Reliable Innovation: A Tecchie's Travels in the Land of Translators

Alain Désilets [1], Louise Brunette [2], Christiane Melançon [2], Geneviève Patenaude [1]

Institute for Information Technology
National Research Council of Canada
Ottawa, ON, K1A 0R6, Canada
{alain.desilets; genevieve.patenaude}@nrc-cnrc.gc.ca

Département d'études langagières
Université du Québec en Outaouais
Gatineau, QC, J8X 3X7, Canada
{louise.brunette; christiane.melancon}@uqo.ca

Abstract

Machine Translation (MT) is rapidly progressing towards quality levels that might make it appropriate for broad user populations in a range of scenarios, including gisting and post-editing in unconstrained domains. For this to happen, the field may however need to switch gear and move away from its current *technology driven* paradigm to a more *user-centered* approach. In this paper, we discuss how ethnographic techniques like *Contextual Inquiry* could help in that respect, by providing researchers and developers with rich information about the world and needs of potential end-users. We discuss how data from Contextual Inquiries with professional translators was used to concretely and positively influence several R&D projects in the area of Computer Assisted Translation technology. These inquiries had many benefits, including: (i) grounding developers and researchers in the world of their end-users, (ii) generating new technology ideas, (iii) selecting between competing development project ideas, (iv) finding how to alleviate friction for important ideas that go against the grain of current user practices, (v) evaluating existing or experimental technologies, (vi) helping with micro level design decision, (vii) building credibility with translators, and (viii) fostering multidisciplinary discussion between researchers.

1 Introduction

Research and Development in Machine Translation (MT) has traditionally been very much a *technology driven* affair, fueled by a quest for the elusive grail of Fully Automatic High Quality Translation. Even 15 years after Church and Hovy's call to look for “good applications for crummy Machine Translation” (Church and Hovy, 1993), no application of MT have reached the kind of widespread adoption that one would hope for (with the possible exception of Translation Memories). This situation could however change rapidly. Indeed, recent anecdotal reports indicate that MT technology may finally have reached sufficient quality levels to make it useful for broad user populations in a range of scenarios, including gisting and post-editing in unconstrained domains. For this to happen however, the field may need to switch gear and adopt a more *user centered* paradigm. In this paper, we discuss how ethnographic techniques like Contextual Inquiry could help in that respect, by providing researchers and developers with rich information about the world and needs of potential end-users.

Contextual Inquiry[1] is a well known technique in Human Computer Interaction, where researchers observe and interview potential end-users while they are involved in their normal day to day work. One advantage of this technique, compared to other requirement elicitation methods, is that it is *generative*. Instead of just asking end-users what they *think* they need (which often turns out to be different from what they *actually* need), researchers aim for a deep understanding of important details of the end-user's world, including: top level goals, recur-

[1] Contextual Inquiry: http://en.wikipedia.org/w/index.php?title=Contextual_inquiry&oldid=190730351

rent workflow sequences, artifacts used, culture and values, and even physical environment. This intimate knowledge can then be leveraged to design new and possibly disruptive technologies and processes, which are nevertheless well grounded in actual end-user needs and context, and have therefore a greater chance of being adopted.

We report on our own experience in using this technique at the LOPLT (Laboratoire d'observation des pratiques langagières technologisées), and how it concretely affected several R&D projects in the area of Computer Assisted Translation (CAT).

2 The LOPLT Contextual Inquiry Study

LOPLT is a multidisciplinary team at the Language Technology Research Centre[2], comprised of researchers from both translation and technology fields. It aims at better understanding current workpractices of translators, and how new technologies could best support or augment them.

In this project, we realized early on that there was a lack of well-validated and actionable knowledge about the technological workpractices of translators. Although the act of translation has been thoroughly investigated through Think Aloud Protocols (see Jääskeläinen, 2002 for a review of this literature), these studies are not particularly useful for investigating how technology could better support the work of professional translators. For one thing, few of the studies were done in a normal professional translation work environment. Indeed, most of them involved subjects who were student translators or even, students learning a second language. Even those that looked at professional translator were generally carried out in an artificial controlled environment, as opposed to the subject’s actual work environment. In most of these artificial environments, subjects did not have access to any technological aids. On a different note, these studies all focus on high-level psycho-linguistic processes as opposed to lower level pragmatic work practices and processes. In particular, they did not look at how translators use technology in the course of their work. Because of those limitations we found the results of previous studies to be of limited use for making design decisions about CAT technology.

