TransBank: Metadata as the Missing Link between NLP and Traditional Translation Studies

Michael Ustaszewski **Andy Stauder**
University of Innsbruck, Department of Translation Studies
michael.ustaszewski@uibk.ac.at andy.stauder@uibk.ac.at

Abstract

Despite the growing importance of data in translation, there is no data repository that equally meets the requirements of translation industry and academia alike. Therefore, we plan to develop a freely available, multilingual and expandable bank of translations and their source texts aligned at the sentence level. Special emphasis will be placed on the labelling of metadata that precisely describe the relations between translated texts and their originals. This metadata-centric approach gives users the opportunity to compile and download custom corpora on demand. Such a general-purpose data repository may help to bridge the gap between translation theory and the language industry, including translation technology providers and NLP.

1 Introduction

The breakthroughs in machine learning of the past years have made statistical approaches the dominant paradigm in Natural Language Processing (NLP) in general and in Machine Translation (MT) in particular. As a consequence, translation data such as parallel corpora, translation memories (TM) and postediting data have become of utmost importance to the increasingly (semi-)automated translation industry. The "datafication of translation" (Wang, 2015; van der Meer, 2016) observed in the industry is no exception in the era of big data.

Driven by the irreversible trend towards translation automation and the resulting demand for more and more data, several industry-led projects for the large-scale collection of translation data have been launched (see section 2.1). These projects mainly aim to boost productivity in the language industry and are therefore oriented towards the needs of language service providers (LSPs), translation technology developers and, to a lesser extent, of translators. While such data repositories are without doubt useful not only for the translation industry but also for a range of lines of research in Translation Studies (TS), we believe that they do not fully accommodate the needs of the latter. The reason for this is that although they cover a large variety of language pairs, subject domains and text types, they pay insufficient attention to metadata about the included texts and text segments. After all, the object of study in TS is not limited to translation products but extends to translation processes and the linguistic, cognitive, socio-cultural, socio-economic and technological factors that influence translation. Translation data without metadata is only of limited use for research into the complex nature of translation because metadata describe how, when and under what circumstances a given text has been produced. For the training of NLP systems and for productivity gains in the translation industry, these questions are less of a concern, at least for the present.

In the academic discipline of TS, the trend towards data-driven methods is far less pronounced than in the translation industry and among translation technology providers. A bibliometric analysis has revealed that corpus-based research is gaining momentum in TS but still noticeably trails behind the most popular lines of research (Zanettin et al., 2015). To avoid losing relevance to the translation industry and translation technology providers, TS needs to continue adopting and developing more rigorous, empirically sound and objective methods of scientific inquiry. Furthermore, as Way and Hearne (2011) highlighted, in order to advance MT, we need to make explicit the phenomena that occur in real translated data and model these phenomena more effectively in a joint effort of MT developers, translators and translation scholars. A closer collaboration between these major stakeholders in translation would be a big step towards narrowing the yawning and growing gap between translation theory and translation practice, identified i.a. by Sun (2014), and to eventually provide innovative solutions to unresolved problems in MT. For this endeavor to be successful, the availability of high-quality, labelled real-world translation data is paramount.

Against this background, we aim to develop a novel resource that equally meets the requirements of all translation stakeholders, be it translation scholars, translation technology developers, LSPs, translators, translation trainers and trainees, or researchers from neighboring disciplines interested in translation and translated language. The key to such a general-purpose collection of translation data is a precise and comprehensive set of metadata labels that help capture the relation between translated texts and their originals, including the circumstances under which the translations were produced. Translated target texts (TTs) and their source texts (STs) will be stored in a freely available, open-ended and growing bank of translation data – hence the project name TransBank. The dynamic and metadata-centric approach is expected to combine the advantages of pre-existing data collections and to give users the opportunity to compile and download translational corpora on demand, tailored to the requirements of specific research questions or applications.

In this article, which is mainly meant to be a vision paper, we aim to outline the goals, concept, and planned architecture of TransBank, whose development will start in September, 2017.

2 Related Work

Given the vast amount of bi- and multilingual corpora, an overview of related work is necessarily selective and incomplete.[1] For the sake of simplicity, existing repositories can be grouped into resources oriented mainly towards the language industry and/or NLP one the one hand, and towards academia on the other. The following two subsections summarize, in accordance with this distinction, several selected resources that share some but not all features with TransBank.

