MICHAEL CYSOUW (Leipzig)
BERNHARD WÄLCHLI (Leipzig)

Parallel texts: Using translational equivalents in linguistic typology

Parallel texts are texts in different languages that can be considered translational equivalent. We introduce the notion 'massively parallel text' for such texts that have translations into very many languages. In this introduction we discuss some massively parallel texts that might be used for the investigation of linguistic diversity. Further, a short summary of the articles in this issue is provided, finishing with a prospect on where the investigation of parallel texts might lead us.

1. Introduction

This issue grew out of a workshop with the same title held on April fool's day 2005 at the Max Planck Institute for Evolutionary Anthropology in Leipzig. Besides the present contributors, there was also a presentation by JOHAN VAN DER AUWERA on his work with parallel texts, which has already been published elsewhere (VAN DER AUWERA *et al.* 2005). The main goal of the workshop, and of this issue, was to bring together typologists that have been working with translated texts. The articles in this issue give a survey of past experiences, some words of caution for future aspirants in this line of research, but also various bold attempts to employ this rich source of data in spite of all possible problems.

2. Massively parallel texts: a selection

According to Wikipedia "a parallel text is a text in one language together with its translation in another language".[1] Parallel texts have played an essential role in philology (often referred to there as BILINGUALS) mainly for deciphering ancient languages, the most famous example being the Rosetta Stone. The currently most widespread scientific use of parallel texts is related to the study of (automatic) translation. Yet, in both literary and computationally oriented approaches to translation mostly parallel texts are used with translational equivalents in only two languages. For linguistic typology such pairwise comparisons are of limited value. If one wants to compare large sets of languages, then mainly such texts are of interest of which translations exist in very many, and ideally also very diverse, languages. We propose to use the term 'massively parallel text' (MPT) for such texts of which many different translations are available. Here, we would like to present a few texts that might be useful in future typological investigations. This summary only raise some possibilities and does not aspire any completeness

Probably the most widely used MPTs in computational approaches are the verbatim reports of the proceedings of the European Parliament. These reports are freely available online.[2] The earliest proceedings were translated into nine languages

[1] http://en.wikipedia.org/wiki/parallel_text_alignment
[2] http://www.europarl.eu.int/activities/archive.do

(French, Italian, Spanish, Portuguese, English, German, Dutch, Danish and Greek), somewhat later joined by Finnish and Swedish. Recently, the number of languages into which the reports are translated was extended to twenty (added were Czech, Estonian, Latvian, Lithuanian, Hungarian, Maltese, Polish, Slovak and Slovene). Bulgarian, Irish and Romanian are planned to be included in 2007. Although this is clearly a massively parallel text—in number of languages but even more so in the sheer amount of text—the diversity of languages available is too narrow for many typological purposes.

An even much more massively multilingual organization is the United Nations. Here the most well-known MPT is the *Universal Declaration of Human Rights*, currently available online in 332 different languages.[3] The usage of this text for typology is somewhat restricted because of the rather legalese language-variety used in this document. Still, for some linguistic domains this MPT can be fruitfully applied (cf. WÄLCHLI 2005, Ch. 6). Less well-known is the online database of literary translations of the UNESCO: the *Index Translationum*.[4] This database contains about 1.5 Million entries about translated works. For example, 51 translations in thirteen different languages of Agatha Christie's *Partners in Crime* are listed (German, Czech, Portuguese, Spanish, Norwegian, French, Finnish, Indonesian, Italian, Bulgarian, Hungarian, Korean, and Lithuanian). This database can be a fine starting point to find references for MPTs.