In order to fill this knowledge gap, we conducted a series of Contextual Inquiries with 11 translators coming from a broad range of work environments (home based freelancers, medium sized agencies, large government translation department, academia, and even amateur communities of volunteer translators). Each subject was interviewed in the context of carrying out two translation tasks: a *natural* and a *controlled* task (50 minutes each). In the *natural* task, we asked the subject to work on whatever document was in her in-tray at the time. The purpose of this task was to maximize the ecological validity of the data by making sure that we observed the subject working on a document which is representative of what he usually translates. In the *controlled* task, we asked all subjects to translate a same short document (a nontechnical newspaper article). The purpose of this task was to provide a common point of reference across all subjects.

During the interviews, the translator and interviewer's voices were recorded and the translator's screen was captured on video. The audio was later transcribed to text. Copies of the source documents being translated as well as the translations produced by the subjects were also collected. The audio, video and text documents collected during our contextual inquiries provide us with a detailed account of what the subjects did and why.

This rich data was analyzed as follows. First, time-synchronized verbatim transcriptions of the audio were created using the Transana video analysis software[3]. These transcriptions were augmented with notes describing important events that were visible on the screen capture, but were not apparent from what the translator or interviewers said. A *Grounded Theory* approach (Strauss and Corbin, 1998) was then used to identify recurrent themes and phenomenons from the ground up. During this *open coding* phase, we made sure that every passage of every transcript was analyzed by at least two researchers, one of them from the field of translation, and one from the field of technology. This was done in order to ensure that both perspectives were brought to bear on every data item we recorded. Often this analysis was done collaboratively by two or more researchers looking at the

[2] Language Technology Research Center: a joint center of the National Research Council of Canada, Université du Québec en Outaouais and Translation Bureau of Canada (http://www.crtl-ltrc.ca/).

[3] Transana: http://www.transana.org/.

same video together, and discussing what they saw and what it might mean.

Although we used an open coding approach, we also paid particular attention to pre-established categories which have been found helpful in usability methodologies like Contextual Design (Beyer and Holtzblatt, 1998) and Usage Centered Design (Constantine and Lockwood, 1999). These included: top level user goals, recurrent workflow sequences, physical or virtual artifacts, culture and values, and physical environment layout. Open coding also yielded additional categories which are more specific to the task of translation. For example, each episode of a transcript where the subject was trying to solve a particular translation problem[4] was coded along three categories. These were: PROBLEM-TYPE (ex: finding an equivalent for a term, understanding the meaning of a part of the source text), PROBLEM-RESOLUTION-APPROACH (ex: consult a Terminology Database, search in a Translation Memory) and LINGUISTIC-RESOURCE-EMPLOYED (ex: TERMIUM, TransSearch, Google). The Transana software was again used to tag specific sections of the transcripts and video with such codes, making it easy for us to view parts of the data that relate to specific themes.

3 Leveraging Contextual Inquiry to make research and development decisions

Researchers and developers are constantly faced with a myriad of decisions, some minor, some major, which can greatly impact adoption of the systems they build.

In the context of our Computer Assisted Translation projects, we have found the above Contextual Inquiry data to be invaluable to help us make such decisions in an informed and rational way. Our ultimate goal at LOPLT is to generate knowledge that can provide those same benefits to other researchers and developers, without requiring them to conduct their own field interviews with translators. We have however found this to be highly challenging. The most direct way to achieve this goal would be to grant access to the annotated and coded transcripts to any researcher who cares to look at it, but this would be in clear breach of confidentiality to the subjects who participated in the study. The alternative is for us to extract the most important and significant trends from the data, and present them in a short and concise compendium. However, we are finding that this is difficult to do in a generic way. For example, because we have a particular interest in tools that might help translators collaborate in a massively online, Wikipedia-like fashion (Désilets, 2007), we are particularly attentive to any detail related to that theme. But most of those may not be as interesting to researchers working on other types of technologies. Conversely, we might overlook details that seem irrelevant to us, but turn out to be important to other researchers.

Notwithstanding these challenges, we are currently working on such a generic summary of our observations, and it should become available in the next 12 months. For now however, this paper will limit itself to providing "teasers" which illustrate the different ways in which this data has helped us in our own CAT development and research efforts.

3.1 Grounding developers and researchers in the world of their end-users

One of the outcomes of our Contextual Inquiry work is a list of 120 well validated claims about the world of translators. Although many of them do not come as a surprise to people with a translation background, we have found that technology developers and researchers are often not aware of them.