2.1 Industry- and NLP-Oriented Resources

The TAUS Data Cloud[2] is the largest industry-shared TM repository, exceeding 79 billion words in more than 2,200 language pairs. Its main aim is to boost the language service sector, which is why it focuses mainly on economically relevant aspects of the industry. It is mainly used to train MT systems, in both industry and academia. Users can download data if they have enough credits, which can be either bought or earned by uploading one's own translation data. Data is searchable on the basis of a relatively small set of metadata labels: source and target language, domain, content type, data owner, and data provider.

MyMemory[3], operated by the LSP *Translated*, also claims to be the world's largest TM. It comprises the TMs of the European Union and United Nations as well as data retrieved from multilingual websites. Its main target groups are, again, the language industry, NLP developers and translators. The download of TMs is free of charge. The search options based on metadata are limited to language pairs and subject domains.

The European Parliament Proceedings Parallel Corpus (EuroParl, Koehn, 2005) has been highly influential in statistical MT due to its large size, multilingual and sentence-aligned architecture. However, it includes a rather limited range of subject domains and text types. Metadata are very scarce (languages and language directions, speaker turns), which limits its usefulness for many potential research questions outside NLP.

The European Commission made a number of its large multilingual TMs and corpora from the domains of politics, law and economy freely available[4]. These resources are often used to train MT systems and to feed TMs in the industry. They, too, provide rather scarce metadata.

Finally, OPUS (Tiedemann, 2012) is a freely available, growing collection of pre-existing, automatically enriched (e.g. POS-tagged) corpora and TMs retrieved from the web. Its main assets are size, the large number of language pairs, textual diversity in terms of subject domains and text types, variation of translation modes (written translation, localization, and subtitles), as well as its open-ended nature. On the downside, metadata are scarce: it provides varying but small label sets depending on the respective processed corpus that is accessed through the OPUS interface, e.g. the EuroParl corpus. Due to its size, variation and free availability, it has proved useful for NLP and MT systems.

[1] For a more comprehensive yet somewhat outdated corpus survey, see Xiao (2008).

[2] https://www.taus.net/data/taus-data-cloud

[3] http://mymemory.translated.net

[4] https://ec.europa.eu/jrc/en/language-technologies

2.2 Academia-Oriented Resources

The Dutch Parallel Corpus (Macken et al., 2011) is a balanced, high-quality and non-growing parallel corpus covering Dutch, English and French, with translations both from and into Dutch. Its advantages are textual variation (19 different types), translational variation (human translation, translation using TMs, postediting and relay translation), as well as variation of ST context. Rich metadata, such as author, text type, publishing information, domain, translation mode, etc., are available.

The Translational English Corpus[5] comprises several sub-corpora of different textual genres and provides a wealth of metadata about texts' extralinguistic parameters, including data about the translators who produced the texts. Contrary to the other resources outlined here, this corpus is not a parallel but a monolingual comparable one, contrasting original English with translated English. It has been influential in TS, especially for research into translationese and stylistic variation.

The MeLLANGE Learner Translator Corpus[6] is a notable representative of multilingual learner corpora and therefore does not include professional translations, but translations produced by translators-in-training. While its size is comparatively small, it provides a wealth of translation-specific metadata on various levels, including information about translators' profiles and the translation processes (e.g. time spent and type of translation aids used). It is most suitable for the study of didactic aspects of translation and of translation quality.

Finally, the Human Language Project (Abney and Bird, 2010) attempted to build a universal corpus of all of the world's languages for the study of language universals and for language documentation. Although not a translation corpus as such, it is of interest to TransBank due to its aim to include as many languages as possible.

From the short survey, it should have become obvious that the reviewed resources differ considerably from each other and that they address different target groups. To a certain extent, their design and architecture reflects the underlying research question or application they have been developed for. In other words, there is no universal, general-purpose repository of translation data suitable for a maximally broad variety of translation-related problems. We believe that the availability of such a resource may both spur data-driven research in TS and foster the collaboration between different stakeholders in translation. To this aim, we plan to develop TransBank as outlined in the following section.

