Towards a Research Infrastructure for Translation Studies

Lars Ahrenberg

Department of Computer and Information Science
Linköping University

lars.ahrenberg@liu.se

Abstract

In principle the CLARIN research infrastructure provides a good environment to support research on translation. In reality, the progress within CLARIN in this area seems to be fairly slow. In this paper I will give examples of the resources currently available, and suggest what is needed to achieve a relevant research infrastructure for translation studies. Also, I argue that translation studies has more to gain from language technology, and statistical machine translation in particular, than what is generally assumed, and give some examples.

Keywords: language technology, translation studies, research infrastructure, CLARIN.

1 Introduction: translation studies

Translation studies is a field of research that aims to understand the processes and products of translation. It is a relatively new field of the Humanities that have seen a rapid development after the Second World War. In this short period of time it has changed its focus several times and developed in many different directions [6]. It has been approached from many disciplines including literary studies, linguistics, cultural studies, sociology, and cognitive science. The quest for empirical grounding has also meant that some scholars have taken an interest in corpus linguistics and argued for the usefulness of corpora and computational tools in the study of translation, in particular for investigations into 'translation universals' and features of 'translationese' (e.g. [1, 4]).

For good reasons, literature is usually regarded as the genre that is the least suitable for machine translation. From the perspective of translation studies, however, literature might be the genre that could benefit the most from tools and methods used in machine translation. One reason is that both the content and the style of the texts are important in the study of literature. While culture and norms are emphasized as explanatory factors in recent theories of translation, the author's style and the rendering of the source text in the target language are still very much in the researchers' focus of attention.

It would appear, though, that in spite of the efforts spent on arguing for the relevance of translation corpora in translation studies, large-scale studies of translation based on parallel or comparable corpora are quite scarce. I can only guess why this is so. Maybe the relevant university departments, often departments for the study of literature, do not have the necessary funds for computational resources and personnel, maybe the researchers do not have the training or interest in corpus analyses, or don't find it worthwhile. At the same time, though, there is a growing interest in literary history itself to apply methods from corpus linguistics and statistical analysis to hugh corpora of literary works [3].

In my view, the arguments for corpora in translation studies still hold strong and CLARIN could be a suitable environment for demonstrating it. This is true not least in a Swedish context, where at least the concept of a corpus is well understood [2]. But language technology has actually more to offer than tools and annotated corpora. After a look at CLARIN's goals and achievements so far as it relates to translation and translation studies, I will then outline what I think language technology, and translation technology in particular, can contribute to translation studies.

2 Some relevant CLARIN resources

CLARIN's general infrastructure framework provides for federated access to data and resources that actually reside in different centers, provided you are a researcher associated with a CLARIN center. This is all very well.

The resources currently available for the study of translations are quite limited. The Virtual Language Observer[1] gives few matches to search words such as 'translation' or 'translation studies' mostly producing theses rather than resources. A search for 'parallel corpora' gave 53 results[2] including many duplicates, and the majority being references to tools rather than corpora. Some of the tools available, e.g., for word and sentence alignment, are of

[1] http://catalog.clarin.eu/vlo/

[2] 2014-10-24, actually an increase from 38 6 weeks earlier

course useful in this context, but if there are no corpora, and no suggested workflows, they are of little use to a researcher, especially one who is unfamiliar with such tools.

Altogether I found references to 12 different parallel corpus collections. Only one of them, the ECI Multilingual Text, has Swedish parallel data. More parallel corpora may of course be added in the future, but a problem with the current resources is that they do not result from coordinated efforts. Some of these corpora contain resources for several related language pairs including translations of the same source text, but that is then only on a small scale, using a single source text or extracts from a limited number of different source texts.

As for corpus search tools I haven't been able to find one that is currently offering federated search in parallel corpora. The Text Laboratory at the University of Oslo is developing a new version of their corpus search system, **Glossa**, that will support federated corpus search. Support for search in parallel or multilingual corpora is said to be an item for the future. **WebLicht**, developed at the University of Tübingen, is a system for search and annotation that allows a user to configure his or her own tool chain for a specific purpose, but so far only for monolingual annotation. Most of the corpora available for search are German, a few have texts in different languages, but there is currently no support for parallel search results. **Keeleveeb**, developed by the Estonian company "Filosoft", does allow for search in bi-lingual resources and in multiple resources at the same time. So far, all bilingual resources available are dictionaries, while available corpora are monolingual Estonian.