Below is an example of the type of claims we are able to make based on our Contextual Inquiry:

> *"When translators consult a resource (e.g. Terminology Database, Translation Memory) to resolve a translation problem, they seem to care more about recall than precision. In other words, translators do not mind seeing a list of mostly poor suggestions, as long as it contains at least a few good ones. Translators are highly skilled at quickly scanning lists of potential solutions to a translation problem, and identifying which ones (if any) are most appropriate for their current situation."*

[4] By translation problem, we mean a word or expression which a subject had difficulty translating, and for which he had to consult various resources.

We have found that this type of information provides useful background that researchers and developers can use to at least form a general mental picture of their end-users. And as any usability practitioner will tell you, just getting a development team to start thinking about the end-user is half the battle.

3.2 Generating new technology ideas

We have also found Contextual Inquiry data to be a great source of inspiration for new technological ideas that are well grounded in the needs of translator. Indeed, in the course of this project, our team collectively generated a list of 30 ideas for technological innovations, all of which were directly triggered by something we read in the transcripts of our translator interviews. This represents a staggering average of 3 ideas per subject. Indeed, it has been rare for us to come out of an interview without at least one new idea in our head. Many of those are smallish ideas that only propose tweaks on existing technologies, but some are bigger and more disruptive innovations.

As an example, we noticed that, in the course of most of our 100 minute interviews, the translator manually carried out a search for aligned bilingual sentences using Google. We also found that, to find even a single pair of aligned sentences, this manual process required a minimum of 30 seconds. Some of the subjects we observed doing this already had access to a proper Translation Memory, but would turn to this strategy when they did not find any answers in the TM.

This inspired us to start a new project, called *WeBiText*, which aims at building a “Google of parallel search”, i.e., a large, heterogeneous Translation Memory based on parallel content mined from the Web (Désilets et al., 2008).

3.3 Selecting between competing project ideas

One of the curses of working in a R&D environment is that, in the course of a year, one is exposed to more “cool” technology ideas than one can possibly explore. We have found user observation data to be very useful for deciding which ideas are most likely to lead to technologies that translators will want to adopt.

As an example, one of the ideas we had before observing translators was the concept of a synchronized parallel translation editor, which would help the translator keep track of where she is in both the source and target text. One implementations we had contemplated consisted of a side by side view, with source text on the left and target on the right, where scrolling or clicking anywhere in the target side would automatically move the cursor accordingly in the source text (and vice-versa).

However, when we observed professional translators, we noticed that they were very good at orienting themselves around documents, and that they could, in a matter of a few seconds, easily locate the sentence in the source text that corresponds to a specific sentence in the target text. When we asked our subjects if they found this to be cognitively demanding, they tended to respond that it wasn't, and that orienting themselves around source and target text had become second nature to them.

This of course does not necessarily mean that the idea of a synchronized parallel translation editor is bad per se. It does however mean, that there is less evidence in our data to support that particular idea, compared to say, the WeBiText idea which we described previously.

3.4 Knowing how to alleviate friction for important ideas that go against the grain

Of course, important innovations must sometimes go against the grain of current practices in order achieve ground-breaking impact. Machine Translation, in particular, may very well be an example of such an important disruptive technology. But even in this type of situation, it is still important to be aware of that friction and to understand its exact nature, so that one can take measures to alleviate it wherever possible.

As an example, one project idea we are currently investigating is one which we call *WikiTerm* (Désilets et al., 2007). The concept is that of a large, open, Wikipedia-like terminology database covering all domains and languages. While the recent success of Wikipedia makes this a very compelling idea, our user observation data leads us to believe that it may be fighting an uphill battle against important current practices and values of translators.

Indeed, with a few exceptions, none of the subjects we observed consulted any of the open terminology resources that already exist (ex: Wikipedia, Wiktionary, OmegaWiki, ProZ). Also, translators often talked about the importance of using trustworthy sources when looking for solutions to translation problems. When discussing this, some of our users even explicitly mentioned Wikipedia as an example of a resource that might contain less-than-reliable information. Altogether, these observations tell us that a WikiTerm might be up against a significant perception of unreliability on the part of translators.

