Chapter 1

Introduction

Stella Neumann
IFAAR, RWTH Aachen

Silvia Hansen-Schirra

Oliver Čulo
Johannes Gutenberg-Universität Mainz in Germersheim

1 Parallel corpora in Translation Studies

Parallel corpora, i.e. collections of originals and their translations, can be used in various ways for the benefit of Translation Studies, Machine Translation, Linguistics, Computational Linguistics or simply the human translator. In Computational Linguistics, translation corpora have been employed for Machine Translation but also for term extraction, word sense disambiguation etc. as early as the 1980s (important milestones being Nagao 1984 and Brown et al. 1990). One of the early electronic resources is the Canadian Hansard, which was initially used for implementing sentence alignment (Gale & Church 1991), a task that is now a standard feature of applications such as translation memories. Moreover, parallel corpora are used as data basis for multilingual grammar induction, automatic lexicography and many other tasks in information extraction and language processing across different languages.

In Translation Studies, the focus is more on identifying features that distinguish translations from original texts. From this perspective, the main research interest lies in the detection of patterns of (inevitable) modifications introduced by the translator(s) along the way in terms of local solutions, added information or even larger changes in the register of the text. These modifications may be individual to a given translation task or a translation pair but they may also instantiate typical features of translated text that make translations different from non-translated texts in a wide range of linguistic features. The investigation of

Stella Neumann, Silvia Hansen-Schirra & Oliver Čulo. 2017. Introduction. In Silvia Hansen-Schirra, Stella Neumann & Oliver Čulo (eds.), *Annotation, exploitation and evaluation of parallel corpora*, 1–8. Berlin: Language Science Press. DOI:10.5281/zenodo.283408

corpora is an obvious method to detect these distinctive properties of translations empirically and has been employed since the 1990s as witnessed by Baker (1993; 1996); Johansson & Ebeling (1996) and more recently by Hansen (2003); Teich (2003); Mauranen & Kujamäki (2004) and Hansen-Schirra, Neumann & Steiner (2012). Furthermore, parallel corpora are used as reference works for translation teaching and in professional translation settings since they enable quick and interactive access to translation solutions (e.g. translation memories).

Exchange between the Translation Studies and the Computational Linguistics communities has traditionally not been very intense. Among other things, this is reflected by the different views on parallel corpora. While Computational Linguistics does not always strictly pay attention to the translation direction (e.g. when translation rules are extracted from (sub)corpora which actually only consist of translations), Translation Studies is amongst other things concerned with exactly comparing source and target texts (e.g. to draw conclusions on interference and standardisation effects). However, there has recently been more exchange between the two fields – especially when it comes to the annotation of parallel corpora. This special issue brings together the different research perspectives. Its contributions show – from both perspectives – how the communities have come to interact in recent years.

With issues of the creation of large parallel data collections including multiple annotations and alignments largely solved, the exploitation of these collections remains a bottleneck. In order to use annotated and aligned parallel corpora effectively, the interaction of the different disciplines involved addresses the following issues:

- Query tools: We can expect basic computer literacy from researchers nowadays. However, the gap between writing query or evaluation scripts and program usability is immense. One way to address this is by building web query interfaces. Yet, in general, what are the claims and possibilities for creating interfaces that address a broader public of researchers using multiply annotated and aligned corpora? An additional ongoing question is the most efficient storage form: are database formats superior to other formats?

- Information extraction strategies: The quality of the information extracted by a query heavily depends on the quality of the annotation of the underlying corpus, i.e. on precision and recall of annotation and alignment. Furthermore, the question that arises is how we can ensure high precision and recall of queries (while possibly keeping query construction efficient).

What are the strategies to compose queries which produce high-quality results? How can the query software contribute to this goal?

- Corpus quality: Several criteria for corpus quality have been developed (e.g. in the context of standardisation initiatives). Quality can be influenced before compilation by ensuring the balance of the corpus (in terms of register and sample size), its representativeness etc. Also, inter-annotator agreement and – to a lesser extent – intra-annotator agreement are an issue. But, how can we make the corpora thus created fit for automatic exploitation? This involves issues such as data format validity throughout the corpus, robust (if not 100% correct) processing with corpus tools/APIs and the like. What are relevant criteria and how can they be addressed?

