design a new logged interface that would meet our precise needs. For it to be as close as possible to the translator's usual environment, we studied existing CAT tools because they are organized in an ergonomic design translators are familiar with. With all this in mind, we created Argos[8].

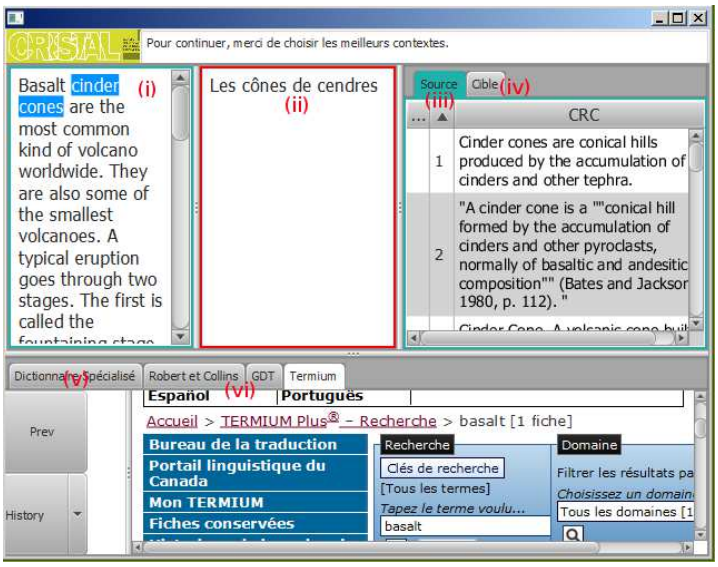

Fig. 1. The Argos interface

Argos is composed of (*i*) a source window in which the source text (ST) is displayed; (*ii*) a target window in which the translator can type in his translation (TT); (*iii*) a window where a list of source contexts are displayed when the translator selects

[6] http://www.omegat.org (last consulted: 28 February 2014)

[7] http://www.translationzone.com/trados.html (last consulted: 28 February 2014)

[8] Argos is coded in Java 1.7 and has been tested on Linux, Mac OS and Windows.

a term in the ST (like *cinder cones* in the screen shot); (*iv*) a window where a list of target contexts is displayed when the translator types a target text in a specific input field; (*v*) a dedicated window for a bilingual specialized dictionary, where target terms are displayed when source terms are entered in a specific field; (*vi*) 4 logged tabbed windows connected to specific translation resource URLs (e.g. *Termium*) (see §5.3).

All keyboard activity is recorded, whichever window is being used: characters, deletion, etc. When the translator selects a (simple or complex) term in the ST, a list of KRCs is displayed in the KRC window. This blocks the TT window until the translator chooses at least one KRC with a simple click, forcing him to explicit which KRC was useful. The same mechanism is set for the target KRCs. The translator's queries about terms and his KRC choices are recorded into the log.

5 Pilot Study

The experiment we present here is a pilot study that aimed at testing our protocol and the translation interface with a small number of participants. Two larger scale studies are also planned, involving 20 participants each.

5.1 Participants

7 students from the Faculty of Translation and Interpreting at the University of Geneva –4 Master's students, and 3 PhD students– participated in the experiment. We felt that these students would be good candidates as they all had French as a mother tongue, had followed translation courses from English to French and were all familiar with CAT tools. The PhD students had some professional experience in translation.

5.2 Text

The text to be translated was chosen based on previous experiments with translators (e.g. [13][17][7]). Its main features are: (*i*) written in English, to be translated into French; (*ii*) 150-word long, to be translated in less than 2 hours; (*iii*) dealing with a subject that was (a) technical enough for fostering terminology search, but (b) not highly technical for the students, and (c) familiar enough to us to ensure we would be able to assess the quality of the translations at the end of the experiment, (*iv*) containing a number of collocational and syntactic difficulties, (v) structured in a very logical way. We picked an extract from a popular-science book on volcanology[9] that describes the 2 phases in which *cinder cones* are built.

[9] *What's so hot about volcanoes?,* Wendell A. Duffield (2011), Mountain Press.

5.3 Resources

Lexicographic resources. Participants had access to: *the Robert & Collins* (English-French, French-English), *Termium*, the *Grand Dictionnaire Terminologique*, (all three online), and some entries of a specialized bilingual dictionary of volcanology[10].

KRCs. For some terms in the text, we selected different types of supposed KRCs[11]. First, some contexts were selected in the source language (English). Second, we tried to anticipate possible equivalents in the target language (French) for each term and provided contexts for each. A dozen one-sentence long KRCs were provided to the participants for each term. We tried to put together different types of KRCs for each term, such as definitions, hyperonymy, synonyms, collocations, etc. We added some "Knowledge Poor Contexts", supposed to be of no use to the translator (Table 1).

