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

THOMAS STOLZ (Bremen)

Harry Potter meets *Le petit prince*: On the usefulness of parallel corpora in crosslinguistic investigations[1]

This paper documents some of the experiences I have made in the course of my (areal) typological research projects. The empirical basis of these projects stems from the analysis of two large parallel literary corpora. The texts involved are original and translations of Antoine de Saint-Exupéry's *Le petit prince* and Joanne Rowling's *Harry Potter* series. The paper addresses a selection of issues touching upon methodological, theoretical and practical problems of this kind of corpus-based linguistic research. Parallel corpora offer interesting possibilities for typological research. However, working with parallel literary corpora often imposes severe restrictions upon sample size and sample composition as there is a clear European bias in terms of available translations.

1. A long introductory lament

This paper is meant as a general comment on the state-of-the-art of crosslinguistic methodology by way of weighing the pros and cons of typologically-minded research based on parallel corpora. In this section, I start with a selection of critical remarks referring to what might be called received common practice in typology and universals research. Many of my observations are well-known facts and thus may sound trivial. However, I consider it useful to review these facts together in order to prepare a checklist for the work with parallel corpora which is still in its infancy. In Sections 2 and 3, I present glimpses of my own experience with two distinct parallel literary corpora, viz. the translations of *Le Petit Prince* and the ones of *Harry Potter*. In the final Section 4, I draw the necessary methodological conclusions.

Both typologists and universals researchers are eager to make sure that the empirical basis on which they build their theories is such that it guarantees the highest possible degree of comparability of the languages sampled. Questions of optimal sample size and composition have been amply discussed in the literature (PERKINS 1989; RIJKHOFF & BAKKER 1998). Besides all sample-related problems of which languages to compare, we have to decide whether or not the data we draw from different languages are indeed in a relation of equivalence among each other and thus allow for being compared at all. The notorious *tertium comparationis* (SEILER 2000: 28-9) enters the scene: if two or more phenomena are to be compared to each other in order to yield generalisations, there must be a language-independent yardstick.

It has become common practice in crosslinguistic research to use grammatical categories (say, comitatives, STOLZ 1997), functions (say, possession, HEINE 1997), construction types (say, co-compounds, WÄLCHLI 2005), or word-classes (say, numerals, HANKE 2005) as a *tertium comparationis*. Literally hundreds of

[1] When I use the 1st person singular in this contribution, I do this not without expressing my gratitude to the members of my research team at the University of Bremen for their help in all sorts of matters: TAMAR KHIZANISHVILI, NATALIYA LEVKOVYCH, SONJA KETTLER, CORNELIA STROH, and AINA URDZE. I also like to thank MICHAEL CYSOUW and BERNHARD WÄLCHLI for inviting me to participate in their project. If there is anything wrong with this article, the blame should be put on me alone.

languages world-wide have been checked for the presence/absence, the distributional and formal properties of the said categories, functions or constructions. More often than not, the researcher's language expertise is limited to only a small sub-set of his sample. Therefore, he has to rely heavily on the information available elsewhere. Extant descriptive material (grammars, dictionaries, monographs and articles devoted to selected topics) is often enough perused hurriedly to get hold of as many examples of the item searched for as possible in the shortest possible time. Besides widely acknowledged advantages, this method also has its pitfalls as all descriptive grammars leak, in a manner of speaking. Furthermore, many descriptive grammars have a prescriptive touch too in the sense that the authors make a conscious choice of observable phenomena—a choice that may be motivated by puristic ideas or other ideologies (sometimes dictated by the theoretical framework one has opted for).[2] Thus, chances are that what the researcher who uses these grammars is looking for simply is not dealt with in some of them. Or it may be there but hidden in an unexpected context in a given grammar. Terminological mannerisms and unfamiliar descriptive formats may also lead to oversight or misinterpretation.[3] With reference to these and other potential sources of error, proponents of the method described claim that, for statistical reasons, a large sample can make up for the occasional mistake because the probability of grave errors diminishes with a high number of recurrent instances of the same phenomenon.

Basing oneself entirely on the extant descriptive literature is tantamount to pretending that (the) languages have already been described exhaustively—which, of course, is nothing but an illusion to which nobody would dare to subscribe (CROFT 2001: 3-46). Moreover, this method is at its best when it comes to determining whether a well-defined phenomenon is frequent or not among the languages of the world. It may also shed some light on certain patterns as, for example, the co-occurrence or incompatibility of phenomena. The dozens of contributions to the *World Atlas of Language Structure* (HASPELMATH *et al.* 2005) demonstrate that the massive application of the grammar-perusal method yields highly interesting results especially as to the geolinguistics of human languages. However, if one wants to know about the full range of uses of the phenomena in order to put these bits of information in crosslinguistic perspective, checking grammars for the presence/absence of something alone does not suffice. The method works sufficiently well with established categories but fails to account for emergent ones (BYBEE & HOPPER 2001). It is prone to ignore language-internal variation—not only diatopical but also stylistic and context-dependent variation. The usual decontextualised

[2] A good example of the repercussions purist-mindedness may have on descriptive linguistics is the deliberate omission of Spanish-derived function words in many grammars of indigenous languages of the Americas although these borrowings are fully integrated in the language system (BRODY 1998).

[3] For example, Basque receives a dot with the wrong colour on the WALS-map on reduplication (RUBINO 2005: 116): SALTARELLI *et al.* (1988) is RUBINO's main source of information on Basque and it does not systematically describe reduplication processes. A look at LAFITTE (1998), however ,reveals that total reduplication is a highly productive process in all varieties of Basque. STOLZ (1997) and STOLZ *et al.* (forthcoming-b) differ as to the classification of Bambara in their typology of comitatives/instrumentals because different grammars (BRAUNER 1974 vs. KASTENHOLZ 1989) gave widely divergent descriptions of this area of grammar (perhaps because the grammars are based on different regional varieties).

examples provided by the average descriptive grammar make it difficult to find any correlation that reaches beyond the sentence level.

How can this dissatisfying situation be remedied? There are of course other established methods of data collection. Suffice it to mention (a) direct consultation of informants (native speakers or language experts) using questionnaires, (b) recording of stimulus-based natural discourse, or (c) analysis of extant texts. Including the above mentioned grammar perusal, all of these methods have their merits alongside a variety of disadvantages which cannot easily be overcome. For the sake of brevity, I will mention only a selection of the characteristics of each. For (a), the usual problems fieldworkers have to face when they interact with native speakers come to the fore (VAUX & COOPER 1999) and thus relatively large populations of informants are needed before one can be sure that a given response is valid. Questionnaires are problematic too as their design in itself constitutes a potential case of researcher-guided prejudice. Spontaneously produced original discourse (b) has the advantage that there is no interference whatsoever by the researcher but comparing sets of data of this kind from many languages leaves us with the problem to find those aspects that can be compared. That is why linguists often resort to stimulus-based original discourse—for instance, the famous *Pear Stories*. Thus, common ground is created by way of referring to one and the same general topic. However, because informants are free to speak about a given topic, the various results may happen to be largely incommensurable in terms of size, form and content. Last but not least, (a) shares with (b) the tediousness, consumption of time and manpower that are necessary to carry out the data collection and subsequent analysis thereof.

Thus, the idea suggests itself to circumvent the actual hunter-gatherer footwork by way of exploiting already existing corpora (c). Yet, even in this case, some work still has to be done beforehand. One cannot simply take any random assortment of texts and start comparing because the heterogeneity of the corpora would be an obstacle. With a view to facilitating comparison, the texts used should ideally be identical for all the sample languages. The easiest way to achieve this identity is translation. Translation, however, reeks of non-authentic language, meaning: one can never be sure whether a given phenomenon that is attested in a translation would ever have been produced in the same way in an original text of the same language. Clearly, this is the same problem as the one mentioned for the questionnaires above. Further, parallel corpora made up of translations of one and the same text are almost exclusively specimens of written language. This restriction to only one register is in itself a problem—a problem that is aggravated by the fact that written language very often seems to obey rules of its own which do not necessarily reflect what speakers do when they talk. On top of that, translation-based parallel corpora normally comprise only one idiolect per sample language as long as there are no competing translations of the same text. Thus, the population of native speakers per language represented in the corpus is minimal. Admittedly, translators will probably follow the model of a standard (if there is any) because they intend to be understood by their readers.

With a view to making statements which are generally valid for languages, and not for "marked" varieties thereof, we have to find texts which bear close resemblance to actual language use. This can be achieved by gathering texts which in-

clude frequent passages of direct speech. Ideally, the chosen texts should reflect contemporary usage as this allows for additional checks with native speakers. Realistic narrative prose is the best candidate whereas poetry, bound poetic form in general and any kind of avant-garde or *l'art pour l'art* kind of literature with surrealistic formal ambitions are ruled out as sources for our generalisations about human language. Likewise, special language for certain technical or other disciplines is not suited for our comparative endeavour.

On first view, the Bible appears to be an good example of such a text. Apart from the general peculiarities of so-called holy books/scriptures,[4] the factor time is one of the problems which render working with a parallel corpus of Bible translations difficult.[5] Not all of the many extant translations are recent. The chronology of translations covers half a millennium at least. Moreover, there are several originals from which the translations could have been made (Latin, Greek, Hebrew/Aramaic etc.) and thus apparent differences among the sample languages might turn out to reflect differences in the originals. As to length, however, the Bible is almost ideal. Other aspects which influence the selection of sample texts are, on the one hand, their availability and, on the other, legal problems such as copyright regulations. Thus, there is a variety of factors which together or in isolation may have the undesired effect of restricting our choice of languages for the planned sample—both for its size and for its potential members. Such externally imposed restrictions are of course in conflict with the linguist's genuine interest in studying certain phenomena. These phenomena themselves may require the inclusion/exclusion of certain languages and/or the use of particular texts.[6] Quantitative issues are probably less of a problem than qualitative ones.

Is it at all feasible then to use parallel corpora? Does it make sense at all to carry out crosslinguistic research on their basis? In the light of my longish list of complaints above, this answer may come as a surprise as it is yes, nevertheless. In the subsequent sections, I explain this positive turn indirectly by way of self-reviewing two of my own typological projects for which parallel corpora have proved to be handy and ultimately indispensable tools. Note that Sections 2 and 3 are only meant to hint at some general problems. The data discussed are neither analysed exhaustively nor do they represent more than a fragment of the whole stock of data available to us.

---

[4] Missionaries are notorious for creating new varieties of the languages into which they translate the Bible or parts thereof. If we accept the Bible style as representative of a given language, we might run the risk of working with a partially artificial (non-)language in our sample (ZIMMERMANN 1997). On general aspects of distorted hagiolects, see ENNINGER & RAITH (1981).

[5] However, KAISER (2005) demonstrates that Bible translation can successfully employed for solving diachronic riddles.

[6] Consider, for example, the dual in Latvian (ENDZELĪNS 1951: 450-60 and 702). Any corpus of contemporary texts will yield the same statistical result, namely that the category is marginal, if present at all, because it is no longer part of standard language. However, if one wants to know about the functions in which the remnants of the dual are still involved, one has to resort to texts with a somewhat rural (and perhaps anachronistic) flavour such that they include direct speech of elderly inhabitants of the countryside.

2. *Le petit prince*

I have used the parallel corpus of translations of ANTOINE DE SAINT-EXUPÉRY's *Le petit prince* for three major typological projects, viz. one on comitatives and instrumentals (STOLZ *et al.* 2005; forthcoming-b), one on alienability (STOLZ *et al.* forthcoming-a) and a third one on total reduplication (STOLZ 2004).[7] The corpus became necessary because these topics do not belong to the canon of phenomena accounted for by each and every grammar. The text has been translated into more than 150 languages (including regional/dialectal substandard varieties) of Europe, Asia (including the Philippines), Africa and the Americas. More translations are to be expected in the near future.[8] However, there are lamentable gaps on the global map: neither Australian nor Oceanian languages are represented. Of the indigenous languages of the Americas, a translation only exists for Quechua.

Under these circumstances, the focus was narrowed down to the areal-typology of Europe.[9] What compelled me to this drastic change of perspective was the disproportion of readily available translations for the languages of the various continents. If we wanted to get working at all, we had to start from a European-biased set of translations anyway. Otherwise we would have been forced to reduce the sample size in terms of numbers of languages in order to avoid the unwanted effects of areal under-representation and over-representation, respectively. The sample consists of the 64 translations as shown in Table 1. Italics marks those languages for which a translation of *Harry Potter* (at least the first book) is available, too.

With 86% of the sample, Indo-European languages clearly outrank the members of other genealogic groups. Thus, the sample is biased to the detriment of the non-Indo-European languages of Europe. Additionally, two of the major Indo-European phyla are over-represented to some extent as Germanic and Romance (but also Albanian) substandard/regional varieties form part of the sample alongside the respective standard languages, whereas the Slavic phylum is represented exclusively by standard varieties. With the objective to create a genetic balance, the sample would have had to be reduced considerably—a consequence which conflicted with our wish to cover as many languages as possible. We felt entitled to use this sample because areally-minded studies of a comparatively small region are exempt from the requirements of genetically unbiased sample composition, not the least because phyla-internal divergent behaviour of varieties is a valuable piece of evidence for areality (COMRIE 1993).