At the same time, we have also noticed that when translators do not find what they need in reliable sources, they have no qualms about searching in less reputable ones (ex: doing a Google search on the whole Web). And as we mentioned earlier, translators are highly skilled at quickly sifting grain from chaff in lists of suggested solutions. Therefore, it may be that a WikiTerm can still achieve adoption by translators, as long as it provides features to address the perception of unreliability. For example, the system might provide automatic ratings of terminology entries, based on metrics that have been found to correlate with quality in Wikipedia, such as: age of the entry, number of edits, number of unique contributors, amount of discussion on the talk page, and reputation of contributors (Wilkinson and Huberman, 2007; De Alfaro and Adler, 2007).

3.5 Evaluating existing or experimental technologies

We have also found our Contextual Inquiry data to be useful for evaluating the usefulness of particular technologies, whether they be established ones or experimental prototypes.

For example, in the context of the WikiTerm project, we needed to evaluate the extent to which existing wiki resources such as Wikipedia, Wiktionary and OmegaWiki, already meet the linguistic needs of professional translators. Using data extracted from our transcripts, we were able to demonstrate that, in their current state, they lack sufficient coverage of typical translation problems. This analysis was based on a list of 59 instances of translation problems encountered by our subjects. Using qualitative data from our transcripts, we were also able to show that the user interface of existing wiki resources does not make it easy to carry out key translation related tasks such as: finding an appropriate solution for a translation problem, adding a new solution for a problem, and assessing the trustworthiness of a particular solution to a problem (Désilets et al., 2007).

The list of translation problems encountered by our subjects has also been useful to evaluate coverage of the WeBiText system (Désilets et al., 2008), and how it is affected by the day-to-day tweaks that we effect on the code, in an attempt to improve it. This ensures that we are always moving in the direction of better coverage.

It is interesting to contrast the kind of data we collect through Contextual Inquiry, to that which can be collected through log analysis on existing CAT tools. While analyzing the logs of a particular tool has the advantage of providing larger quantities of data, it may be strongly biased towards translation problems which are particularly appropriate for that one tool. Indeed, our interviews with translators clearly reveal that they know exactly the strengths and weaknesses of each tool, and that that they do not waste time submitting queries to tools which are not appropriate for that particular type of problem. Thus, if one was to look at the log of a Terminology Database like TERMIUM[5], one might conclude that problems related to phraseology and idiomatic expression (ex: "he is out to lunch") are uncommon. Yet, if you looked in the logs of a general domain bitext like TransSearch (Macklovitch et al., 2000), you might conclude that on the contrary, terminology problems are rare, compared to those related to phraseology and idiomatic expressions. In contrast, our Contextual Inquiry data captures the whole range of problems encountered by translators in their day to day work. It may therefore serve as a better basis for evaluating the overall usefulness of a particular tool for translators.

3.6 Helping with micro level design decisions

When developing new technologies, one is faced with a myriad of small decisions which together, can significantly affect adoption of the system. Often these decisions are taken in the absence of data or knowledge about the target users. Not surpris-

[5] TERMIUM: The Terminology Database of the Government of Canada (http://www.termiumplus.gc.ca/)

ingly, this often leads to long debates in the development teams, about what is best for the user. We have found that our Contextual Inquiry data helps us make those decisions in a more informed and rational way. Whenever we find ourselves arguing about what the system should do for the end-user, we can usually resolve the dispute by discussing the issue in terms of things we have *actually seen* in our data, instead of things that we *hypothesize* about the end-user.

For example, at some point in the context of the WeBiText project, we were faced with a choice between working on improving the sentence alignment algorithm, or increasing the size and variety of the corpus used by the system. The issue with alignment was that the system often presented sentence pairs that were in fact not aligned. Most of the time, this was due to the fact that the web pages containing those sentences were not parallel texts to start with. This is one of the technical challenges of building a Translation Memory based on open web content, as opposed to the carefully controlled parallel texts which are typically poured into more conventional systems, and we felt that WeBiText needed to be able to deal better with this reality.

At the same time, we also wanted to crawl more parallel web sites in order to increase the size and variability of the corpus, and we did not have sufficient human resources to carry out work in both those areas at once.

We found we could choose between those two directions more confidently by turning to our user data. Indeed, given the claim we made earlier to the effect that translators seem to care more about recall than precision, we felt confident that increasing the size and breadth of the corpus would provide more immediate value to the end-users than improving the accuracy of sentence alignment.

3.7 Building credibility with translators

As a result of having participated in this Contextual Inquiry study, the main developer on the LOPLT team (Désilets) finds he can now interact much more constructively with professional translators. We have found that a developer equipped with this sort of knowledge can discuss technological innovations with translators without being perceived as threatening, or even worse, being disregarded as yet another naïve, uninformed tecchie. This can put him in an ideal position of influence, which he can leverage to push for new, and possibly disruptive technological ideas, yet do it in a way that is more likely to lead to adoption.