3 TransBank

TransBank is conceived as a large, multilingual, open and expandable collection of translated texts and their originals aligned at sentence level. The core feature is the ability to compile and download parallel sub-corpora on demand, tailored to the requirements regarding specific questions of translation research or translation technology development. TransBank is a meta-corpus in a double sense: firstly, it is a corpus of corpora, and, secondly it provides a rich description of metadata. The analogy is that of a library: When scientists try to answer research questions, they first query the library catalogue to find relevant literature. The library does not provide the answers, but the materials that may contain the answers. In the same way, scientists will be able to turn to TransBank in search of data that may help them to solve their translation-related problems. Note, however, that TransBank will *not* be a metasearch engine, because it will contain the data itself rather than searching for data in third-party resources. TransBank thus closely resembles so-called biobanks, which store biological specimens (e.g. tissue samples) for future use in biomedical research.

The goals of TransBank can be summarized as follows: First, to build an open-ended collection of translations and their STs, aligned at the sentence level. Second, to define a finite set of metadata labels that help capture the distinctive features of translations as highly intertextual language material. Third, to label the data in the bank using the defined metadata labels. And, fourth, to make the data freely available in a highly user-friendly and interoperable form.

The universality and usefulness of TransBank for both academia and the industry is to be ensured by means of the following features:

- **Size**: No size limits are to be set, which implies that the collection will not be a balanced one.

- **Open-endedness**: The collection will grow dynamically, which allows diachronic studies of language and translation over time.

[5] http://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/

[6] http://mellange.eila.jussieu.fr/public_doc.en.shtml

- **“Omnilinguality”**: The collection will not be limited to certain language pairs; on the contrary, the more languages, the more useful the resource will become.
- **Authenticity:** The project is empirical, yet not experimental in nature: it does not aim to include any translations specially made for the project, but to collect real-world translation data only.
- **Textual variation** in terms of genre, subject domains and text types.
- **Translational variation** in terms of different modes of translation (specialized translation, literary translation, localization, subtitling, relay translation, retranslation, post-edited texts, etc.). Any text that on whatever grounds is representative of the population of translations in the broadest sense may be included into the collection. A yet-to-be-developed definition will help to deal with borderline cases on theoretical grounds.
- **Translator Variation**: Not only translations carried out by professionals are to be included, but also trainee translators’ works (as in the case of learner corpora) or amateur translations (as in the case of crowdsourced translation or fansubbing corpora, e.g. OpenSubtitles[7]).
- **Alignment** of texts at sentence level.
- **Metadata labelling** using a fine-grained and objectivity-oriented label set.
- **Download** of TMX and plain text files.
- **High quality** ensured by semi-automatic data collection and manual processing.
- **User-Friendliness** of search interface.
- **Free availability** under CC license.

3.1 Data Collection

As TransBank is only in its pre-project stage, only a few thousand words of test data for the DE-EN language direction have been collected. However, the following is to outline the collection principles to be used during the actual project.

The platform is to impose no restrictions regarding language pairs and directions, genres, text types, subject domains, translator experience and status, time of production, etc. The only two criteria for selection are that texts must be translations paired with their respective originals, and that texts are legally unproblematic in terms of copyright laws and data privacy. Data Selection will be guided by a data harvesting plan that comprises two components: the retrospective collection of legacy data on the one hand and the prospective collection of new data on the other. The retrospective collection consists in enriching existing resources (hence legacy data) with metadata, whereas the prospective collection aims to source and label new data not yet included in any repository. Collecting new data presupposes that the respective translators or LSPs are prepared to provide data and metadata corresponding to our label set already from the start. In this regard, our prospective data collection resembles the TAUS Data Cloud or Translated’s MyMemory, where users contribute their own data, with the important difference that submitted data will undergo metadata labelling prior to inclusion. The prospective approach is expected to yield significant quantities of non-literary texts, i.e. text types that constitute the bulk the translation industry’s output but that are often underrepresented in translational corpora. As a matter of fact, the focus on literary translation in TS is deemed to be a major factor contributing to the theory-practice gap (Sun 2014); therefore we hope that TransBank may help to narrow the gap. The prospective collection of data will require a community effort between the TransBank project and partners from academia and industry. An incentive for LSPs, especially small ones, is that they will receive manually checked sentence alignment for the data they provide, free of cost, which they can then ingest into their CAT systems.

3.2 Metadata

At this point, the question why metadata is crucial to the TransBank project remains to be answered. The answer has a theoretical and a practical perspective.