---

[7] My choice of the sample text was inspired by a similar attempt of a typological sister-project headed by HANS-JÜRGEN SASSE. I seize the opportunity to express my gratitude to YANN VINCENT, France, and GERHARD VOLZ, Austria, two private collectors of translations of *Le petit prince*, who lent me a hand in my search for rare items. For reasons of space, I do not provide the full list of bibliographic references of the translations used for this contribution. The relevant data can be found in STOLZ *et al.* (forthcoming-b).

[8] After our sample was considered complete, several translations into regional varieties of German, Italian and Spanish were published. In addition, there are now several Saami versions, and Udmurt and Tatar have also joined the club.

[9] For the geographic details of our interpretation of the term Europe see STOLZ *et al.* (2003).

Table 1. The sample according to genetic affiliation and status.

Affiliation	Standard	Substandard/Regional
Romance	*Catalan, French, Italian, Portuguese, Rumanian, Spanish*	*Galego*, Aragonese, Asturian, Badiota, Corsican, Friulian, Gascon, Gherdëina, Languedocien, Moldavian, Provençal, Sardinian, Surselvan, Vallader
Germanic	*Danish, Dutch, English, Faroese, German, Icelandic, Norwegian (Bokmål), Swedish*, Letzebuergesh,	Alsatian, Frisian (West), Limburgian (North), Limburgian (South), Yiddish
Slavic	*Bulgarian, Bielorussian, Croatian, Czech, Macedonian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian*	–
Other Indo-European	*Albanian (Tosk), Greek, Latvian, Lithuanian, Welsh,* Armenian (East), Breton, Kurdish (Kurmanči),	Albanian (Gheg), Romany (Lovari)
Uralic	*Estonian, Finnish, Hungarian,* Saami	–
Various	*Basque, Georgian, Turkish,* Azeri, Maltese	–

What can be done with a parallel corpus of this kind? The size of the text (the number of pages oscillating around 100 depending on the translation) is not sufficiently long to yield many substantial insights into qualities, but it has just the right length to be easy to handle and to allow for reliable quantitative statements (ALTMANN & LEHFELDT 1973). To do the comparison properly, equal length of the compared texts is required (TULDAVA 1995: 151-2). For the translations of *Le petit prince*, however, identical length can only be achieved by cutting off the text at a pre-determined mark because the languages differ widely as to the number of pages, words, or sentences they use. I will demonstrate these discrepancies between the different translations for the number of sentences, which we determined on the basis of a purely orthographic criterion, namely the occurrence of full stops, question marks and exclamation marks. The French original contains 1,652 sentences. This number is exceeded only by Greek. Six texts (among them four close relatives of French) display exactly the same number of sentences as the original whereas the bulk of the sample texts (56 languages = 87.5%) fail to reach this number by a margin of minimally one and maximally 124 sentences (see Table 2). The languages with the four lowest scores as to the number of sentences all belong to non-Romance phyla and, geographically, are far removed from French as they are spoken in the European East.

Table 2. Number of sentences per language in *Le Petit Prince.**

No.	Languages
1,663	Greek
1,652	**French, Languedocien, Provençal, Friulan, Rumanian**; Serbian; *Hungarian*
1,651	**Gherdëina**
1,650	**Spanish**; German; Bulgarian, Ukrainian; *Finnish*
1,649	**Italian**, **Vallader;** Frisian; Slovenian; *Basque*
1,648	Czech
1,647	English
1,646	*Turkish*
1,645	Albanian (Gheg); Breton; Danish, Icelandic; *Estonian*
1,644	**Gascognian, Surselvian**
1,643	**Moldavian**; Dutch
1,642	Latvian, Lithuanian
1,641	**Aragonese, Badiota, Portuguese**; Welsh
1,640	**Sardinian**; Norwegian, Swedish; Macedonian, Slovak; *Maltese*
1,639	**Galego**; Faroese; Polish
1,638	Croatian, Russian
1,637	**Asturian**; *Saami*
1,636	Letzebuergesch, Limburgian (North), Limburgian (South)
1,634	**Catalan**; Lovari; *Georgian*
1,633	**Corsican**
1,631	Bielorussian
1,628	Albanian (Tosk)
1,623	Armenian
1,528	*Azeri*

* For easy reference, Romance languages are marked boldface and positioned leftmost in a line, non-Indo-European ones appear in italics and on the right of other languages. Members of different phyla are separated by a semi-colon<;>, members of the same phylum by a comma <,>.

The mere numbers do not necessarily mean that lower figures imply a loss of content (or, for higher figures, a gain in content) as opposed to the original because the rules for using punctuation devices of the individual languages may diverge in such a way that several sentences of the original fuse into one in the translation, or one French sentence may correspond to several sentences in the translation. Becuase of this shuffling about of sentence boundaries, we accepted the possibility of comparing texts of different length as long as the content is kept constant (TULDAVA 1995: 155-9). Furthermore, the above figures suggest the impact of the French original on the translators' choices is not strong enough to determine every structural aspect of the translation. At the same time, the parallel divergence of several languages (for instance, the two Baltic languages with 1,642 sentences each) from the French model is also indicative of something else, namely that despite their claims, the translators have probably not always exclusively translated from the French original, but used another language with which they both were more familiar. Note that the use of one or more additional languages does not al-

ways mean that the translator follows their lead. in the case of Galego, the translator tried hard to find solutions which were sufficiently dissimilar from both Portuguese and Spanish to mark Galego's distinctness (LUNA ALONSO 2000).[10] The third insight to be gained from Table 2 is the fact that genetically closely related languages do not necessarily display identical nor similar results. The differences between the two Turkic languages, Turkish and Azeri, and those of the East Slavic languages, Ukrainian, Russian and Bielorussian, support the idea that members of one and the same genus are still individual languages and behave as such.

This view of things is corroborated by other phenomena which can be ascertained by statistical means. As an example, I present the token frequencies for the primary translations equivalents of French *avec* 'with' in the translations. It is important to note that none of the other languages displays values as low as the 37 attestations of *avec* in the original. Table 3 informs about the token frequency of the translation equivalents of *avec* and their ratio to the number of occurrences of *avec* in the French original. The languages are ordered according to this ratio. Boldface again identifies Romance languages whereas italics are used for glossonyms of non-Indo-European languages.

Not only is it normal for the sample languages to use their equivalents of French *avec* much more frequently than the French original uses *avec* itself, but also genetic affiliation is only mildly indicative of the frequency with which the items under scrutiny are used in a given language. Closely related languages such as Lithuanian and Latvian wind up on different ranks because of their surprisingly divergent token-frequency values which differ by 40 tokens. The gap is even more pronounced for Faroese and Icelandic with 87 tokens more for the former. The Baltic case is especially intriguing because Table 2 still shows Lithuanian and Latvian to behave in a predictably similar way. Azeri and Turkish (which were already dissimilar as to the number of sentences) go again different ways, which is the more remarkable as Azeri (in spite of the lower number of sentences) has the higher token frequency for the translation equivalent of French *avec*.

The patterns of genetically unexpected behaviour, however, are by no means random. At closer inspection, they can be shown to obey an areal logic according to which those languages which deviate from their next of kin behave more like their genetically unrelated next-door neighbours. All in all, there is a kind of cline from the European Southwest to the Northeast, including a center-periphery dichotomy (STOLZ *et al.* 2003). The same applies to our project on total reduplication phenomena which, primarily on the basis of the same parallel corpus has revealed that there is a clear North-South divide in Europe as to the readiness of languages to employ reduplication strategies (STOLZ 2004). Thus, the parallel corpus of translations of *Le petit prince* has made it possible for us to gain relevant insights into the areal-typological structure of Europe.

---

[10] Only anecdotally, I like to point out that native speakers, when confronted with the translated texts, relatively often expressed their dissatisfaction with the translator's choices. For Faroese, for example, our two informants (ZAKARIS HANSEN and VÁR Í OLAVSTOVU) complain unanimously about the overlong sentences, which, to their mind, are not in line with the Faroese speech rhythm favouring short to medium sized sentences. According to their intuition, a better Faroese translation would split up many of the sentences of the French original.

Table 3. Token frequency and ratio of the translational equivalent of *avec*.

Language	Tokens	Ratio
Albanian (Gheg)	403	10.9
Basque	360	9.7
Kurdish	341	9.2
Bielorussian	225	6.1
Maltese	224	6.0
Albanian (Tosk)	219	5.9
Polish	213	5.7
Russian	201	5.4
Rumanian	198	5.3
Ukrainian	192	5.2
Moldavian	177	4.8
Faroese	166	4.5
Armenian (East)	165	4.4
Vallader	157	4.2
Finnish	152	4.1
Welsh	145	3.9
Hungarian	138	3.7
Greek	134	3.6
Swedish	133	3.5
Limburgian (South), Lithuanian	129	3.4
Azeri, Danish	125	3.3
Yiddish	124	3.3
Letzebuergesh	121	3.2
Norwegian (Bokmål)	120	3.2
Portuguese, Limburgian (North)	118	3.1
Asturian	113	3.1
Frisian, Romany (Lovari), *Georgian*	111	3.0
Breton	108	2.9
Bulgarian, Dutch	106	2.8
Gherdëina, Serbian	101	2.7
Badiota, Surselvan	99	2.7
Estonian	96	2.6
Slovenian	95	2.5
Czech	94	2.5
Galego, English, German, *Turkish*	91	2.4
Friulian, Latvian	89	2.4
Aragonese	88	2.3
Croatian, Macedonian	84	2.2
Slovak	80	2.1
Icelandic	79	2.1
Catalan	78	2.1
Alsatian	77	2.0
Spanish	73	1.9
Sardinian	65	1.7
Italian, *Saami*	60	1.6
Provençal	55	1.5
Corsican	53	1.4
Gascon, Languedocien	52	1.4
French	37	1.0

Owing to the limited length of the sample text, many problems connected with determining the exact functional range of an item remain unsolved. A typical example is the difficulty to clarify with certainty whether a given grammeme translating French *avec* is (over-)syncretistic in the sense that it not only encodes instrumental and/or comitative but also what we call the ornative (STOLZ *et al.* forthcoming-a). Consider Sentence 29 of Chapter 14 of *Le petit prince* in the various sample language, as presented in full in Appendix 1. The French original sentence is given here as (1a), and the English and the Croatian equivalents in (1b) and (1c), respectively. Boldface marks the grammemes under scrutiny. Instrumental NPs and ornative NPs are identified by labelled square brackets (excluding governing adpositions but including bound case markers), labelled 'tool' and 'ornative', respectively. Numerical indexes distinguish instrumental markers (lower case 1) from ornative ones (lower case 2).

(1) a. French (Romance)

Puis	*il*	*s-épongea*		*le*	*front*
then	he	REF.3-mop:PAST		DET.M	forehead
avec₁	*[un*	*mouchoir*	*à*	*carreaux*	*rouges]*^{TOOL}.
with	a	handkerchief	at	square.PL	red

b. English (Germanic)
Then he mopped his forehead
with₁ *[a handkerchief decorated* ***with₂*** *[red squares]*^{ORNATIVE}*]*^{TOOL}.

c. Croatian (Slavic)

Zatim	*obrise*	*čelo*	
then	wipe:REF	forehead	
*[rupčić**em₁***	***s₂***	*[crven**im₂***	*kvadrat**ima₂**]*^{ORNATIVE}*]*^{TOOL}.
handkerchief:**INS**	**with**	red:**INS.PL**	square:**INS.PL**

Taken at face value, these sentences are suggestive of a partition into three groups. The largest one comprises almost 80% of the entire sample: 51 out of the total 64 languages make use of only one translation equivalent of French *avec*—and this equivalent always encodes the instrumental relation. 13 languages (or 20% of the sample) overtly mark two relations, namely instrumental and ornative. However, ten of those (= 15.6%) use the same grammeme twice,[11] i.e. the grammeme is polysemous as it encodes both instrumental and ornative, like English. A minority of three languages (= 4.4%) use two distinct constructions each, namely the simple inflectional instrumental for the instrumental relation and a PP headed by a preposition which also governs the inflectional instrumental for the ornative relation. These last mentioned languages belong to the Slavic phylum, more precisely to its Western and Southern branches. However, on closer inspection, this supposed typology starts to crumble. Native speakers confirm that for practically all members

[11] For the interpretation of the allographs <*â*> and <*a*> in Welsh as representatives of one and the same grammeme, cf. STOLZ (1998).

of the Germanic and Romance phyla, the constructions reported for English in (1b) are also fully acceptable. Moreover, we also learned that various Slavic languages—especially Russian—display a growing tendency to replace the constructions of (1a) by those of (1c) although this is still stigmatised by normative grammar which favours ornative adjectives in lieu of a PP (ZEMSKAJA 2004).

Another sentence, no. 150 in Chapter 26, shows that 58 (= 90%) out of 64 languages construe the relation *well(s) with a (rusty) pulley* identically as they use the grammeme translating French *avec* as relator in the construction. What this fact implies is that some of the languages (e.g. Lithuanian, Bielorussian, Czech, Russian, Serbian) in this sentence behave differently from the pattern they follow in the example aboce. Since the sample text is too short to contain sufficient cases of ornative-like relations, it cannot be decided whether the observed variation reflects stylistic options or obeys other more strict rules. Thus, we have reached the limits of this corpus of parallel translations. For some questions, *Le petit prince* surely has the right answers handy but not for all.