3.8 Fostering multidisciplinary discussion between researchers

A final benefit of this Contextual Inquiry approach, is that it can be a great tool for fostering collaborative, multidisciplinary research. Indeed, in our study, we have opted for a truly multidisciplinary approach, where interviews are carried out, transcribed and analyzed collaboratively by teams comprised of researchers from both the worlds of translation and technology. This has resulted in very interesting discussions between researchers of the two disciplines, and some of the more valuable insights and technology ideas could not have emerged without this sort of interplay.

4 Conclusion and Future Work

In short, we have found ethnographic methods like Contextual Inquiry to be very useful for grounding Language Technology R&D in the actual needs of translators. We believe these techniques can play an important role in the coming years, and help move MT technology out of the labs, and into the hands of end-users. This is particularly important given that MT is currently at a critical crossroad where it might become appropriate for use by broad user populations in a range of scenarios, including gisting and post-editing in unconstrained domains.

We plan to continue using this type of technique to advance knowledge along three axes. The first direction is to continue investigating how translators work. This will involve conducting additional Context Inquiries with translators in work environments that we have not yet covered (ex: countries other than Canada, highly technical translation), as well as coming up with ways to summarize our observations so that they can be useful to other researchers and developers in the area of Computer Assisted Translation.

Secondly, we will continue to use this data ourselves to advance our own CAT projects such as WeBiText and WikiTerm.

Finally, we plan to conduct a whole different series of Contextual Inquiries with revisers and post-editors as subjects. This seems particularly relevant and timely, given that MT is rapidly progressing towards a level of quality sufficient to make post-editing a reasonable and economical alternative to translation from scratch. Yet, empirical studies of MT post-editing have focused almost exclusively on evaluating the productivity gains of this new paradigm (Krings, 2001, Guerra, 2004). While this is for sure an important question, it is more focused on the needs of managers than those of translators and post-editors. We believe it is equally important to understand how post-editors carry out their work, and how MT and post-editing technologies could be better tailored to meet their needs and culture.

References

Beyer H., Holtzblatt K. *Context Design: A Customer-Centered Approach to Systems Designs.*, Morgan Kauffman. 1998

Church, K. W., Hovy, E. H. 1993. *Good applications for crummy Machine Translation*. Machine Translation, 8:239-258.

Constantine, L. L., Lockwood, L. A. D., 1999. *Software for use: a practical guide to the models and methods of usage-centered design*. ACM Press/Addison-Wesley Publishing Co. New York, NY, USA

De Alfaro and L., Adler, T, 2007. *A Content-Driven Reputation System for Wikipedia*. Proceedings of WikiMania 2007, Taipei, Taiwan, August 3-5, 2007.

Désilets, A., Farley, B., Stojanovic, M. 2008. *WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content*. In Proc. of Translating and the Computer 30, London, UK, 27-28 November 2008.

Désilets, *Translation Wikified: How will Massive Online Collaboration Impact the World of Translation?* Proc. of Translating and the Computer (29). November 29-30, 2007. London, United Kingdom.

Désilets, A., Barrière, C., Quirion, J., 2007. *Making WikiMedia resources more useful for translators*. Proceedings of Wikimania 2007, The International Wikimedia Conference. Taipei, Taiwan. August 3-5, 2007. NRC Publication Number: NRC 50383.

Guerra, L. 2003. *Machine Translation: an Imperfect but Evolving Technology*. In special supplement of Multilingual Computing and Technology magazine, Number 62, March 2004.

Jääskeläinen, R. 2002. *Think-aloud studies into translation: An annotated bibliography*. Target, Volume 14, Number 1, 2002 , pp. 107-136(30) .

Krings, H., translated and edited by Koby, G.S. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*. Kent State UP, Ohio, USA

Macklovitch, E., Simard, M., Langlais, P, 2000. *TransSearch: A Free Translation Memory on the World Wide Web*. In Proceedings of the LREC 2000, Athens, Greece

Strauss, L., Corbin, J. M. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* Sage Publications Inc, ISBN 0803959400, 9780803959408.

Wilkinson, D. M., Huberman, B. A., 2007. *Cooperation and quality in Wikipedia.* In Proc. of the 2007 international symposium on Wikis, Montréal, Canada, Oct 21-23, 2007