The **NoSketch Engine** [5], a thinner version of the commercial Sketch Engine, but with support for federated search and parallel corpora, is apparently in use at the LINDAT/CLARIN Centre for Language Research Infrastructure, though there is no description of it in the CLARIN VLO. Thus, the current situation as regards resources for translation studies leaves quite a lot to be desired.

3 Research questions in translation studies

The research questions in translation studies are many and varied. Very often, however, they concern comparisons, for example comparisons between source and target texts, comparisons of different translations, or of different translators, or of translations with original texts.

While sometimes only one source text is of interest in a particular study, it is more common that different translations of it, not to say ALL its different translations into a given target language, are included. The goal may be to compare the translations for quality, or to compare different strategies in solving translation problems relating to specific phenomena in the source text. At other times translations of different source texts by the same translator is the focus of research, e.g., in order to characterize the translator's 'voice'. But this inevitably involves a comparison with other existing translations, produced by other translators.

Thus, a corpus collected for the study of a particular issue in translation studies, normally consists of many texts, that are to be accessed, studied and compared from a number of perspectives. If the question is very general, pertaining to such matters as translation universals or translation norms in a given culture at a certain time period, both source texts and translations need to be numerous. This suggests that accessing, annotating, searching

and reorganizing these texts would be vastly more efficient on the desk-top compared to the desk.

4 What language technology can contribute

From the above it is clear that a very important possible contribution from CLARIN is making parallel and comparable texts available for search. While texts and translations that are protected by copyright will, as usual, be hard to come by, there are plenty of classical literary works that are no longer copyright-protected and for which copyable translations should also exist in abundance. To harvest them into CLARIN, however, requires collaboration and concerted efforts among interested researchers from different centers.

Of course, to make this data useful for translation studies, corpus and language technology tools for processes such as part-of-speech tagging, lemmatization, alignment, parallel concordancing and search are required. Such tools nowadays exist in large numbers for many languages but the problem for CLARIN I suppose is to make them communicate well with one another. There is also a problem of scale. An integrated environment such as *ParaConc* allows up to four parallel texts in the system at one time[3]. A large-scale project in translation studies might involve several source texts with over ten translations per source. This puts different requirements on underlying representations, storage, and formatting of search results.

In addition, I believe that language technology can benefit translation studies by introducing supplementing methodologies. Translation studies seldom bother to quantify its data and while examples are analyzed with ingenuity and precision one often wonders how much of all relevant data these example cover, and about the possible existence of data that speak against a tentative hypothesis or conclusion. Thus, computational tools and resources could help translation studies acquire and utilize more of statistical methods. This macro-analytic perspective on texts, presented in [3] for the study of literature, can be applied to the study of translations as well.

Translation studies and machine translation research share a common interest in explaining translations. Machine translation systems, and statistical systems in particular, predict translations, but in so doing they also supply an explanation for how the translation came about. A limitation with current statistical MT is that all models are generated from textual data only, while in translation studies individual, contextual, and cultural factors are also taken into account. But this is no limitation in principle. Models could well be developed that take contextual factors into account and are able to distinguish translations performed by different translators, or translations that are more or less in line with different translation norms.

Whether researchers in translation studies find value in probabilistic models or not, they are often forced to limit their conclusions to a single work, a limited range of constructions, or hedge their conclusions with reference to the scarcity of textual data on which they are based. Open common resources and language technology certainly have the potential to overcome that limitation.

[3] www.athel.com/paraflyer2.pdf

References

[1] Mona Baker. Corpora in translation studies. *Target*, 7(2):223–244, 1995.

[2] Elisabeth Bladh and Magnus P. Ängsal, editors. *Översättning, stil och lingvistiska metoder*. Studia Interdisciplinaria Linguistica et Litteraria. Göteborgs Universitet, Göteborg, Sweden, 2013.

[3] Matthew L. Jockers. *Macroanalysis: Digital Methods & Literary History*. University of Illinois Press, Urbana/Chicago/Springfield, 2013.

[4] Sara Laviosa. *Corpus-Based Translation Studies: Theory, Findings, Applications*. Rodopi, Amsterdam/New York, 2002.

[5] Pavel Rychlý. Advance search in clarin text corpora. In *Extended abstract. CLARIN Annual Conference (CAC2014), Soesterberg, the Netherlands*, 2014.

[6] Mary Snell-Hornby. *The Turns of Translation Studies*. John Benjamins, Amsterdam/Philadelphia, 2006.