3. *Harry Potter*

The books of the *Harry Potter* series fulfil the same criteria as *Le petit prince.*[12] In contradistinction to the latter, the still unfinished series provides a rather large amount of text already going far beyond 2,400 pages (= 24 times as large as *Le petit prince* in terms of pages) by now although this length has not been reached yet by all potential members of a sample as not all of the books are already translated into every language.[13] Nevertheless, even the first book *Harry Potter and the Philosopher's stone* alone exceeds the length of *Le petit prince* by 230%. Furthermore, *Harry Potter* translations exist only for a relatively small set of languages in comparison to the impressive numbers reported for *Le petit prince*. Discounting the occasional plus for *Harry Potter* (e.g. a Greenlandic translation of the first book), the best one can have is a subset of the sample based on *Le petit prince*—again with a clear bias for European languages. Owing to the fact that regional and substandard varieties are particularly scarce for *Harry Potter*, the resulting sample is strongly standard-oriented. The languages marked by boldface in Table (1) above together with Low German and Irish form the European *Harry Potter* sample. Among these 38 languages, there are only six non-Indo-European ones (= 16%) which is only a slightly better ratio than the one observed for *Le petit prince* (where the nine non-Indo-European languages account for 14%). Galego, Ukrainian, Irish and Welsh determine the upper limit of the text length to be used in the comparative investigation because these are the languages for which only the first book has been translated so far.

The first book of *Harry Potter* is certainly sizeable enough to provide a suitable basis for a quantitative comparative study. But what about an investigation into the qualitative side of linguistic phenomena? Is it possible to uncover, say, categories and their distribution across the sample languages? For our project on possessive

[12] VAN DER AUWERA *et al.* (2005) demonstrate that a comparative linguistic study based on a Harry Potter parallel corpus is perfectly feasible.

[13] For bibliographic details for the translations of *Harry Potter* see STOLZ *et al.* (forthcoming-a).

relations in the languages of Europe, we needed to know whether languages formally distinguish between a possessive modifier (= English *my*) and a proper possessive pronoun (= English *mine*). Descriptive grammars often are not very specific about this distinction and thus we have looked in vain for pertinent information especially for what concerns the presence/absence of a possessive pronoun of the *mine*-type. In chapter 3 of Harry Potter, first book, there are two relevant clauses? in the English original. For the possessive modifier, we have *I want **my** letter* and for the possessive pronoun, *(I want to read it)…as it's **mine***. The syntactic difference between the two categories is straightforward: the possessive modifier co-occurs NP-internally with a head-noun representing the possessee whereas the possessive pronoun is itself the head and only constituent of an NP. As with *Le petit prince*, it is relatively easy to produce a concordance of sentences which are in a equivalence relation among each other (STOLZ *et al.* 2003). Nine languages fail to fulfil the necessary conditions because they use constructions which do not fit the above description.

Luckily, the remaining 29 languages use two syntactically different construction types, as shown in (2)—for a full listing, see Appendix 2. Boldface is used as above. The example from English in (2a) contains a formal distinction of the two categories whereas the example from Catalan in (2b) does not keep MY and MINE formally apart.

(2) a. English (Germanic)

I want	***my***	*letter*		*... as it's*	***mine***	

b. Catalan (Romance)

Vull	*la*	***meva***	*carta*	*... que*	*és*	***meva***
want	DET.F	**my:F**	letter	… that	is	my:F

c. Slovenian (Slavic)

Ho⌊em	***svoje***	*pismo*		*... saj*	*je*	***moje***
want:1SG	**REF:NT**	letter		…because	is	**my:NT**

The example from Slovenian in (2c) appears to be a minority subtype of identity. As a matter of fact, Slovenian, Czech, Slovak and Estonian share the rule according to which there is a general possessive modifier for all those cases where the clause subject and the possessor are identical. In the MINE-versions however, subject-possessor co-referentiality is blocked and thus a different construction has to be used—a construction which specifies the possessor person. These forms then are identical to the ones to be used as possessive modifiers in sentences without subject-possessor identity. For Finnish, the possessor is marked twice in the NP that contains the possessee—the possessor suffixes cannot occur in the MINE-version because there is no host available. The possessive modifier, however, has word status itself and thus also occurs in the MINE-version. Taking the (2b) and (2c) cases together, the percentages for the two groups are almost equal: 48% for MY ≠ MINE and 52% for MY = MINE.

These bits of knowledge can be retrieved relatively easily from the extant descriptive literature while it takes some effort to come to similar results by the strictly corpus-based analysis. For the sake of the argument, let us pretend that we do not know what the grammars could tell us about the sample languages. With a view to verifying whether or not there is a MY-MINE-distinction at all and if so, whether the distinction is compulsory, two sentences are not enough, of course. However, frequency is a factor that should not be underestimated. In this case, proper possessive pronouns occur seldom enough in the text whereas possessive modifiers are commonplace. The next example of the proper possessive pronoun is to be found 41 pages further in chapter 5: pronominal possessor *All **yours** ... (smiled Hagrid)* versus nominal possessor *All **Harry's** ... (it was incredible)*. And again, the languages present a variety of solutions. Of the problematic cases listed in (6), the two insular Scandinavian languages Faroese and Icelandic remain mysteries because the translators avoid the pronominal strategy. Instead, a predicative possessive construction similar to English *to own* is employed. A BELONG-construction (in some cases only for one of the two additional sentences) is attested for Czech, Ukrainian, Latvian, German, French, and Welsh. Thus, we cannot say anything definite about Welsh either. Polish, and Rumanian use completely different constructions for one of the sentences. However, the additional evidence helps to clarify the position of Danish which behaves (expectedly) like Swedish and Norwegian, i.e. it belongs to the type exemplified by Catalan. The same holds for the intermediate Slavic cases Croatian, Serbian and Russian, all of which turn out to follow the pattern of Slovenian. Likewise, Lithuanian displays properties of Slovene. Of all the problematic cases, only Rumanian can be shown to belong to the same class as English. For none of the other languages is there compelling evidence that would justify a reclassification.

4. Conclusions

Working with parallel corpora in typological linguistics has its limitations when we simply try to adopt the principles of the grammar-perusal method. While, in the latter case, one searches for information about a given phenomenon in chapters with similar subtitles in the grammars of as many languages as possible, the search in parallel corpora focuses on checking things in the same sentence in many languages. The above case studies suggest that there are several factors which might cause confusion. These problems notwithstanding, the sentence-by-sentence concordance is perfectly viable method which helps to uncover patterns including those of language-internal variation. The method, however, depends crucially upon the availability of a sizeable number of equivalent sentences in which a given phenomenon is attested in order to determine how to interpret consistency and variation. Without any doubt, parallel corpora are excellent bases for investigations inspired by quantitative typology (ALTMANN & LEHFELDT 1973). All kinds of interesting questions can be tackled with a statistically sound methods of quantitative linguistics (BEST 2001). To some extent, frequency counts also allow us to formulate hypotheses about the markedness values of phenomena. The identifica-

tion of correlations between categories are also in the scope of quantitative investigations.

If both sentence-by-sentence concordance and quantitative methods fail to meet all of our expectations, one might ask whether working typologically with parallel corpora should better be done in a different way. An alternative that suggests itself is the following: *in lieu* of going through the texts sentence by sentence, a full-blown corpus analysis should be carried out separately for each of the various sample languages—and only after their completion, the results of these separate studies can be compared to each other in order to allow for generalisations. This approach requires the application of the principles of corpus linguistics (BIBER *et al.* 1998) first whereas proper typological or universalist-minded criteria may then be applied to the results of the corpus analyses.

However, if the researcher aims at comprehensiveness, none of the above options alone can guarantee that one ever comes near this goal. Only a combination of many and diverse sources of information will allow us to gain sufficiently secure insights into the nature of human languages. Parallel literary corpora are a long overdue and valuable addition to the toolkit of empirical linguistics but they do not necessarily replace any of the more traditional ways and means of cross-linguistic research.

Abbreviations

DET determiner, F feminine, INS instrumental, M masculine, NT neuter, PL plural, REF reflexive.

Appendix 1

A. Languages with one comitative/instrumental relation [51 languages out of 64]

A.1. Germanic phylum [13 languages (out of 14)]

Alsatian	*D'rno het'r sini Stirn* ***mit*** *[me rotkàrrierte Nàstüech]*^{TOOL} *àbg'wischt.*
Danish	*Så tørrede han sig i panden* ***med*** *[et rødternet lommetørklæde]*^{TOOL}.
Dutch	*Toen veegde hij zich het voorhoofd* ***met*** *[een roodgeruite zakdoek]*^{TOOL}.
Faroese	*Hann turkaði sveittan av enninum* ***við*** *[einum reyðpuntutum lummaklúti]*^{TOOL}.
Frisian	*Doe switfage er syn foarholle* ***mei*** *[in rearútsjese bûsdoek]*^{TOOL}.
German	*Dann trocknete er sich die Stirn* ***mit*** *[einem rotkarierten Taschentuch]*^{TOOL}.
Icelandic	*Síðan þerraði hann sér um ennið* ***með*** *[rauðtiglóttum vasaklút]*^{TOOL}.
Letzebuergesh	*an sech duerno d'Stir* ***mat*** *[engem routkaréierten Duch]*^{TOOL} *ofgebotzt.*
Limburgian (North)	*Doe vaegdje hae ziene kop aaf* ***mèt*** *[eine roeëje, geroete tesseplak]*^{TOOL}.
Limburgian (South)	*Doew vreef heë zich d'r kop drueg* ***mit*** *[inne roewe gerüdde sjnoefplak]*^{TOOL}.
Norwegian	*Etterpå tørket han pannen* ***med*** *[et rødrutet lommetørkle]*^{TOOL}.
Swedish	*Sedan torkade han svetten ur pannan* ***med*** *[en rödrutig näsduk]*^{TOOL}.
Yiddish	*Nokh dem hot er zikh opgevisht dem shtern* ***mit*** *[a royt-kvadratn tikhl]*^{TOOL}.

A.2. Romance phylum [16 languages (out of 20)]

Aragonese	*Dimpués s'ixugó a fren* ***con*** *[un moquero de cuadros royos]*^{TOOL}.
Asturian	*Llueu llimpióse la frente* ***con*** *[un pañuelu pintu]*^{TOOL}.
Badiota	*Y* ***cun*** *[n fazorel da cadrí cöci]*^{TOOL} *s'âl spo assuié ía la frunt.*
Catalan	*Després s'eixuga el front* ***amb*** *[un mocador de quadres vermells]*^{TOOL}.
Corsican	*Dopu s'asciuvò u fronte* ***incù*** *[un mandigliulu quadrittatu rossu]*^{TOOL}.
French	*Puis il s'épongea le front* ***avec*** *[un mouchoir à carreaux rouges]*^{TOOL}.
Friulan	*Po al sujà il cernêli* ***cun*** t_1*[un fassolet a cuadris ros]*^{TOOL}.

Galego	*Despois enxugou a fronte* ***c*** *[un pano de cadros vermellos]*[TOOL].
Gascognian	*Puish que's boishè lo temp* ***dab*** *[un mocader de quarrèus rotges]*[TOOL].
Gherdëina	*L s'ova pò suiá jú l fruent* ***cun*** *[n fazulët da chedri cueceni]*[TOOL].
Italian	*Poi si asciugò la fronte* ***con*** *[un fazzoletto a quadri rossi]*[TOOL].
Languedocien	*Puèi se freguèt lo front* ***amb*** *[un mocador dels carrèus roges]*[TOOL].
Portuguese	*Depois enxugou a testa* ***com*** *[um lenço aos quadrados vermelhos]*[TOOL].
Provençal	*Pièi s'espounguè lou front* ***em*** *'[un moucadou di carrèu rouge]*[TOOL].
Sardinian	*Tando s'at assuttadu su sudore de cara* ***chin d'*** *[unu muccadore a quadros rujos]*[TOOL].
Spanish	*Luego se enjugó la frente* ***con*** *[un pañuelo a cuadros rojos]*[TOOL].

A.3. Slavic phylum [8 languages (out of 11)]

Bielorussian	*Potym vytser uspatsely lob [čyrvon****aj*** *kljatčast****aj*** *nasowk****aj****]*[TOOL]
Bulgarian	*Setne izbărsa čelo* ***s*** *[edna kărpa na červeni kvadrati]*[TOOL].
Czech	*Potom si otřel čelo [červen****ě*** *kostkovan****ým*** *kapesník****em****]*[TOOL].
Macedonian	*Potoa go izbrišal čeloto* ***so*** *[edno karirano tsrveno ša miče]*[TOOL].
Russian	*Potom [krasn****ym*** *kletčat****ym*** *platk****om****]*[TOOL] *utjor pot so lba i skazal*:
Serbian	*Zatim obrisa čelo [tsrven****om*** *karиran****om*** *maramits****om****]*[TOOL].
Slovenian	*Nato si je* ***z*** *[rdeč****e*** *kockast****im*** *robc****em****]*[TOOL] *otrl čelo*.
Ukrainian	*Potim [kartatoj****u*** *červonoj****u*** *xustynkoj****u****]*[TOOL] *vyter z litsja pit i skazav*:

A.4. Minor Indo-European phyla [6 languages (out of 10)]

Albanian (Gheg)	*Mandej fshiu ballin* ***me*** *[nji faculetë të kuqe, kutija-kutija]*[TOOL].
Armenian	*Heto [karmir vandakavor taškinak****ov****]*[TOOL] *čakati k'rtnink'6 srbec' w asac':*
Breton	*Hag e sec'has e dal* ***gant*** *[ur frilien karrezennoù ruz]*[TOOL].
Latvian	*Pēc tam viņš noslaucīja no pieres sviedrus* ***ar*** *[sārti rūtotu kabatas lakatiņu]*[TOOL].
Lithuanian	*Paskui [raudon****a*** *languot****a*** *nosin****e****]*[TOOL] *nusišluostė kaktą.*
Romany (Lovari)	*Palakodi [jekha posotyake kotores****a****]*[TOOL] *khoslas pesko chikat*.