Definition	**Scoria** is very vesicular, low density basalt.
Hyperonymy	Volcano type: **scoria** cone, shield volcano, stratovolcano.
Property + collocation	Many **scoria** cones are monogenetic in that they only erupt once, in contrast to shield volcanoes and stratovolcanoes.
"Knowledge *poor* context"	The second moai has a Pukao which is made of red **scoria**.

Table 1. Examples of KRCs for *scoria* (source language, English)

Internet Access. Unlike Master's students, PhD students had access to Google through the interface, in order to search for resources we would have not thought of.

5.4 Questionnaire and Interviews

The translation task was completed by an online questionnaire about the main translation difficulties, the use of resources and KRCs, the relevance of KRCs, the stages of the translation process when KRCs were needed most, the interface, and general information (age, experience, degrees, etc.). Then, an approx. 20-minute semi-structured interview was conducted with all the participants.

5.5 Experiment

After a 15-minute test of the environment, the students were allocated 2 hours to translate the text, and to indicate which KRCs were the most useful. Their activity

[10] *Dictionnaire bilingue des Sciences de la Terre* (anglais / français) (2013), Michel J.-P. *et al*., Dunod, 5th edition. Relevant entries were converted into electronic form.

[11] These were taken partly from a comparable, French-English, popular-science corpus compiled by Josselin-Leray [24], partly from reliable documents found on the Internet.

was recorded and saved. Immediately after the translation task, we asked them to fill in the questionnaire. We then conducted the recorded interviews.

6 Preliminary Results

The nature of the preliminary results of the pilot study is twofold: (*i*) they provide feedback regarding the validity of our protocol; (*ii*) they allow us to identify some preliminary tendencies about the use of KRCs during the translation process.

Data Analysis. The expected analysis of the results obtained through this protocol relies on the complementarity of different types of data: questionnaires, video recording, logs, and final translations. Alves [24] showed how the combination of these different techniques –which he calls "triangulation"– leads to more explicit results. To help us read the logs, we created automatic post-logging compilation processes. These gather all the translators' individual logs in one file containing: the terms that were searched, the resources they were searched in, the KRCs that were selected, the (anonymized) translators that selected them.

Validation of the protocol. Our protocol is operational and everything went smoothly during the experiment, without any interfering on the translation process. All the data was saved, and the log compiled all the results to be observed. All participants warmly welcomed the protocol they considered user-friendly and respectful of most of their environment, especially regarding (*i*) the resources provided, (*ii*) the usability of the interface, (*iii*) the appearance of the interface which was close to existing CAT tools and (*iv*) the level of difficulty of the text. The difficulties we had anticipated in the text we chose were identified as such and treated by all the participants with all the resources provided.

First results on KRCs. The most important finding is that the participants overwhelmingly chose knowledge-rich contexts and discarded "knowledge-poor" contexts: out of 92 contexts that were selected by the participants, only 5 were "knowledge-poor" contexts. In addition, 6 participants out of 7 clearly stated that the KRCs selected in the interface were very useful and used them to translate, especially KRCs that contain information about collocations. However, even if KRCs are indeed valuable, their usefulness decreases when the information they present is either irrelevant or not easily accessible. It is then of prime importance to work on the diversity of KRCs, on their quality, but also on their layout.

Future adjustments. Feedback from the participants leads us to operate several adjustments for the full-scale experiments to come. Among them, participants suggested to better select the "target" KRCs, and to complete the list provided. We are working on better anticipation of the types of target KRCs to display. Second, even if the size of the text was suitable, the participants voiced concern about the fact

that it was only an excerpt of a chapter, which hampered their translation. We will then provide the full text, but ask the participants to translate only the chosen excerpt. Last but not least, some features suggested by the participants will be added to the interface, such as an electronic notepad and keyboard shortcuts.

7 Conclusive Remarks and Perspectives

The pilot study presented in this paper is a stand-alone experiment designed to test logistics and gather information prior to a larger study, in order to improve its quality and reliability. The results from the pilot study show that our protocol and the Interface Argos sound promising to assess the use and usefulness of KRCs in translation on a large scale (over 40 participants), in an environment which is as "ecological" as possible. What makes the future results worthy of interest is the quantity, the quality and the diversity of the data (cf. §6 "triangulation" [24]). At this point of our research, the analysis of all the data collected still needs to be refined. The subsequent experiments will enable us to quantify and generalize some tendencies and complement this with fine-grained observations especially through the analysis of the quality of the translations obtained, the viewing of the video recordings and the semi-structured interviews. Finally, the replicability of the method, which also allows one to compare several groups of translators, several types of KRCs, and several types of texts, seems to guarantee a refined comprehension of the linguistic phenomena that are at stake in specialized translation.