At the theoretical end, one must recall the difference between data and information: data is only the body of recorded facts, whereas information refers to the patterns that underlie these facts and that are a prerequisite to gain knowledge about something. In order to identify patterns, in turn, one must be able to group a series of observed facts according to some shared ontological charac-

[7] www.opensubtitles.org

teristics. This is especially true of corpora, which consist of samples of language in use (*parole*) taken out of their natural communicative context. Metadata help to restore that context (Burnard, 2005) and to form subgroups of language samples by relating them to the population they were originally taken from.

On the practical end, sample grouping according to some shared features corresponds to the compilation of sub-corpora tailored to certain previously specified criteria, for example for the training of domain-specific MT systems from the perspective of NLP, for the compilation of domain-specific TMs in the translation industry, or for the gathering of research data to investigate a specific translation-related problem from the perspective of TS. No less important, fine-grained metadata allows to filter data and thus to reduce noise in subsamples, which is very important in the case of very large data collections.

TransBank metadata are to include all major aspects relevant to the production of the translation, such as target language, text type, subject domain, intended use, translator experience and education, use of translation aids, time of production etc. What is decidedly not going to be labelled is translation quality, as this is an issue that has still not been resolved by the scientific community: the translation bank would provide a valuable, re-usable resource for tackling this research question. A separate subset of the labels will have to be defined for STs as they, too, have a number of key features relevant to translation, e.g. source language, year, place and channel of publication, genre and text type, intended audience, if they are translations themselves (resulting in intermediary translation), and so forth. Summing up, the set of metadata labels will provide a precise and generally valid tool to describe the intertextual relation between STs and TTs.

As for the collection of the metadata, these will in part be provided by text donors and reviewed/completed by trained project staff, researching missing entries as well as possible. In this regard, cooperation by data donors is again expected to be improved by the added value provided to them in the form of cost-free sentence alignment.

3.3 Data Storage and Access

Data Storage will be provided in the form of TMX files for the aligned text and METS as a container format for metadata. The web-based search and presentation platform is to provide output options for the download of the texts, which can be generated via XSLT from the above XML formats: plain text of STs and TTs, and TEI compliant XML-files for those who want to label data within the texts as well, as opposed to our metadata about the texts. The TMX/METS files will be available for download as well.

The platform will allow for faceted search operations, which can be used for downloading specific sub-corpora. This means that search parameters can be combined instead of only used in a mutually exclusive manner, as is the case with fixed, separate (sub-)corpora or hierarchically labelled data. One of the most common use cases of faceted search is the narrowing of search results in online-shopping: e-commerce platforms allow users to tick categories, i.e. search facets, such as *manufacturer*, *price range*, *user rating*, *shipment options*, etc. In the discussed meta-corpus, the combination is not only one of various labels for one group of texts, but for two: users have to choose a combination of metadata labels for the STs on the one hand and for the pertinent TTs on the other. The resulting search mask is therefore two-sided. Table 1 shows an example of a query to compile a parallel Bulgarian-English corpus of fictional texts published in Bulgaria between 1995 and 2017 and translated by female translators into their native language English.

Source texts (included in download [**yes**] / [no])	Target texts (included in download [**yes**] / [no])
[language (Bulgarian)] [published from (1995) to (2017)] [published in (Bulgaria)] [genre (fictional)]	[language (English)] [translator (female)] [translation into (native language)]

Table 1: Example query in two-sided mask.

As can be seen in Table 1, users can also choose if STs, TTs or both are to be included in the download, i.e., corpora consisting of only STs or only TTs are an option, too, both of which not necessarily monolingual. This makes it possible to generate comparable corpora as well, e.g. by searching for all original texts from a certain subject domain in various languages, without considering the translations included in the bank. The search engine to be used is Elasticsearch[8].

[8] https://www.elastic.co/products/elasticsearch

4 Expected Outcomes

While sharing a number of features with each of the resources reviewed in section 2, the combination of features, together with a new way of accessing translation data via the envisaged web-platform for faceted search, is expected to provide a genuinely new repository of translation data. The main innovative feature is the ability to compile and download parallel or comparable sub-corpora on demand, tailored to the requirements of specific translation-related problems. This may be beneficial to all stakeholders in translation.