A.5. Non-Indo-European phyla [8 languages (out of 9)]

Azeri	*Sonra [qırmızı dama-dama däsmal****la****]*[TOOL] *alnının tärini silib dedi*.
Basque	*[Sudur-zapi gorri-koadratu bat****ez****]*[TOOL]*, bekokiko izerdia txukatu zuen*.
Estonian	*Siis ta pühkis [punaseruudulise taskurätiku****ga****]*[TOOL] *oma otsaesist*.
Finnish	*Sitten hän pyyhkäisi hien otsaltaan [punaruutuise****lla*** *nenäliina****lla****]*[TOOL].
Georgian	*shemdeg [c'iteludjredebiani cxvirsaxoc****it****]*[TOOL] *shublze opli sheimshrala da tkva*:
Hungarian	*Aztán [egy piros kockás zsebkendö****vel****]*[TOOL] *törölgetni kezdte a homlokát.*
Saami	*Son sihkui bivastaga [gállus ruksesruvttot njunneliinn****iin****]*[TOOL]**.**
Turkish	*Sonra [kırmızı kareli bir mendil****le****]*[TOOL] *alnını sildi.*

B. Languages with two comitative/instrumental relations [13 languages out of 64]

B.1. One polysemous marker [10 languages]

English	*Then he mopped his forehead* ***with*** *[a handkerchief decorated* ***with*** *[red squares]*[ORNATIVE]*]*[TOOL].
Moldavian	*Apoi îşi şterse fruntea* ***cu*** *[o batistă cadrilată* ***cu*** *[roşu]*[ORNATIVE]*]*[TOOL].
Rumanian	*Apoi îşi şterse fruntea* ***cu*** *[o batistă* ***cu*** *[pătrăţele roşii]*[ORNATIVE]*]*[TOOL].
Surselvian	*Lu schigenta el siu frunt* ***cun*** *[in fazalet* ***cun*** *[quadrels cotschens]*[ORNATIVE]*]*[TOOL].
Vallader	*Lura ha'l süantà seis frunt* ***cun*** *[ün fazöl* ***cun*** *[quaders cotschens]*[ORNATIVE]*]*[TOOL].
Albanian (Tosk)	*Pastaj fshiu ballin* ***me*** *[një shami* ***me*** *[kutia të kuqe]*[ORNATIVE]*]*[TOOL].
Greek	*Épeita skoúpise to métōpó tou* ***m'*** *[éna mantēli* ***me*** *[kókkina karrō]*[ORNATIVE]*]*[TOOL].
Kurdish	*Paşê wî xwêdana aniya xwe* ***bi*** *[destmaleke* ***bi*** *[damikên sor]*[ORNATIVE]*]*[TOOL] *zu ha kir.*
Welsh	*Yna sychodd ei dalcen* ***â*** *[chadach* ***a*** *[sgwarau cochion arno]*[ORNATIVE]*]*[TOOL].
Maltese	*Imbagħad mesaħ moħħu* ***b'*** *[maktur* ***bi*** *[l-kaxxi ħomor]*[ORNATIVE]*]*[TOOL].

B.2. Two specialised constructions [3 languages]

Croatian	*Zatim obrise celo [rupčić**em**$_1$ **s**$_2$ [crven**im**$_2$ kvadrat**ima**$_2$]^{ORNATIVE}]^{TOOL}.*
Slovak	*Potom si utrel čelo [vreckovk**ou**$_1$ **s**$_2$ [červený**mi**$_2$ kock**ami**$_2$]^{ORNATIVE}]^{TOOL}.*
Polish	*Następnie otarł sobie czoło [chustk**ą**$_1$ **w**$_2$ [czerwon**ą**$_2$ krat**ę**$_2$]^{ORNATIVE}]^{TOOL}.*

Appendix 2

Grey shading indicates those cases where, notwithstanding formal differences between the two versions, the examples do not instantiate the MY-MINE-distinction. Listed here are only the (seemingly) unproblematic cases [29 languages].

A. MY ≠ MINE [14 languages]

Language	MY	MINE
French	*Je veux **ma** lettre.*	*Elle est **à moi***
Spanish	*Quiero **mi** carta*	*Es **mía***
Dutch	*Ik will **mijn** brief terug*	*... want hij is **van mij***
German	*Ich will **meinen** Brief*	*... es ist nämlich **meiner***
English	*I want **my** letter*	*... as it's **mine***
Bulgarian	*Iskam **si** pismoto*	*... t©j kato to e **do men***
Polish	*Chc\ **mój** list*	*... bo to list **do mnie***
Ukrainian	*Ja xo\u **svogo** lista*	*... bo **vin mij***
Albanian	*Dua letr[n **time***	*... [sht[**imja***
Latvian	*Atdodiet **manu** v\stuli*	*... jo t® ir **man\j®***
Irish	*Teastaíonn an litir **uaim***	*... mar gur **liomsa** í*
Basque	***Neure** gutuna nahi dut*	*... **nirea** da eta*
Hungarian	*A level**em**et akarom*	*... mivel az **enyém***
Turkish	*Mektubu**mu** istiyorum*	*... çünkü o **benim***

B. MY = MINE [15 LANGUAGES]

Language	MY	MINE
Slovenian	*Ho\em **svoje** pismo*	*... saj je **moje***
Czech	*Chci **sv]j** dopis*	*... pon[vad je **m]j***
Slovak	*Chcem **svoj** list*	*... je **môj***
Estonian	*Ma tahan **oma** kirja*	*... sest see on **minu oma***
Finnish	*Anna tänne **minun** kirjee**ni***	*... koska se on **minun***
Catalan	*Vull la **meva** carta*	*... que és **meva***
Galego	*Quero a **miña** carta*	*... porque é **miña***
Italian	*Voglio la **mia** lettera*	*... è **mia***
Portuguese	*Quero a **minha** carta*	*Ela é **minha***
Low German	*Ik will **mien** breef hebben*	*... denn dat is **mien***
Norwegian	*Jeg vil ha brevet **mitt***	*... det er nemlig **mitt***
Swedish	*Jag vill ha **mitt** brev*	*... eftersom det är **mitt***
Greek	*Thél| to grámma **mou***	*Aphoú eínai dikó **mou***
Georgian	*Momecit **emi** c'erili*	*... c'erili **emia***

References

ALTMANN, GABRIEL & LEHFELDT, WERNER (1973): *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren*. München: Fink.

BEST, KARL-HEINZ (2001): *Quantitative Linguistik—eine Annäherung*. Göttingen: Peust & Gutschmidt.

BIBER, DOUGLAS; CONRAD, SUSAN & REPPEN, RANDI (1998): *Corpus linguistics. Investigating language structure and use*. Cambridge: Cambridge University Press.
BRAUNER, SIEGMUND (1974): *Lehrbuch des Bambara*. Leipzig: Enzyklopädie Verlag.
BRODY, JILL (1998): On Hispanisms in elicitation. In: KOECHERT, ANDREAS & STOLZ, THOMAS (eds.), *Convergencia e individualidad. Las lenguas mayas entre hispanización e indigenismo*. Hannover: Verlag für Ethnologie, 61-84.
BYBEE, JOAN L. & HOPPER, PAUL (eds.) (2001): *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.
CHUNG, SANDRA (1998): *The design of agreement. Evidence from Chamorro*. Chicago: The University of Chicago Press.
COMRIE, BERNARD (1993): Language universals and linguistic theory: data-base and explanations. *Sprachtypologie und Universalienforschung* 46 (1), 3-14.
CROFT, WILLIAM (2001): *Radical Construction Grammar. Syntactic theory in typological perspective.* Oxford: Oxford Unversity Press.
ENDZELĪNS, JĀNIS (1951): *Latviešu valodas gramatika.* Rīgā: Latvijas Valsts Izdevniecība.
ENNINGER, WERNER & RAITH, JOACHIM (1981): Linguistic modalities of liturgical registers: the case of the Old Order Amish Church Service. *Yearbook of German-American Studies* 16, 115-29.
HAARMANN, HARALD (2004): *Elementare Wortordnung in den Sprachen der Welt*. Hamburg: Helmut Buske.
HANDBOOK (1999): *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press.
HANKE, THOMAS (2005): *Bildungsweisen von Numeralia. Eine typologische Untersuchung*. Berlin: Weißensee.
HASPELMATH, MARTIN; DRYER, MATTHEW S.; GIL, DAVID & COMRIE, BERNARD (eds.) (2005): *The World Atlas of Language Structure*. Oxford: Oxford University Press.
HEINE, BERND (1997): *Possession. Cognitive sources, forces, and grammaticalization*. Cambridge: Cambridge University Press.
KAISER, GEORG (2005): Bibelübersetzungen als Grundlage für empirische Sprachwandeluntersuchungen. In: PUSCH, CLAUS D.; KABATEK, JOHANNES & RAIBLE, WOLFGANG (eds.), *Romanistische Korpuslinguistik II: Korpora und diachrone Sprachwissenschaft.* Tübingen: Narr, 71-84.
KASTENHOLZ, RAIMUND (1989): *Grundkurs Bambara (Manding) mit Texten.* Köln: Köppen.
LAFITTE, PIERRE (1998): *Grammaire basque (navarro-labourdin littéraire).* Donostia/San Sebastian: Elkarlanean.
LUNA ALONSO, ANA (2000): Contrastes estilísticos: le petit Prince en lingua galega. In: CASAL SILVA, MARIA LUZ *et al.* (eds.), *La lingüística francesa en España. Camino del siglo XXI, II: Jesús Lago Garabatos in memoriam.* Arrecife, 647-53.
PERKINS, REVERE D. (1989): Statistical techniques for determining language sample size. *Studies in Language* 13, 293-315.
RIJKHOFF, JAN & BAKKER, DIK (1998): Language sampling. *Linguistic Typology* 2, 263-314.
RUBINO, CARL (2005): Reduplication. In: HASPELMATH ET AL. 2005, 114-7.
SAKAYAN, DORA (2000): *Modern Western Armenian for the English-speaking world. A contrastive approach.* Montreal: Arod Books.
SALTARELLI, MARIO (1988): *Basque*. London: Croom Helm.
SEILER, HANSJAKOB (2000): *Language universals research: a synthesis*. Tübingen: Gunter Narr.
SIEWIERSKA, ANNA; RIJKHOFF, JAN & BAKKER, DIK (1998): Appendix—12 word order variables in the languages of Europe. In: SIEWIERSKA, ANNA (ed.), Constituent order in the languages of Europe. Berlin: Mouton de Gruyter, 783-812.
STOLZ, THOMAS (1996): Some instruments are really good companions—some are not. On syncretism and the typology of instrumentals and comitatives. *Theoretical Linguistics* 23 (1/2), 113-200.
STOLZ, THOMAS (1998): UND, MIT und/oder UND/MIT?—Koordination, Instrumental und Komitativ—kymrisch, typologisch und universell. *Sprachtypologie und Universalienforschung* 51 (2), 107-30.
STOLZ, THOMAS (2004): A new Mediterraneanism: word iteration in an areal perspective. A pilot-study. *Mediterranean Language Review* 15, 1-47.

STOLZ, THOMAS & GUGELER, TRAUDE (2000): Comitative typology—nothing about the ape, but something about king-size samples, the European community, and the little prince. *Sprachtypologie und Universalienforschung* 53 (1), 53-61.
STOLZ, THOMAS; STROH, CORNELIA & URDZE, AINA (2003): Solidaritäten. *Lingua Posnaniensis* 45, 68-92.
STOLZ, THOMAS; STROH, CORNELIA & URDZE, AINA (2005): Comitatives and instrumentals. In: HASPELMATH ET AL. 2005, 214-7.
STOLZ, THOMAS; KETTLER, SONJA & URDZE, AINA (forthcoming-a): *Split possession. An areal-linguistic study of the alienability correlation and related phenomena in the languages of Europe*.
STOLZ, THOMAS; STROH, CORNELIA & URDZE, AINA (forthcoming-b): *On comitatives and related categories. A typological study with special focus on the languages of Europe*. Berlin: Mouton de Gruyter.
TULDAVA, JUHAN (1995): *Methods in quantitative linguistics*. Trier: Wissenschaftlicher Verlag.
VAN DER AUWERA, JOHAN; SCHALLEY, EWA & NUYTS, JAN (2005): Epistemic possibility in a Slavonic parallel corpus—a pilot study. In: HANSEN, BJÖRN & KARLIK, PETR (eds.), *Modality in Slavonic languages. New perspectives*. München: Otto Sagner, 201-17.
VAUX, BERT & COOPER, JUSTIN (1999): *Introduction to linguistic field methods*. München: LINCOM Europa.
WÄLCHLI, BERNHARD (2005): *Co-compounds and natural coordination.* Oxford: Oxford University Press.
ZEMSKAJA, ELENA A. (2004): Analytische und agglutinative Tendenzen im Russischen. In: HINRICHS, UWE (ed.), *Die europäischen Sprachen auf dem Wege zum analytischen Sprachtyp*. Wiesbaden: Harrassowitz, 285-92.
ZIMMERMANN, KLAUS (1997): Introducción: apuntes para la historia de la lingüística de las lenguas amerindias. In: ZIMMERMANN, KLAUS (ed.), *La descripción de las lenguas amerindias en la época colonial.* Frankfurt a.M.: Vervuert, 9-17.