The most famous MPT is of course the christian Bible (see DE VRIES, CYSOUW *et al.*, DAHL, and WÄLCHLI, this issue). There is a long tradition of using Bible texts for language comparison, the most famous multi-lingual text being the Lord's Prayer (see ADELUNG 1806-1817 [1970]). A collection of the Lord's Prayer is online available in more than 1,300 languages.[5] The merits of this particular MPT is restricted because of the short size and the strong theological impact of the exact wording of the translation. More interesting are the various active endeavors to translate the whole Bible, or at least large parts of it, into as many of the world's languages. It is difficult to assess how many translations have been made, but the Wycliffe Bible Translators website estimates that the whole Bible is translated 'only' in about 400 languages.[6] However, they also estimate that there are a further 1,000 languages in which at least the New Testament is translated, and about 800 languages in which at least some parts of the scripture is available. Further, they claim that in more than 1,500 languages Bible translations are in progress. Most of these translations only exist as hard-copy published versions. These are often difficult to obtain because most public libraries do not collect translations of the Bible. As for online availability, the Sword Project[7] and the Zefania Project[8] both give access to various freely available Bible translations. Further, the Rosetta Project has about 1,200 scanned versions of different genesis translations in more than 1,000 languages. Pending some copyright issues, these should become available

[3] http://www.unhchr.ch/udhr/navigate/alpha.htm
[4] http://databases.unesco.org/xtrans/
[5] http://www.christusrex.org/www1/pater/
[6] http://www.wycliffe.org/wbt-usa/trangoal.htm
[7] http://www.crosswire.org/sword/
[8] http://sourceforge.net/projects/zefania-sharp/

online soon. [9] Besides the Bible, but also in the Christian realm, another MPT is a collection of some (short) introductory texts of the Jehovah's Witnesses, which are available online in 264 different languages.[10]

As another MPT, many translation are available of key Marxist's texts. In the former Soviet Union, a major effort has been made to translate various important Marxists' texts into many different languages. For example, the *Index Translationum* lists 71 translations in 36 languages of LENIN's *State and Revolution*. Even better, the Marxist's Internet Archive provides direct online access to 24 of these translations in different languages.[11] There are definitively more translations of LENIN in printed versions, though it might be difficult to get hold of them after the demise of the Soviet Union.

Two MPTs have already been used to some extent in typological investigations: ANTOINE DE SAINT-EXUPÉRY's *Le Petit Prince* and the books of *Harry Potter* by J. K. ROWLING (see e.g. STOLZ and DA MILANO, this issue). Not yet used in typological research, as far as we know, are the fairy tales of HANS-CHRISTIAN ANDERSEN. The Andersen Museum in Odense actively collects translation of his stories, and they claim to have translations in as much as 123 languages.[12] Their website provide some scanned pages, though apparently not everything they have collected is available online. Also it seems that not always the same stories that have been translated, which diminishes their utility as a MPT. Further, some interesting fairy tale-like MPTs can be found on the UNILANG webpage.[13] On this community-driven collection of multilingual resources there is a collection of short stories that are being translated by internet users. These stories are supposed to be used in language learning, and therefore deliberately evade complex linguistic constructions. Among these stories is also the infamous Æsop fable *The North Wind and the Sun*, which got some recognition in linguistics because the International Phonetic Association uses it to exemplify the usage of the International Phonetic Alphabet (cf. HANDBOOK 1999).[14]

Finally, a possibly interesting source of MPTs is movie subtitles. There is an active online community where subtitles for movies are exchanged.[15] These subtitles are partly ripped from DVDs, but often self-made by fans of a particular movie. The more popular films will therefore be available in various languages, but also in multiple versions of the same language. For example, there are 76 different subtitles in 21 different languages listed for the film *Harry Potter and the Sorcerer's Stone*. Although there are many restrictions on the languages used in subtitles (like the length of the phrase, which has to fit on the screen), this source of information might be interesting because most of the text is direct speech—in contrast to all other MPTs discussed previously, in which the majority of the text are reports.

---

[9] http://www.rosettaproject.org/
[10] http://www.watchtower.org/languages/languages.htm
[11] http://www.marxists.org/xlang/index.htm
[12] http://webpartner.odmus.dk/andersen/eventyr/
[13] http://home.unilang.org/
[14] http://web.uvic.ca/ling/resources/ipa/handbook.htm
[15] http://divxstation.com/subtitles.asp

3. Survey of this issue

This issue opens with a paper by THOMAS STOLZ in which he discusses his experiences with using parallel texts in his typological research over the past decade. Although he notes many possible pitfalls and drawbacks in this kind of research, the actual examples discussed show that there is definitively great value in using massively parallel texts.