- Corpus maintenance: Beyond the validity of the data format, maintenance of consistent data collections is a more complex task, particularly if the data collection is continually expanded. A change of the annotation scheme entails adjustments in the existing annotation. Questions to this end include whether automatic adjustment is possible and how it can be achieved. Maintenance may also involve compatibility with and/or adaptations to new data formats. How can we ensure sustainability of the data formats?

A colloquium held at the Corpus Linguistics 2009 Conference at the University of Liverpool was concerned with the interface between the requirements of linguists and Translation Studies working with parallel corpora and computational linguists providing the tools and exploiting the corpora for their purposes. In this sense, it was closely related to and a continuation of the workshop "Multilingual Corpora: Linguistic Requirements and Technical Perspectives" held at the Corpus Linguistics 2003 Conference at Lancaster University (see Neumann and Hansen-Schirra Neumann & Hansen-Schirra 2003).

The present special issue is a collection of contributions arising out of this Colloquium. In what follows we outline the contributions responding to some of the questions posed above. The volume sets off with a focus on annotation, alignment and query on the syntactic level: Volk, Marek and Samuelsson discuss a trilingual parallel treebank, the Stockholm Multilingual Treebank SMULTRON. The ultimate purpose of the resource is its exploitation for Machine Translation, a typical application scenario for parallel treebanks. Interestingly, the resource only consists of translations in the three languages English, German and Swedish. The authors discuss solutions for some important questions in querying the treebank, thus focussing on an issue in working with parallel corpora that typically only arises at a later stage of corpus construction but that is not the least trivial.

In their contribution, Vintar and Fišer discuss the exploitation of multilingual resources – and translations in particular – for a monolingual computational linguistic task, the construction and enrichment of the Slovene WordNet. They turn the problem of a lesser-studied language into an advantage in drawing on the rich body of translations existing for Slovene. At various stages of their work, parallel corpora are used to disambiguate word senses with the help of translations – making use of a typical feature of translation, namely settling on one interpretation of ambiguous items in the source text – as well as to extract a bilingual lexicon of word-aligned items in order to enrich the resource with domain-specific lexical items. Vintar and Fišer show how monolingual resources can be successfully exploited with the help of parallel corpora that contain the required information.

Fantinuoli's contribution demonstrates an even more practice-oriented exploitation of corpora, both monolingual and parallel. Fantinuoli describes the design of a software, InterpretBank, which assists conference interpreters in all stages of their work. Based on Baroni and Bernardini's Baroni & Bernardini (2004) BootCat mechanism, it harvests the web for domain-specific documents given a set of search terms, performs term extraction on them and uses additional resources, e.g. Wikipedia or bilingual online dictionaries, to propose definitions, translations, collocations and keyword-in-context information. All available modules, for harvesting, management and retrieval, are adapted to the specific needs of interpreters, reducing the time needed for preparation and allowing for efficient retrieval while interpreting. A pilot module adds the possibility to include parallel resources, e.g. translation memories or the OPUS corpora, in the preparation phase.

The contribution by Čulo, Hansen-Schirra, Maksymski and Neumann revisits a more theory-oriented topic. It discusses the analysis of the bilingual CroCo Corpus, a richly annotated and aligned corpus of English and German translations and originals, with respect to a translation-specific research question. It exemplifies the exploitation of a resource that comes close to a parallel treebank for a research question that has a long history in Translation Studies, namely the study of shifts (e.g. Vinay and Darbelnet Vinay & Darbelnet 1958, Catford Catford 1965 etc.). The goal of this contribution is a heuristic identification of shifts in translation that can then be interpreted as properties of translations. While the main aim of the study is to advance empirical knowledge in the field of Translation Studies, it also has some clear implications for computational handling of translation shifts – for instance, in Machine Translation.