Correspondence address

Thomas Stolz
FB 10: Linguistik
Universität Bremen
PF 330 440
D-28334 Bremen
stolz@uni-bremen.de

BERNHARD WÄLCHLI (Leipzig)

Advantages and disadvantages of using parallel texts in typological investigations[1]

In this paper, advantages and disadvantages of using parallel texts in typological studies are considered according to the criteria of diversity, domains, analysis, perspective, quality, representativity, and comparability. It is shown in a case study of multi-verb constructions (including serial verb constructions, converb constructions, etc.) in two motion event domains (BRING and RUN) how typology can profit from parallel texts especially in the investigation of quantitative variables. A method is introduced to transform features with continuous distributions into ternary features with low, intermediate, and high values which can then be tested for correlations.

1. Introduction

Typology has often been criticized for the bad quality of the data used. Consider a particular case of such critique—NEWMEYER's (1998: 329f) discussion of STASSEN's (1985) typology of comparative constructions:

> "Specialists [...] have pointed out to me, however, that Classical Greek, Latin, and Classical Tibetan [...] manifest a wide range of comparatives of the 'Exceed' type. How could Stassen have missed noting this fact about the two former languages, which are both in his sample? Reliance on secondary sources is to blame—the existence of the Exceed Comparative in these languages is virtually never mentioned in their published grammars. The reasons for their omission are not difficult to understand: for one thing, verbal constructions are quite often discussed exclusively in the context of the adjective. What this means is that Stassen probably greatly underestimates the full range of possibilities for comparison in the world's languages [...]. Now Stassen cannot be faulted personally for not having taken the time to actually learn all the languages in his sample, instead of merely thumbing through the odd grammars. *Nobody* has that kind of time. But if he had done so, one feels that he would have ended up with a radically different set of statements concerning the universals of comparative constructions from that which he proposes in his book. In sum, reference to secondary sources and reliance on consultants in typological research may be more than a necessary evil—it may point to the shaky foundations of the entire enterprise."

However, parallel texts indicate that the situation is more in line with Stassen's classification. In a set of 12 instances in which a comparative construction can be found in the Gospel according to Mark (henceforth Mark), none of them is an example of the Exceed Comparative (standard of comparison marked by a verb such as '(sur)pass') in English, Classical Greek, Latin, or Written Tibetan. The English examples are given in (1) with the markers of comparison presented in boldface.

(1) Comparative constructions in Mark in Early Modern English (King James)
*1:7 ...mighti**er than** I...; 4:31. ...**less than** all the seeds...; 4:32 ...great**er than** all herbs...; 8:14 ...**more than** one loaf; 9:43 ... it is **better** for thee to enter...**than**...to go...; 9:45 ...it is **better** for thee to enter...**than**...to be cast...;*

[1] I would like to thank MICHAEL CYSOUW for many useful comments. My research is supported by the Swiss National Science Foundation (2004-6, Nr. 001-104983 "The encoding of displacement in the languages of the world").

*9:47 ...it is **better** for thee to enter into...**than**...to be cast...; 10:25 It is eas**ier** for a camel to go...**than** for a rich man to enter...; 12:31 ...great**er than** these; 12:33 ...**more than** all whole burnt offerings...; 12:43 ...hath cast **more** in, **than** all they...; 14:5 ...might have been sold for **more than** three hundred pence.*

The Exceed Comparative, however, does occur systematically in other translations. For example, in Haitian Creole it is found in all 9 instances in which a comparative construction is used in the translation. One example is shown in (2).

(2) Exceed comparative in Haitian Creole [Mark 10:25]

Lap	*pi*	*fasil*	*pou*	*gro*	*bèt*	*yo*	*rélé*	*chamo-a*
it:PROG	more	easy	for	big	animal	they	call	camel-DEF
pasé	*nan*	*jé*	*you*	*zégoui*	***pasé***	*pou*	*you*	*moun*	*rich*
pass	in	eye	one	needle	pass	for	one	person	rich
antré	*nan*	*péyi*	*koté*	*Bondié*	*Roua-a*				
enter	in	land	side	God	kingdom-DEF				

'It is easier for a camel to go through the eye of a needle, than for a rich man to enter into the kingdom of God.'

Unfortunately, NEWMEYER does not tell which constructions have been pointed out to him, but it is obvious from looking at the data from parallel texts that Exceed Comparatives must be rare (see also ANDERSEN 1983: 131) in Greek, Latin, and Written Tibetan. The dominant constructions are of the Separative type (the standard of comparison is in the Ablative in Latin, in the Genitive in Classical Greek and marked by the Ablative *las* in Tibetan). However, the parallel texts also indicate language internal diversity. All three languages have an alternative construction where the standard of comparison is a clause, marked by the Particle construction in Latin (10:25 ***quam** divitem intrare in regnum*) and Classical Greek (*ḗ* 'than'), and by *bas* in Written Tibetan (consisting of a nominalizer *ba* in the Instrumental case). The parallel text material thus suggests that comparative encoding in Latin, Greek and Tibetan is split, while being more consistent in English and Haitian Creole.

In this simple example several advantages of using of parallel texts have become manifest. The question whether a certain construction type is present in a particular language cannot be answered negatively on an empirical basis, one can always have missed some rare examples and NEWMEYER plays with this fact. However, most typological investigations are implicitly or explicitly about frequently instantiated constructions and dominant construction types, which is much firmer ground from an empirical point of view.

Linguistic structure cannot be accessed directly, it can be investigated only in particular utterances and so linguistic typology is always a typology of texts. Parallel texts allow for a strict definition of typological domains by extension (translation equivalents of a certain number of particular clauses in a text which instantiate a semantic domain) rather than by intension (abstract semantic definition of a domain). In practice, domains should always consist of several places in order to

minimize accidental bias. The extensional domains in parallel text studies are thus internally complex and allow for an investigation of the internal consistency of a chosen domain. The parallel text method shares some of these properties with the questionnaire method, which has been used more often in typology (see, e.g., DAHL 1985). However, questionnaire studies are dependent on informants and this strongly limits the number and diversity of languages that can be considered. We know from recent developments in typology and especially areal typology that large and diverse samples are needed.

In spite of many available translations, typology has little experience with using parallel texts.[2] So the title of this paper is actually premature: it is still unknown how valuable parallel texts can be in typological investigations. Also, when I speak in this paper of the 'parallel text method' the reader should be aware that there is no such thing as an established single method. Parallel texts simply lend themselves for certain kinds of analysis which cannot be done as easily with other kinds of material. There is only one way to find out how valuable parallel texts can be in typological investigations: we must try. I have made use of parallel texts in typological studies in several ways essentially due to a lack of other possibilities to address certain research questions, notably in investigating co-compounds (WÄLCHLI 2005), 'again' expressions (WÄLCHLI 2006), and some aspects of motion events (WÄLCHLI 2001, WÄLCHLI & ZÚÑIGA forthcoming). But rather than summarizing results published elsewhere I would like to present another investigation here to illustrate the parallel text method. In Section 2, I will present some first results from an investigation of multi-verb constructions in two lexical domains of motion events. Following this example, I will discuss some advantages and disadvantages of the method in more general terms in Section 3.

2. Multi-verb constructions in motion events. A case study

In this section, two lexical domains of motion events are discussed where multi-verb constructions based on motion verbs are common, (a) directed transport (BRING), and (b) directed race (RUN). It is shown in this particular example how typology using parallel text data can deal with non-discrete variables and how the cross-linguistic consistency of a feature can be tested. A method is introduced to transform features with continuous distributions into ternary features with low, intermediate, and high values which can then be tested for correlations.

2.1. Multi-verb constructions

Multi-verb constructions (MVCs) are clauses that contain more than one lexical verb irrespective of the type of chaining between the verbs. In the two domains considered, the second verb is mostly 'go' or 'come'. Auxiliaries expressing TMA categories and other meanings not related to motion events (even if deriving from motion verbs) are disregarded. Put differently, only lexical multi-verb construc-

[2] According to HASPELMATH (1997: 17) translations of the New Testament are an innovative source of data in typology "which has not to my knowledge been made use of in typological work before".

tions are considered, multi-verb constructions with grammatical or modal functions are not considered. So, for example, English *is running, is going to run, will run, wants to run, starts running* will not be considered MVCs here.

Examples (3)-(9) illustrate various kinds of chaining in directed transport: verb serialization (3) and (4), overt coordination (5), converb construction (6), medial-final chaining (7), and root serialization (8) and (9). An English (King James) translation is given only for the first example since all examples are from the same place in Mark [9:19], the parallel text serving as material for this study. Verbs are marked boldface.

(3) Haitian Creole (French-based creole) [Mark 9:19]

Minnin	*ti-bouay*	*la*	***ban***	*mouin.*
lead	little-boy	DEF	give	I

'...bring him unto me.'

(4) Yabem (Austronesian)

*...**a-kôc***	*eŋ*	***a-n-dêŋ***	*aê*	***a-mêŋ**.*
2PL-take	he	2PL-IRR-go.to	I	2PL-come

(5) Moore (Niger-Congo, Gur)

***Tall**-y*	*biigã*	*n*	***wa***	*ka.*
transport-2PL	child	and	come	here

(6) Chuvash (Turkic)

Ač-i-ne	*Man*	*pat-ăm-a*	***il-se***	***kil-ĕr**.*
child-POSS3-DAT/ACC	I.GEN	to-POSS1SG-DAT	take-CONV	come-IMP2PL

(7) Choctaw (Muskogean)

Isht	*hʊs*	*sʊm*	***ʊla**.shke,*	*achi*	*tok*
take.'NOM'	2PL'NOM'	I.DAT	come.to-INTENS	say	REM.PST

(8) Khoekhoe/Nama (Khoisan)

Tita	*!oa*	***u-ha***	*bi!*
I	to	take-come	he.OBJ

(9) Khasi (Austro-Asiatic)

...to	***wal-lam***	*ia*	*u*	*ha*	*nga*
IMP	come-lead	OBJ	he	to	I

Clauses lacking multi-verb constructions (where other languages have multi-verb constructions) are called **verb solitarizing** (a term coined by GIL 1999).[3] Here I have to come back to the notion of 'clause' as used in the definition of multi-verb constructions above. Clauses are viewed here as functional rather than purely structural units, as far as they occur within a single sentence. A clause is a sequence within a sentence that is a recurrent translational equivalent of a verb-solitarizing construction. Even if the terms clause and verb solitarization as I use them refer to each other, this definition is not circular since verb solitarizing constructions can be easily established in the considered domains in parallel texts. Translations having always verb solitarizing constructions in the two domains are, for example, Russian and Navajo. English, however, even if strictly verb solitarizing in many domains, is not fully solitarizing in the RUN domain, which can be seen in the example as shown in (10).

(10) English [Mark 10:17]
*...there **came** one **running**, and kneeled to him...*

If we now compare the two domains BRING and RUN, we find that there is no implicational universal. Multi-verb constructions in the two domains are not obviously dependent on each other. Some examples are shown in Table 1.

Table 1. Cross-linguistic diversity in multi-verb constructions.

		BRING	
		Verb solitarizing	Multi-verb constructions
RUN	Solit.	Dinka, Navajo, Russian	Ainu, Ewe, Khasi
	MVC	English, Guaraní, Maltese	Choctaw, Chuvash, Khoekhoe

2.2. Data collection

Can it be concluded from Table 1 that the two domains are completely unrelated? No, let us have a closer look. First of all, we have to choose sets of clauses and a sample of languages. As for sampling, the parallel text method is different from other typological studies in that the possible diversity of the sample is more limited by the availability of parallel texts than is the case when using reference grammars. Here, a convenience sample with a strong Eurasian bias consisting of 165 languages (listed in Table 2) has been chosen. Also, the notion 'language' is very narrowly defined as the variety used in the chosen texts.

[3] The underlying idea is that it is not at all clear that serialization is the special case and that non-serializing languages are the normal case. It might just as well also be the other way round. Actually, languages without any multi-verb constructions seem to form a minority.

Table 2. Sample of languages.