From the perspective of TS, a universal translation repository may promote data-driven research. This, in turn, may increase objectivity, validity and reliability of research and eventually make TS more compliant with the scientific method – a desideratum well in line with the growing awareness of the importance of more rigorous research in TS (Künzli, 2013). The range of possible studies using TransBank data is virtually limitless. Studies may include, for example, diachronic issues such as translation-induced language change; the impact of translation tools on linguistic features of written text; contrastive questions regarding differences between text-type norms and conventions across languages; explorations of the characteristics of crowdsourced translation; or cognitive research interests in connection with the differences between texts produced by trainees and experienced practitioners. It is important to note at this juncture that TransBank does not itself aim to answer such questions, but to provide a resource that facilitates such studies. Therefore, the planned first version does not include any tools aimed at complex analyses of translation material (e.g. collocations), only for searching and compiling it.

From the perspective of NLP and translation technology providers, the openness and targeted high quality of sentence-aligned translation data is expected to make the repository useful for the training and testing of new systems. The focus on metadata will facilitate the collection of custom domain-specific data. The dynamic and open-ended nature will yield a sufficiently large data quantity for big data approaches.

From the perspective of the industry, it is again the availability of domain-specific TMs and training data for MT what makes the repository interesting to LSPs as well as individual translators.

Finally, from the perspective of translator training, the analysis of authentic data rather than made-up examples has great didactic potential, especially when contrasting professional with non-professional translations. Similarly, the use of TransBank as a learner corpus to find recurrent translation errors in groups of trainee translators may help to improve translator training.

On a more general level, the approach to metadata labelling applied to TransBank has the potential of becoming a generally valid translation-specific metadata standard in academia and the industry. The importance of standardization is not to be underestimated in view of the ever increasing amounts of translation data being produced and demanded in the digital era.

5 Conclusion

The main problem to be addressed by the TransBank project is a lack of translation data that may be used to make explicit real-world translation phenomena and provide sound theoretical models capable of explaining them. This would benefit not only TS as a discipline in the tradition of the humanities, but the language industry, including translation technology providers and NLP developers, as well. TransBank is therefore conceived as a universal one-stop repository to serve the needs of all stakeholders in translation.

In summary, what we are aiming to provide is a re-usable, open, sustainable and dynamic collection of real-world translation data covering a large variety of languages, genres, subject domains, text types, translation modes, translator profiles and text production settings. Users will be able to download parallel and/or comparable corpora on demand, tailored to their specific translation-related problems. The key to such a customizable flexibility is a precise set of metadata labels that capture the relation between translated texts and their originals, including the circumstances under which the translations were produced.

Given its universal nature, TransBank may promote the collaboration between various interest groups in translation. It is therefore a link between the translation industry, NLP and academia in a data-centric world.

Acknowledgments

TransBank is funded by the *go!digital 2.0* program of the Austrian Academy of Sciences. We also thank the three anonymous reviewers for their helpful comments.

References

Steven Abney and Steven Bird. 2010. The human language project: building a Universal Corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 88-97. http://aclanthology.coli.uni-saarland.de/pdf/P/P10/P10-1010.pdf.

Lou Burnard. 2005. Metadata for Corpus Work. In Martin Wynne, editor, *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books, Oxford, pages 30-46.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79-86.

Alexander Künzli. 2013. Empirical approaches. In Yves Gambier and Luc van Doorslaer, editors, *Handbook of Translation Studies. Volume 4*. John Benjamins, Amsterdam, pages 53-58.

Lieve Macken, Orphée De Clercq and Hans Paulussen. 2011. Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus. *Meta*, 56(2):374-390. http://dx.doi.org/10.7202/1006182ar.

Sanjun Sun. 2014. Rethinking translation studies. *Translation Spaces*, 3:167-191. dx.doi.org/10.1075/ts.3.08sun.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages. 2214-2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Jaap van der Meer. 2016. Datafication of Translation. In *TAUS Blog*. www.taus.net/blog/datafication-of-translation.

Peng Wang. 2015. Datafication of Translation. In *Keynotes 2015. A Review of TAUS October Events*. TAUS Signature Editions, Amsterdam, pages 11-14. www.taus.net/think-tank/reports/event-reports/keynotes-2015.

Andy Way and Mary Hearne. 2011. On the Role of Translations in State-of-the-Art Statistical Machine Translation. *Language and Linguistics Compass*, 5:227-248.

Richard Z. Xiao. 2008. Well-known and influential corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook. Volume 1*. Mouton de Gruyter, Berlin, pages 383-457.

Federico Zanettin, Gabriela Saldanha and Sue-Ann Harding. 2015. Sketching landscapes in translation studies: A bibliographic study. *Perspectives*, 23(2):161-182. dx.doi.org/10.1080/0907676X.2015.1010551.