The translation-related research question investigated by Čulo et al. sets the scene for the final paper in this special issue: Alves and Vale introduce an innovative approach to adopting a corpus perspective on psycholinguistic research into the translation process. The authors describe LITTERAE, a computer tool that allows annotating linear representations of the process of producing a translation of a source text. They then proceed to discuss quantitative findings yielded with LITTERAE which suggest certain patterns in target text production. The paper provides a highly interesting way of reducing the gap between corpus-based and process-oriented investigations of translations. It thus rounds off this special issue with a perspective beyond Corpus Linguistics.

The articles in this special issue address a number of the issues discussed above: Vintar and Fišer are concerned with information extraction from various multilingual resources, whereas Čulo et al. exemplify the linguistic interpretation of parallel data on the basis of a heuristic information extraction procedure. Information extraction as well as its interpretation is also exemplified in Alves and Vale's study. Questions of corpus querying are also a major concern of Volk et al, as well as corpus quality, in particular annotation quality. The latter is also addressed by Padó. The only area of interest not covered by one of the contributions is the maintenance of continually expanding resources. This is an area addressed by work in the area of sustainability of corpora, for instance in the framework of the European CLARIN project [1] and similar national initiatives.

2 Acknowledgements

We believe that this volume provides a good overview of some important issues of the operation of parallel corpora, not only focussing on computational issues but also giving insight into the linguistic analysis of translations. If successful, this will not be least thanks to the efforts the reviewers put into providing feedback to the authors and thus ensuring the quality of this issue. The reviewers were: Sabine Bartsch (University of Technology, Darmstadt), Stefan Evert (University of Osnabrück), Johann Haller (IAI, Saarbrücken), Kerstin Kunz (Saarland University, Saarbrücken), Anke Lüdeling (Humboldt University, Berlin), Reinhardt Rapp (University of Mainz, Germersheim), Josef Schmied (University of Technology, Chemnitz), Erich Steiner (Saarland University, Saarbrücken), Elke Teich (Saarland University, Saarbrücken), Mihaela Vela (German Research Center for Artificial Intelligence, Saarbrücken) and Andreas Witt (Institute for the German Language, Mannheim).

We are also grateful to the authors for their contributions and collaboration.

[1] http://www.clarin.eu/external/ (last accessed 9 March 2010)

References

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam & Philadelphia: John Benjamins.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and translation. Studies in language engineering in honour of Juan C. Sager*, 175–186. Amsterdam: John Benjamins.

Baroni, Marco & Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC2004*, 1313–1316. Lisbon: ELDA.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2). 79–85.

Catford, John C. 1965. *A linguistic theory of translation: An essay in applied linguistics*. Oxford: Oxford University Press.

Gale, William A. & Kenneth Ward Church. 1991. Identifying word correspondences in parallel texts. In *Speech and natural language, proceedings of a workshop held at pacific grove, california, usa, february 19-22. 1991*, 152–157. Morgan Kaufmann.

Hansen, Silvia. 2003. *The nature of translated text: An interdisciplinary methodology for the investigation of the specific properties of translations*. Saarbrücken: DFKI/Universität des Saarlandes.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2012. *Cross-linguistic corpora for the study of translations: Insights from the language pair English-German*. Berlin: de Gruyter.

Johansson, Stig & Jarle Ebeling. 1996. Exploring the English-Norwegian parallel corpus. In Carol E. Percy, Charles F. Meyer & Ian Lancashire (eds.), 3–16. Amsterdam: Rodopi.

Mauranen, Anna & Pekka Kujamäki (eds.). 2004. *Translation universals*. Amsterdam & Philadelphia: John Benjamins.

Nagao, Makoto. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn & Ranan Banerji (eds.), *Artificial and human intelligence*, 173–180. Amsterdam: North Holland.

Neumann, Stella & Silvia Hansen-Schirra (eds.). 2003. *Proceedings of the Workshop on Multilingual Corpora, Linguistic Requirements and Technical Perspect-*

ives. Corpus Linguistics Conference 2003. Lancaster. http://www.coli.uni-saarland.de/conf/muco03/Proceedings.htm.

Teich, Elke. 2003. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Berlin & New York: Mouton de Gruyter.

Vinay, Jean-Paul & Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: Méthode de traduction*. Paris: Didier.