Continent *	Languages	No. of lang.
Africa	Acholi, Akan (Twi), Bambara, Bari, Dinka, Efik, Ewe, Hausa, Igbo, Ijo, Kabba-Laka, Kabiyé, Khoekhoe (Nama), Koalib, Kunama, Maltese, Moore, Moru, Murle, Ngambay, Nubian (Kunuz), Pokot (Suk), Sango, Shilluk, Somali, Songhay, Swahili, Yoruba, Zulu	29
Eurasia	Adyghe, Ainu, Albanian, Armenian (Classical), Avar, Basque, Breton, Bulgarian, Chuvash, English, Estonian, Finnish, French, Garo, Georgian (Classical), Georgian (Modern), German (Bernese), Greek (Classical), Greek (Modern), Hindi, Hungarian, Icelandic, Irish, Italian, Kannada, Khalkha Mongolian, Khasi, Komi, Korean, Kurdish (Kurmanji), Lak, Latin, Latvian, Lezgian, Lithuanian, Livonian, Mansi, Mari (Eastern), Mordvin (Erzya), Naga (Tangkhul), Ossetic, Rhaeto-Romance, Romani (Kalderash), Rumanian, Russian, Saami (Northern), Santali, Spanish, Swedish, Tabassaran, Tadzhik, Tamil, Tibetan, Turkish, Tuvan, Udi, Udmurt, Veps	58
SEA & East Asia	Burmese, Cebuano, Chamorro, Fijian, Hawaiian, Hmar, Hmong Njua, Indonesian, Khmer, Lahu, Malagasy, Maori, Marshallese, Mizo, Nicobarese (Car), Ponapean, Samoan, Tagalog, Thai, Timorese (Atoni), To'aba'ita, Toba Batak, Tongan, Ulawa, Vietnamese, Yabem	26
NG & Austr	Burarra, Gumatj, Kâte, Kuku-Yalanji, Kuot, Nunggubuyu, Pitjantjatjara, Toaripi, Tobelo, Waris, Warlpiri, Wik Munkan, Worora	13
N Amer	Cakchiquel, Choctaw, Comanche, Cree (Plains), Dakota, Hopi, Huichol, Inuktitut (Labrador), Mixe (Coatlán), Mixtec (San Miguel el Grande), Muskogee (Creek), Navajo, Ojibwa, Otomí (Mezquital), Purépecha (Tarascan), Totonac (Sierra), Trique, Yucatec Maya, Zapotec (Isthmus), Zoque (Copainalá)	20
S Amer	Aymara, Bribri, Chiquitano, Guaraní, Kuna, Mapudungun, Miskito, Ngäbere (Guaymi), Paumarí, Piro, Quechua (Imbabura), Shipibo, Yanesha'	13
Creole	Haitian Creole, Australian Kriol, Papiamentu, Seychelles Creole, Sranan, Tok Pisin	6

* Continents do not correspond strictly to geographical continents but take into account large genealogic groupings. Thus, Maltese belongs to Eurasia and Malagasy to South East & East Asia.

Further, defining a domain in parallel text studies is different from defining a domain in a reference grammar study. Rather than defining the domain in semantic terms (by intension), the domain is defined as a selection of places in the parallel text which instantiate the intended semantic domain (by extension). Table 3 gives the eighteen places for BRING and the six places for RUN that constitute the two domains in our parallel text study. The different number of clauses is simply due to the fact that BRING is more often represented in the text whereas for RUN all possible examples are taken (the 'flee/run away' domain has not been included). This difference in number of clauses does not create any difficulties for the method used below.[4]

[4] With hindsight, it might have been better to be more restrictive and to exclude 6:55 in the RUN domain which represents undirected rather than directed race.

Table 3. The two multi-verb domains defined by extension as places in Mark.

BRING		RUN	
1:32	*they brought unto him all that were diseased*	5:6	*he ran and worshipped him*
2:03	*bringing one sick of the palsy*	6:33	*and ran afoot thither out of all cities*
6:27	*and commanded his head to be brought*	6:55	*And ran through that whole region round about*
6:28	*And brought his head in a charger*	9:15	*and running to him saluted him*
7:32	*And they bring unto him one that was deaf*	10:17	*there came one running, and kneeled to him*
8:22	*and they bring a blind man unto him*	15:36	*And one ran and filled a spunge full of vinegar*
9:17	*I have brought unto thee my son*		
9:19	*bring him unto me*		
9:20	*And they brought him unto him*		
10:13	*And they brought young children to him*		
11:02	*and bring him*		
11:07	*And they brought the colt to Jesus,*		
12:15	*bring me a penny*		
12:16	*And they brought it*		
15:01	*and carried him away*		
15:16	*And the soldiers led him away into the hall*		
15:20	*and led him out to crucify him*		
15:22	*And they bring him unto the place Golgotha*		

2.3. Some first results

First, we consider only whether there is any multi-verb construction (MVC) in a language; that is, a single occurrence is sufficient for a language to be categorized as having MVC. The results of such a classification for all 165 languages in the sample are shown in Table 4. The distribution is highly significant (Fisher's exact $p < 0.001$), indicating that there is a statistical universal between the two domains BRING and RUN. However, the proportion of non-consistently solitarizing or MVC languages is quite large: 46 + 12 = 58 of the 165 languages (or 35%) behave differently for the two domains.

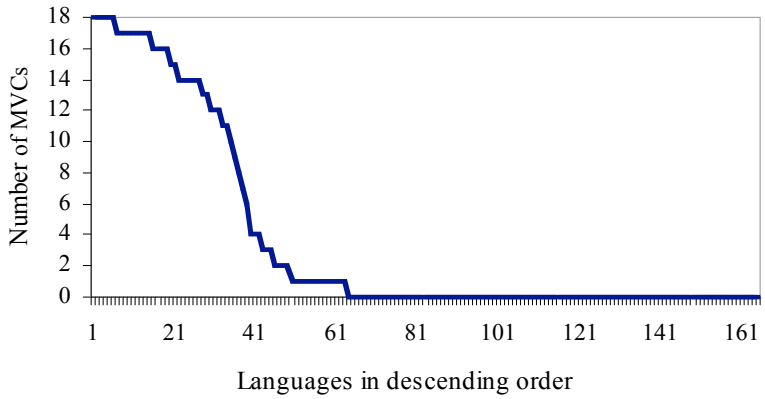
Table 4. Availability of MVC in the two domains.

		BRING	
		Solit.	MVC
RUN	Solit.	65	12
	MVC	46	42

While Table 4 shows that MVC in the two domains are not distributed randomly, I have actually not shown yet whether or not the two MVC domains are consistent

features from a cross-linguistic point of view. Considering whether or not a property occurs in a domain is useful only when this property represents a discrete binary feature (the classification always goes one or the other way in a given language). Multi-verb constructions in the two domains are far from being a discrete feature, there is a continuous distribution between fully solitarizing and fully MVC languages, without any clear cut-off line as can be seen for BRING in Figure 1.[5] In the BRING domain, there are many Intermediate values (57 out of 165 languages). In the RUN domain there are even more Intermediate values (in 85 out of 165 languages).

Figure 1. Number of MVC per language in the BRING domain (languages are ordered in descending order of the number of MVC).



The question now is whether multi-verb constructions actually are a feature in the two domains. This will be the case if the distribution is bipolar (higher than expected frequency at the left and right edges). It is assumed that a random distribution of MVCs over the clauses would result in a binomial distribution (see CYSOUW 2002: 74-77 for a related problem). Figure 2 shows that MVC is bipolar in the BRING domain. The value zero on the left side and the observed values above ten on the right side are more frequent than expected. The crossing points between the observed and the expected distributions give us two non-arbitrary cut-off points, which is how the domains are transformed into a feature with three values: High, Intermediate, and Low. Note that Low does not necessarily mean complete absence of the feature. In the BRING domain the crossing point of the lines is between one and two, which is why Low is defined as zero or one instance of MVCs.

[5] Even if there is good reason to call this a continuous variable from the linguistic point of view, statistically we have to do here strictly speaking with discrete measures (occurrence or non-occurrence of MVC in various places in the parallel texts are counted) and the data has undergone a first step of reduction, viz. addition. See CYSOUW (2002: 74) for discussion.

Figure 2. Bipolar structure of the BRING domain (the line shows expected frequencies, the bars show the actual data).

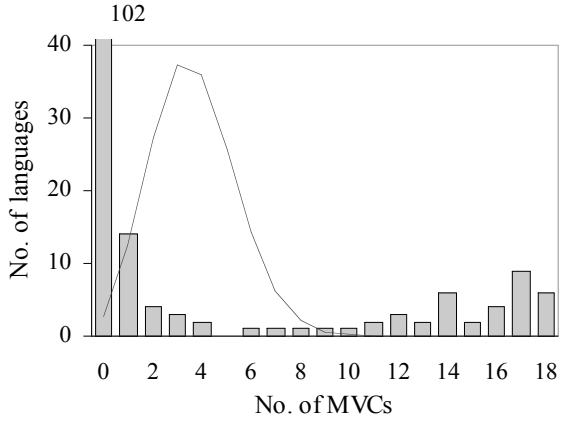


Table 5 gives the number of languages for each value. In brackets the differences to expected values are given. The correlation between MVCs in the two considered domains emerges more clearly when only the extremes are considered (the number of Intermediate cases are all close to the statistical expectation anyhow). Among Low and High values only 14 of 104 languages (or 13%) are non-consistent.

Table 5 substantiates the impression that RUN has more Intermediate values than BRING. The percentage of Intermediate cases is much smaller for BRING (8%) than for RUN (33%). Also areality shows that BRING is a sharper typological feature. MVCs in BRING cluster strongly at various places in the Old World: West Africa (including Haitian Creole and Sranan), South-East, East, and South Asia, and Eastern New Guinea. Intermediate values occur especially at the border of High and Low areas.

Table 5. Number of languages according to MVCs in the two domains BRING and RUN (deviation from statistical expectation in brackets).

		BRING		
		0-1 (Low)	2-8 (Intermed.)	9-18 (High)
RUN	0 (Low)	70 [+15.9]	3 [-3.0]	4 [-12.8]
	1-3 (Intermed.)	36 [-2.0]	6 [+1.7]	12 [+0.2]
	4-6 (High)	10 [-13.9]	4 [+1.3]	20 [+12.6]

2.4. Summary

Investigating variables, such as multi-verb constructions, in various domains in large language samples is important because it shows that linguistic structure is often more irregular cross-linguistically than would have been expected from systematic descriptions in grammars, while at the same time not being randomly distributed but exhibiting strong correlations. The results of this section suggest that multi-verb constructions do not behave parametrical. In other words, languages cannot be said simply to exhibit or lack multi-verb constructions.[6]

It is clear that BRING and RUN are just two of many domains where multi-verb constructions tend to occur. In order to make sure that they correlate (and that multi-verb constructions and its counterpart, verb solitarizing, are consistent cross-linguistic features), all these different domains would have to be investigated as quantitative variables in turn. The purpose of this section has been to show that this can only be done on the basis of quantitative data (since MVC is no discrete variable) and that parallel text studies are a possible way to do this. It has not been shown, however, whether the data used is good enough for this purpose (i.e., whether the texts are representative for the languages they instantiate). The result, however, seems promising, given that the dominant source languages in the translation process, English, French, Russian, Spanish, Classical Greek, and Latin, all have Low values in the BRING domain (all 0) and Low or Intermediate values in the RUN domains (0-1). Thus, the High MVC values found in the two domains in many languages cannot be due to mere peculiarities of the translation process, but represents structural features of the languages into which the text has been translated.

3. Advantages and disadvantages of using parallel texts

Let us now address the potential advantages and disadvantages of using parallel texts in typological studies in more general terms by considering the following criteria: (a) diversity, (b) domains, (c) analysis, (d) perspective, (e) quality, (f) representativity, and (g) comparability.

3.1. Diversity

Irrespective of the sampling procedure applied, it is clear that a typology is the better founded the higher is the degree of diversity of the languages considered. There is no doubt that the reference grammars available in a good linguistic library cover much more genealogic and areal diversity than what questionnaires studies and most parallel texts can cover, which is why reference grammars are the default choice for large-scale typological studies. The only parallel texts available in a

[6] This raises some doubts about the existence of a serial verb parameter, as suggested by STEWARD (2001) on the basis of material from few languages (mostly a single one, Edo). But the results presented here cannot be compared directly to those of STEWARD's study, since he focuses on domains other than those considered here and multi-verb constructions is a much broader term than verb serialization.

sufficiently large number of genealogically diverse languages from all continents are the gospels. There are, however, some areas where Bible translations are underrepresented (due to the fact that in some areas virtually all languages have become moribund before anybody started caring about the Bible). This is the case especially for the linguistically very diverse North American West Coast and for many languages of Australia. But even in Eurasia some isolates and small stocks, such as Burushaski, Ket, and Nivkh, are not represented. Another problem is availability. Even if some texts are easily accessible for some large languages (in published form or electronically on the internet), linguistic libraries usually do not have collections of Bible translations.

3.2. Domains

It depends very much on the domain to be investigated whether a certain parallel text is an appropriate data source. It is clear that the material must represent the domain of a typological research question. Whereas questionnaires can be specially designed to represent all situations relevant for the research question, typologists have no influence on the structure of parallel texts and so many domains are just lacking in available parallel texts. But neither are reference grammars good for all domains. Fortunately, the two sources of material tend to be complementary to a certain extent. Reference grammars are usually better for phonology, morphology and some aspects of syntax. Parallel texts, however, are very good for many lexical domains which are not well represented in grammars.