8. Bibliography

1. Meyer, I.: Extracting Knowledge-Rich Contexts for Terminology: A Conceptual and Methodological Framework. Recent Advances in Computational Terminology. pp. 279–302. Bourigault, D., Jacquemin, C., L'Homme, M.-C. (2001).
2. Auger, A., Barrière, C.: Pattern-based Approaches to Semantic Relation Extraction: A State-of-the-Art. Terminology. 14, 1–19 (2008).
3. Aussenac-Gilles, N., Séguéla, P.: Les relations sémantiques : du linguistique au formel. Cahiers de grammaire. Numéro spécial linguistique de corpus. Toulouse (2000).
4. Bowker, L.: Meeting the Needs of Translators in the Age of e-Lexicography: Exploring the Possibilities. Electronic Lexicography. pp. 379–387. S. Granger & M. Paquot (2012).
5. Ehrensberger-Dow, M., Massey, G.: Exploring Translation Competence by Triangulating Empirical Data. Studies in Translation. 16, 1–20 (2008).
6. Melby, A., Foster, C.: Context in Translation: Definition, Access and Teamwork. The International Journal of Translation and Interpreting. 2, 1–15 (2010).
7. Varantola, K.: Translators and their Use of Dictionaries. User Needs and User Habits. Using Dictionaries. pp. 179–192. B.T.S. Atkins, Tübingen (1998).
8. Barriere, C., Isabelle, P.: Searching Parallel Corpora for Contextually Equivalent Terms. Presented at the 15th Annual Conference of the European Association for Machine Translation , Leuven, Belgique May 30 (2011).

9. Bowker, L.: Off the Record and on the Fly, Corpus-based Translation Studies: Research and Applications. Kruger, A, Wallmach, K., & Munday, J., London/New York (2011).
10. Barriere, C.: Semi-Automatic Corpus Construction from Informative Texts. Lexicography, Terminology, and Translation: Text-based Studies in Honour of Ingrid Meyer. pp. 81–92. Lynne Bowker, Ottawa (2006).
11. Göpferich, S., Jakobsen, A.L., Mees, I.M.: Looking at Eyes Eye-Tracking Studies of Reading and Translation Processing. S. Göpferich, A. Lykke Jakobsen, I.M. Mees (2008).
12. Ehrensberger-Dow, M., Massey, G.: Indicators of translation competence: Translators' self-concepts and the translation of titles. Journal of the Writing Research. 5, 103–131 (2013).
13. Künzli, A.: Experts vs novices: L'utilisation de sources d'information pendant le processus de traduction. Meta. 46, 507–523 (2001).
14. Massey, G., Ehrensberger-Dow, M.: Commenting on translation: implications for translator training. The Journal of Specialised Translation. 16, 27–41 (2011).
15. Jääskeläinen, R.: Tapping the Process: an Explorative Study of the cognitive and affective Factors involved in Translating. Joensuun yliopisto, University of Joensuu (1999).
16. Désilets, A., Melançon, C., Patenaude, G., Brunette, L.: How Translators Use Tools and Resources to Resolve Translation Problems: an Ethnographic Study. Beyond Translation Memories Workshop. Ottawa (2009).
17. Bowker, L.: Using Specialized Monolingual Native-Language Corpora as a Translation Resource: a Pilot Study. Meta. 43, 631–651 (1998).
18. Delpech, E.: Un protocole d'évaluation applicative des terminologies bilingues destinées à la traduction spécialisée. Actes du 7ème Atelier Qualité des Données et des Connaissances, Évaluation des méthodes d'Extraction de Connaissances dans les Données. pp. 37–48. Brest, France (2011).
19. Jääskeläinen, R.: Translation Assignment in Professional vs. Non-Professional Translation: A Think-Aloud Protocol Study. The Translation Process. pp. 87–98. Candace Séguinot, Toronto (1989).
20. Leitjen, M., Van Waes, L.: Inputlog: New Perspectives on the Logging of On-Line Writing. Methods and Applications (Studies in Writing). pp. 73–94. Elsevier Science (2006).
21. Carl, M.: Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. Presented at the LREC (2012).
22. Babych, B., Hartley, A., Kageura, K., Thomas, M., Utiyama, M.: MNH-TT: a collaborative platform for translator training. Presented at Translating and the Computer 34 , London, UK November 29 (2012).
23. Carl, M.: Gaze Activity Patterns in Translation and Post-editing. Workshop on Future Directions in Translation Research. Knowledge Capital, Grand Front Osaka (2013).
24. Alves, F.: Triangulating Translation: Perspectives in process oriented research. Fabio Alves, Federal University of Minas Gerais (2003).
25. Josselin-Leray, A.: Place et rôle des terminologies dans les dictionnaires unilingues et bilingues. Etude d'un domaine de spécialité : volcanologie, Ph.D. Thesis, University Lyon II, (2005).