BERNHARD WÄLCHLI, also drawing on some experience working with parallel texts, presents a new case study, showing how parallel texts offer a possibility to take into account language-internal variation. Notwithstanding this worthwhile addition to the typologist's toolbox, he finishes his paper with some words of caution. Typologists should be aware of the limits of the applicability of parallel texts. Some research topics might profit from such an approach, while others should better refrain from this method.

In the contribution of FEDERICA DA MILANO parallel texts are used to supplement a classical questionnaire study into the structure of demonstratives in the languages of Europe. The insights from the parallel texts are not as compelling as the (more controlled) results from the questionnaire, though they illustrate the earlier findings with 'real' examples.

LOURENS DE VRIES describes in detail some of the processes involved in the translation of the Bible. In particular, he directs attention to its textual multiplicity: there is not one single base text, but rather a number of quite strongly different scriptures, each having its own long tradition. Depending on time, place and Christian church, different version of the Bible were (and still are) the basis for translations. This implies that one cannot automatically assume that different Bible translations are directly equivalent.

The interpretation of the linguistic structure of the multitude of languages involved in an investigation of a massively parallel text is often a tedious and time-consuming affair. MICHAEL CYSOUW, CHRISTIAN BIEMANN and MATTHIAS ONGYERTH investigate a computational approach that automatically suggests a rough gloss for each sentence—based on purely statistical properties of the texts. Although there are various methods available for the automatic alignment of parallel texts, the algorithm presented in this paper has the advantage that it is completely language independent.

Finally, ÖSTEN DAHL approaches parallel texts from the background of his own past research using questionnaires. Massively parallel texts, when available and when applicable, can be a much cheaper method (both money- and laborwise) to reach fine grained typologies. As a first attempt, he presents some insights that can be gained from comparing English Bible translations from different times, showing how linguistic change can be read off differences in the translations.

4. Prospects

Massively parallel texts are an important addition to the kinds of data used in linguistic typology. They are surely not the holy grail of language comparison, but parallel texts are a useful and needed supplement to the traditional data source of

typology (reference grammars, dictionaries, and the interrogation of native speakers using questionnaires). Of course, everyone using translational equivalents should be aware of various inherent biases implied in this kind of data. First, almost all of these texts represent written language, and in most cases also rather standardized registers. In the case of the Bible, the texts often represent even such a specialized register as to make the lect used substantially different from the 'normal' language. However, there is nothing against the inclusion of a great variety of lects—after all, they should all be accounted for in a general theory of linguistic structure. Second, through the process of translation, there is always the chance of inference from the source language. If the topic of investigation is expected to particularly prone to inference, it might be better not to use parallel texts for its investigation. Also, a *post-hoc* control should be performed for any source language influence. If the typology resulting from a parallel text study classifies languages together of which the translations are based on the same source language, this of course disqualifies the validity of the typology.

Still, using parallel texts can have many benefits—and to show this is the major aim of this issue. As the exemplars studied are all contextually situated, it is possible to investigate the influence of context on the structure of the language. Further, by using multiple text passages that are expected to show identical structure, it is possible to investigate language-internal variation—something that is hardly possible by perusing grammars and dictionaries. Finally, by investigating the details of variation between languages it is possible to obtain much more fine-grained typologies. However, all such prospects ask for a much better quantitative interpretation of the data as currently practiced. This is surely a field in which more methodological efforts are needed, too.

References

ADELUNG, JOHANN CHRISTOPH (1806-1817 [1970]): Mithridates oder allgemeine Sprachenkunde mit dem Vater Unser als Sprachprobe in beynahe fünfhundert Sprachen und Mundarten. Fünf Bände. Berlin: Voss / Hildesheim: Olms.

HANDBOOK (1999): *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press.

VAN DER AUWERA, JOHAN; SCHALLEY, EVA & NUYTS, JAN (2005): Epistemic possibility in a Slavonic parallel corpus – a pilot study, in: HANSEN, BJÖRN & KARLÍK, PETR (eds.), *Modality in Slavonic Languages. New Perspectives*. München: Otto Sagner, 201-217.

WÄLCHLI, BERNHARD (2005): *Co-compounds and Natural Coordination*. Oxford: Oxford University Press.

Correspondence address

Michael Cysouw & Bernhard Wälchli
Max Plank Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
cysouw@eva.mpg.de