In databases based on reference grammars there are usually many gaps due to the fact that some relevant information is not found in the grammar (and be it only negative information, that a certain category is lacking). Parallel texts can help especially for research questions that have not been in the center of interest in linguistics and are therefore often not mentioned in grammars. For instance, the excellent grammar of Kuku Yalanji (Pama Nyungan) by PATZ (2002) does not mention co-compounds, the translation of Mark, however, shows that there are co-compounds (WÄLCHLI 2005: 238). A further advantage of using parallel texts is that it gives comparable quantitative data, and often it is even possible to study the context-dependence of certain semantic elements, especially emphatic vs. non-emphatic use (such as light and heavy ‘again’ discussed in WÄLCHLI 2006).

3.3. Analysis

Parallel texts are usually unanalyzed raw text. However, it is much easier to deal with a large number of translations of the same text than with different original texts, first of all, because the meaning of the text is known (except for some surprises due to problems caused by selectivity or underdetermination, cf. DE VRIES, this issue) and, second, because the known structure of the base text makes it possible to look selectively at a small number of passages in which structures relevant for the research question are most likely to occur. Analysis does thus not require segmenting and glossing of all morphemes of the whole text but rather identifying the relevant morphemes and constructions in selected places of the text. It is clear,

however, that an analysis requires additional data sources. Thus, parallel texts are in practice never the only source of information in a typological study. Additional sources, be it a specialist's knowledge about a language, dictionaries, or grammars, are indispensable and additional sources also allow for a first partial evaluation whether the structures present in the text are representative for the language under consideration.

Nevertheless, analysis is a sore point of the parallel text method, given that many languages have (a) non-Latinate writing systems, (b) several completely different orthographies, (c) complex morphonological processes, and (d) a bewildering wealth of affixes and/or function words. Analysis is costly even in the most easily accessible languages. One of the greatest advantages of the method, investigating domain-internal diversity, requires individual coding of each example in a database. If some steps of analysis can be automated, this may make analysis of parallel texts more appealing in the future (see CYSOUW *et al.*, this issue, and DAHL, this issue).

It cannot be denied that the risk of wrong analysis is considerable especially if small differences between morphemes are involved. Here are two examples where I made a wrong analysis in WÄLCHLI (2001: 301, 305). I confounded the Ossetic comitative *-imä* with the dative *-mä*, and I did not realize that Samoan has a verb *o'o* (written *oo*) 'arrive' different from *o* 'go/come.PL'. How big the risk of errors of analysis is can be known only if a substantial number of parallel text studies has been carried out and evaluated. However, the heuristic function of parallel texts is very important. Recurrently finding certain morphemes in a relevant domain calls for looking for them in dictionaries and grammars where they otherwise might have been overlooked.

3.4. Perspective

Linguistic structure is accessed in a different way by typologists depending on the material used. In comparison with grammatical descriptions, texts (with translations) have various advantages that can be subsumed under the heading of perspective, notably function-form orientation and avoidance of system-bias.

Parallel texts studies have a radical domain orientation. This is very useful for typology since typologists often understand the notion of domain as based on the concept of translational equivalence. While most grammars are organized according to formal categories (starting from form class, to particular expressions and then to function), parallel texts lead the investigator from particular textually embedded contexts to form.

Grammars generally tend to be biased (a) toward describing small structural units (morphemes rather than constructions), (b) toward describing systematically behaving structures, and (c) toward describing structures as systematic. Exceptions tend to be downplayed in grammars and simple systematic descriptions are preferred because they are shorter and easier to formulate. Texts lack this kind of system-bias. In texts it can be checked to what extent postulated systems and rules really apply. Especially important is that differences in language use can be studied in parallel texts (see DAHL 1985: 50 for a similar argument for questionnaires).

3.5. Quality

A translation can be wrong or strange in several respects and that can affect a typology based on it in several respects. As soon as frequencies are considered, it does not matter very much whether there are individual errors in few places in a text. More important is whether expressions occur with their natural frequencies throughout the text. It can be assumed that some structures generally will be better represented in translation, even in bad translation, than others, one factor being that some structures are less inert (or more easily convertible) than others in translation. For some structures it has been argued that they are incommensurable. For instance, LEVINSON (2003: 59) argues that frames of reference "are incommensurable (a representation in one framework is not freely convertible into a representation in another)". It is therefore interesting to check how translations into Australian languages (known for their absolute frame of reference in contrast to European languages with relative frame of reference) deal with this incommensurability. In the translation of Mark into Wik Mungkan there are in fact very few absolute location markers, much less than an average narrative text in a language of that region contains. However, (11) shows that the absolute frame occurs:

(11) Wik-Mungkan (Pama Nyungan) [Mark 4:35]
Ngamp *íiy-āmpa,* *kaaw*
PRO go-INFL east
'Let us pass over unto the other side.'

Even if there is incommensurability on the level of the sentence, this does not hold necessarily for the whole text. In example (11), the goal of motion is the east side of the Sea of Galilee, which is why the sentence can be converted into an absolute frame of reference. Thus, rather than discussing the abstract theoretical question whether or not translation is possible—of course, it is always possible with a certain loss due to selectivity and underdetermination, see DE VRIES (this issue)—we have to deal with the question how inert structures are in translations. A feature in the target language is inert if it is likely to be under- or overrepresented (in comparison with original texts) due to the different structures of the source language(s). Features expected to be inert are especially such which are incommensurable at lower level of textual organization and can be rendered correctly only if larger passages or the whole text are considered. For inertness it is of secondary importance whether or not a text is underdetermined and needs interpretation (and on what level there is underdetermination, clause, passage, or whole text). Rather what is relevant is whether a certain structure occurs with its natural frequency in the text as a whole (so that it is balanced in terms of expressivity and fore- vs. backgrounding). If this is not the case, a feature is distorted in the parallel text. It is clear that there will always be some amount of inertness and distortion in translation. Parallel texts are useless for a research question if there is complete distortion, but they can be used to a certain extent even if there is much distortion (as in the case of frames of reference). Moreover, assessing various degrees of distortion for different features is an important research topic in itself.

3.6. Representativity

While the lack of obligatory elements and ungrammatical structures makes a sentence undoubtedly wrong, in many cases there is a choice between using or not using certain elements in a construction. In the domain of motion events this holds for directional particles and affixes in some languages. Mansi (Uralic) is a language with directional prefixes which are not obligatory in many contexts. Examples (12) and (13) give two places from Mark where the old translation (a) from the 19th century has no prefixes but the recent translation (b) has prefixes. (The translations moreover differ in dialect, but this is not relevant here.) Prefixes (boldface) in Mansi are often redundant, but this is not the case in (13b) where the prefix has the particular meaning 'to shore' and not simply 'out'.

(12) Mansi [Mark 3:6]

a. *I* *kval-ïm* *farisej-t*
and rise-PTC:PST Pharisee-PL
'And the Pharisees went forth...'

b. *Farisej-t* ***kon***=*kwāl-s-ət*
Pharisee-PL out-rise-PST-3PL

(13) Mansi [Mark 5:2]

a. *Tau* *kerep-nïl* *sare* *kval-ïm-at* *jipalt...*
he boat-ABL immediately rise-PTC-LOC after
'And when he was come out of the ship...'

b. *Īsus* *xāp-nəl* ***pāγ***=*kwāl-m-ē-t...*
Jesus boat-ABL to:shore-rise-PTC-3SG-LOC

Judging from the occurrence of prefixes in Mansi original texts it seems that the use of prefixes in the recent translation is more representative of Mansi. In Livonian, another Uralic language, directional prefixes are borrowed from the Indo-European contact language Latvian and are completely redundant in most contexts. The translation of the gospels lacks them almost completely due to purism. What we are dealing with here is language-internal variation. Sometimes different registers in the same language have slightly different grammars and especially the frequency of means of expression varies across styles and registers.

Bible translations often create new registers or even new language varieties. Sometimes it is difficult to distinguish the religious variety from standardization since the two often go together. In many languages, "missionary" registers have high prestige and as a consequence an error can become correct first for this register and then for the whole language. Often grammars are based on the prestigious "standard" varieties, which is how "errors" of missionaries can end up in reference grammars. Consider, for instance, BRIGGS' (1993) discussion of "aymara mision-ero". In this "variety" there is a widespread use of TMA forms for direct evidence,

rather than using the colloquial hearsay evidential. An example is EBBING's (1965:83) use of the future instead of an evidential form in a sentence meaning 'The sinners will not enter into heaven' which has the connotation in non-missionary Aymara that the speaker commits himself to take care of making true what he says (BRIGGS 1993: 381). Since Bible translation played an important role in the formation of most modern European standard languages it is an interesting question as to what extent this may have affected their typology. Put differently, wrong translation is a problem for parallel text studies, but it is also a problem for typology in general.

Generally, Bible texts will often have a peripheral status in a typology of texts of particular languages (for the typology of texts see, e.g., BIBER 1995). Put differently, they will not be considered fully representative of a language. However, the problem of representativity is not only an issue for massive parallel texts like the Bible. Every typological classification is ultimately based on concrete examples (texts) and it is always the question to what extent these examples are representative for the language as a whole. Using parallel texts can make typologists more aware that typology is always a typology of texts and only indirectly a typology of languages. An advantage of the parallel text method is that it is more explicit about the concrete text passages considered.

3.7. Comparability

Direct comparability of concrete examples across languages is a strong point of the parallel text method. In the ideal case the same domains, instantiated in the same examples, are represented in the same textual environment with the same degree of emphasis in the same register. This means that, given that the analysis of all examples has been successfully completed, the values for the same features can be determined by applying the same criteria. Most of these advantages apply also to using questionnaires, except that in isolated sentences (as normally used in questionnaires) there is no textual environment which makes it more difficult to assess degrees of emphasis. However, typologists using parallel texts should be aware of the fact that there are no ideal exemplars.

As DE VRIES (this issue) points out, the gospels, the most usable texts in terms of diversity, are not completely parallel in several respects: (a) there is no unique base text, so different translations lack various passages (sometimes passages are given in brackets or footnotes) and (b) there is a wide variety of translational types ranging from highly literal and foreignizing to highly naturalizing and domesticating. These differences will have different effects for each feature to be investigated, so that there is no general answer how good the comparability is in a given set of parallel texts. One way of checking is to measure the variation across different translations representing different translation types in the few languages where more than one translation is available.

Comparability can also be improved by domain selection. Rather than comparing texts as a whole, only a restricted number of clauses is considered which are expected (a) to be represented in all texts and (b) to instantiate the construction or concept to be investigated. This procedure has been used in this paper for the com-

parative construction (Section 1) and for multi-verb constructions (Section 2). Holding the number of places considered constant is important especially when frequencies are compared. However, domain selection is not always possible. Some features with more idiosyncratic distribution due to lexicalization can be investigated only in complete text passages and the type of translation will have some effect on the frequency of occurrence (for co-compounds see WÄLCHLI 2005: 188).

While free translations are a problem inasmuch as it is more difficult to identify domains, literal translations are a problem inasmuch as they reflect at least partly the structure of the source language rather than the target language. This effect can be evaluated to a certain extent by comparing the values of potential source languages in the translation process. If the use of elements (and frequencies) in both source and target languages are strongly alike, this is more likely due to distortion than if there is some variation (philologists speak of *lectio difficilior*).

4. Conclusions

An important advantage of the parallel text method is that, exactly because of all its shortcomings, it requires a strong awareness of the problems involved in comparing languages. Typologists using parallel texts must be aware of a number of biases: (a) written-language bias (LINELL 1982), (b) bias toward planned (conscious) language use (including purism) (MILLER & WEINERT 1998), (c) bias toward religious and legalese registers, (d) narrative register bias, (e) bias toward large languages (in spread zones), (f) bias toward standardized (simplified?) language varieties, (g) bias toward non-native use of languages, (h) bias toward translated language (rather than original language use). However, many of these biases are involved in other sources such as reference grammars and dictionaries as well. There is an astonishing large number of grammars and dictionaries based, at least partly, on translated texts. Not rarely are authors of grammars and dictionaries also involved in Bible translation and it does certainly not hold in general that grammars or dictionaries written by Bible translators are worse in quality than others. It is no secret that much material used in typological studies is not perfect and that typologists are not always the ideal persons to analyze the structure of a particular language. However, the results we can get from typological studies using most different sources of material are so important for linguistics that it must be done even if it cannot be done in a perfect way.

Abbreviations

ABL ablative, ACC accusative, CONV converb, DAT dative, DEF definite article, GEN genitive, IMP imperative, INFL inflection, INTENS intensifier, IRR irrealis, LOC locative, NOM nominative, OBJ object, PL plural, POSS possessive affix, PRO pronoun, PROG progressive, PST past, PTC participle, REM.PST remote past, SG singular.

References

ANDERSEN, PAUL KENT (1983): *Word order typology and comparative constructions*. Amsterdam: Benjamins.

BIBER, DOUGLAS (1995): *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: Cambridge University Press.

BRIGGS, LUCY THERINA (1993): *El idioma aymara. Variantes regionales y sociales*. La Paz: Ediciones ILCA.

CYSOUW, MICHAEL (2002): Interpreting typological clusters, in: *Linguistic typology* 6: 69-93.

CYSOUW, MICHAEL, BIEMANN, CHRISTIAN & ONGYERTH, MATTHIAS (this issue): Using Strong's Numbers in the Bible to test an Automatic Alignment of Parallel texts.

DAHL, ÖSTEN (1985): *Tense and aspect systems*. Oxford: Blackwell.

DAHL, ÖSTEN (this issue): From questionnaires to parallel corpora in typology.

EBBING, JUAN ENRIQUE (1965): *Gramática y diccionario aymara*. La Paz: Don Bosco.

GIL, DAVID (1999): *Verb solitarization*. Paper given at the Mini-Symposium on Verb Serialization, Stockholm University, 18.3.1999.

HASPELMATH, MARTIN (1997): *From space to time. Temporal adverbials in the worlds languages*. München: Lincom.

LEVINSON, STEPHEN C. (2003): *Space in language and cognition. Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

LINELL, PER (1982): *The written language bias in linguistics*. Linköping: University of Linköping.

MILLER, JIM & WEINERT, REGINA (1998): *Spontaneous spoken language*. Oxford: Clarendon.

NEWMEYER, FREDERICK J. (1998): *Language form and language function*. Cambridge, Mass.: MIT Press.

PATZ, ELISABETH (2002): *A grammar of the Kuku Yalanji language of north Queensland*. (Pacific Linguistics 526.) Canberra: Australian National University.

STASSEN, LEON (1985): *Comparison and universal grammar*. Oxford: Blackwell.

STEWARD, OSAMUYMEN THOMPSON (2001): The serial verb construction parameter. New York: Garland.

VRIES, LOURENS DE (this issue): Some remarks on the use of Bible translations as parallel texts in linguistic research.

WÄLCHLI, BERNHARD (2001): A typology of displacement (with special reference to Latvian), in: *Sprachtypologie & Universalienforschung STUF* 54.3, 298-323.

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

WÄLCHLI, BERNHARD (2006): Typology of light and heavy 'again', or, the eternal return of the same, in: *Studies in Language* 30.1: 69-113

WÄLCHLI, BERNHARD & ZÚÑIGA, FERNANDO. (forthcoming): Source-Goal (in)difference and the typology of motion events in the clause, in *Sprachtypologie & Universalienforschung STUF*.

Correspondence Address

Bernhard Wälchli
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
waelchli@eva.mpg.de

FEDERICA DA MILANO (Pavia)

Demonstratives in parallel texts: a case study[1]

The aim of this paper is to show the usefulness of parallel texts for typological investigations. In order to analyze the way in which demonstrative systems of the European languages function, two kinds of data have been considered: first, the results of a questionnaire based on situations represented in 48 pictures, which will be necessarily discussed only in a summarized way here. Second, and this will be the main topic of this paper, a corpus of parallel texts: the translations, in different languages of Europe, of the book *Harry Potter and the Chamber of Secrets*. Parallel texts have been used to verify the generalizations based on the data elicited through the questionnaire.

1. Introduction

In descriptive grammars, terms like "proximal/distal" or "near/far" from the speaker are typically used to define the meaning of demonstratives. However, these definitions are only an approximation of a complex semantic domain. In particular, an important point concerns the distinction, as found in the literature, between so-called "distance-oriented" systems and "person-oriented" systems. The question is: is that a real distinction, or are they two instantiations of a more general system?

In order to answer this question, I have compiled a questionnaire for the elicitation of data. Because demonstratives seem to straddle the boundaries between visual perception, abstract semantic organization and pragmatic context, two parameters have been considered: distance (semantic parameter) and reciprocal orientation between speaker and hearer (pragmatic parameter). The questionnaire includes 48 pictures and is based on the notion of dyad of conversation (JUNGBLUTH 2001). This notion goes beyond the traditional distinction between "person-oriented" and "distance-oriented" systems because it is based on a detailed physical analysis of the orientation of speaker and addressee. The pictures in the questionnaire represent the three main communicative situations: face-to-face conversations, front-to-back conversations and side-by-side conversations.

In order to check the generalizations obtained by the elicited data, I have used a corpus of parallel texts, consisting of translations of *Harry Potter and the Chamber of Secrets* by J.K. Rowling in various European languages.[2] I have chosen this book because it is recent and it has been translated into many languages of the world. Because it is mainly a children's book, conversation is very natural and colloquial and includes a lot of dialogues. Dialogues are particularly interesting because they are a context in which demonstratives are frequently employed in their exophoric use (i.e. with external reference to real objects in space).

[1] I wish to thank MICHAEL CYSOUW and BERNHARD WÄLCHLI for their helpful comments on an earlier version of this paper.

[2] A parallel corpus based on the translations of the same *Harry Potter* book has been used for a typological study of epistemic possibility in the Slavonic languages (VAN DER AUWERA *et al.* 2005). Moreover, STOLZ (this issue) has used another book of the *Harry Potter* series, *Harry Potter and the Philosopher's stone*, to investigate possessive relations in the languages of Europe. The languages of the translations considered in this study are Basque, Catalan, Czech, Dutch, Finnish, French, German, Hungarian, Italian, Polish, Spanish.

2. Methodology

It has to be kept in mind that the use of translations in linguistic research is not unproblematic: the phenomenon of interference from the source language is well known (GELLERSTAM 1996). But this does not mean that translation must be ignored: if controlled, translational equivalents can be a very useful tool in linguistic research, as I will try to show in this paper.

A recent contrastive study of spatial demonstratives in English and Chinese (WU 2004) uses a similar methodology. One set of data was obtained from an experimental procedural task (jigsaw puzzle task). Another set of data came from a corpus formed by two pieces of narrative discourse (*Winnie-The-Pooh* and *Baohulu de Mimi*) with their Chinese and English translations respectively, considering that:

> "Parallel texts make it possible to observe how demonstrative reference in one language is signaled in the other within basically similar or identical propositions. As parallel texts put the discourse contextual factors largely in control, the behaviour of the demonstratives can be observed and compared in a focused manner." (WU 2004: 26)

The generalizations obtained from the analysis of the data elicited through the questionnaire will be discussed necessarily in a summarized way (for further details, see DA MILANO 2005). The attention will be devoted to the verification made possible through the use of parallel texts.

3. The systems of demonstrative pronouns

The topics of the analysis have been on the one hand demonstrative pronouns and, on the other hand, demonstrative adverbs, according to Diessel's definition of demonstratives:

> "[…] demonstratives are deictic expressions serving specific syntactic functions. Many studies confine the notion of demonstrative to deictic expressions such as English *this* and *that*, which are used either as independent pronouns or as modifiers of a coocurring noun, but the notion that I will use is broader. It subsumes not only demonstratives being used as pronouns or nouns modifiers but also locational adverbs such as English *here* and *there*." (DIESSEL 1999: 2)

As far as pronouns are concerned, the data from the questionnaire allowed classifying the languages into eight different types, four two-term demonstrative systems (summarized in 3.1-3.4) and four three-term demonstrative systems (summarized in 3.5-3.8).[3] As far as the parallel corpus is concerned, in the original text all the occurrences of *deictically used* demonstratives have been isolated and for each of the sentences thus isolated, the translational equivalents have been identified. Among these, the total set of sentences with either demonstrative determiners/pronouns or demonstrative adverbs amounted to 83. Looking only at the pronouns, the analysis of the parallel texts confirms the classification obtained through the questionnaire.

[3] Because of limitations of space, it is not possible to show the data obtained from the questionnaire.

3.1. Proximal vs. unmarked two-term systems

Two term systems exist in different variants. One possibility attested is that the two demonstratives show an opposition in locational proximity (i.e. proximal vs. distal), and the term for distal is the unmarked case.[4] From the questionnaire, it turned out that the following languages have such demonstrative systems: Norwegian (proximal *her,* unmarked *der)*, Danish (proximal *den,* unmarked *det*), Dutch (proximal *deze,* unmarked *die*), English (proximal *this,* unmarked *that*), and Northern Italian (proximal *questo,* unmarked *quello*). As shown (1) and (2), with examples from the parallel texts, English and Dutch exhibit a clear preference for the use of the distal term in situations unmarked for proximity.[5]

(1) a. *'Tie **that** round the bars,' said Fred, throwing the end of a rope to Harry.* [English 32]
b. *'Hier, knoop **dat** om de tralies', zei Fred, die Harry een touw toewierp.* [Dutch 23]

(2) a. *'Is **that** supposed to be music?' Ron whispered.* [English 144]
b. *'Moet **dat** muziek voorstellen?' fluisterde Ron*. [Dutch 100]

3.2. Distal vs. unmarked two-term systems

The reverse case was also attested in the questionnaire study. Some languages treat the proximal demonstrative as the unmarked case, in contrast to a marked distal. This was found in Polish (unmarked *ten/ta/to,* distal *tamten*), Russian (unmarked *ètot,* distal *tot*), Czech (unmarked *ten*, distal *tamten*),[6] Hungarian (unmarked *ez*, distal *az*), Bulgarian (unmarked *tazi,* distal *onazi*), and Modern Greek (unmarked *autós*, distal *ekeĩnos*).

In the examples (3)-(6) from the parallel texts, English uses the unmarked distal form. However, in Polish, Czech and Hungarian, the unmarked proximal term is used. Note that (4) and (6) show situations in which the object referred to is not near the speaker and English accordingly uses the (unmarked) distal demonstrative *that*. However, Polish, Czech and Hungarian use the (unmarked) proximal term, which is some evidence that the relation relative to the speaker is not of importance in these languages.

---

[4] The notion of markedness has been considered here as an asymmetric relation among different elements which is determined by various criteria as frequency, semantic generality and use in neutral contexts (GREENBERG 1966). This notion has been relevant in recent studies about demonstrative systems (DIXON 2003; ENFIELD 2003) and it is useful to make an interlinguistic comparison among demonstrative systems.

[5] Numbers behind the citations refer to the pages of the editions consulted.

[6] Note that *té* is a variant of *ten*, and *to* is the neuter form of *ten*. The usage of the suffix *-hle* is not of importance to the present investigation.

(3) a. *'Tie **that** round the bars,' said Fred, throwing the end of a rope to Harry*. [English 32]
b. *'Przywiąż **to** do kraty', powiedział Fred, rzucając Harry'emu koniec liny*. [Polish 32]
c. *To už mu Fred pohotově házel konec provazu a vyzval Harryho: 'Uvaž ho kolem **té** mříže!'* [Czech 27]
d. *'**Ezt** kösd rá a rácsra', szólt Fred, és egy kötelet dobott oda Harrynek.* [Hungarian 30]

(4) a. *'Can I have **that**?' interrupted Draco, pointing at the withered hand on its cushion.* [English 60]
b. *'Mogę **to** dostać?', przerwał im Draco, wskazując na wyschniętą rękęna poduszce.* [Polish 59]
c. *'Koupil bys mi **to**hle?' přerušil je Draco a ukazoval na vyschlou ruku na polštáři.* [Czech 49]
d. *'Vedd meg **ezt** nekem', szólt közbe Draco, és a párnán heverő aszott kézre mutatott.* [Hungarian 53]

(5) a. *'Is **that** supposed to be music?' Ron whispered.* [English 144]
b. *'Czy **to** ma być ich muzyka?' zapytał szeptem Ron.* [Polish 141]
c. *'**To** má být hudba?' šeptl Ron.* [Czech 114]
d. *'**Ezt** nevezik ők zenének?' suttogta Ron.* [Hungarian 126]

(6) *Dumbledore reached across to Professor McGonagall's desk, picked up the blood-stained silver sword and handed it to Harry. […]*
a. *'Only a true Gryffindor could have pulled **that** out of the Hat, Harry', said Dumbledore simply.* [English 358]
b. *'Tylko prawdziwy Gryfon mógł wyciagnąć **ten** miecz z tiary' rzekł profesor Dumbledore.* [Polish 347-348]
c. *'**Ten**hle meč mohl z klobouku vytáhnout jedině ten, kdo do Nebelvíru opravdu patří', řekl prostě Brumbál.* [Czech 280]
d. *'**Ezt** csak olyan ember húzhatta elő a süvegből, aki ízig-vérig griffendéles' szólt Dumbledore.* [Hungarian 309]

3.3. Dyad oriented two-term systems

Prototypically, dyad-oriented systems use the proximal term for referents in the area between speaker and hearer, and the distal term for referents outside this common area. This type is found in Catalan. In the following example from the parallel texts (7), Catalan uses the proximal demonstrative also to refer to an object, the crossbow, which is near the addressee.

(7) a. *'What's **that** for?' said Harry, pointing at the crossbow as they stepped inside.* [English 280]
b. *'¿I **això**?' – va preguntar el Harry, assenyalant la ballesta un cop van ser dins.* [Catalan 255]

3.4. One-term systems

Demonstrative systems of French and German show a tendency toward reduction. In grammars, French is described as having two demonstratives: *ceci* and *celà/ça*[7] and German is described as having a three-term systems: *dieser, der, jener*. But as the results obtained with the questionnaire have shown, and the parallel texts seem to confirm, French and German show a tendency to use only one term, *celà/ça* and *der/die/das*, respectively. In most examples, the two languages use only this demonstrative, as is illustrated here with examples (8) and (9).

(8) a. *'Tie **that** round the bars,' said Fred, throwing the end of a rope to Harry.* [English 32]
b. *'Attache **ça** aux barreaux', dit Fred qui lança à Harry l'extrémité d'une corde.* [French 30]
c. *'Schnür **das** um die Gitterstäbe', sagte Fred und warf Harry das Ende eines Seils zu.* [German 29]

(9) a. *'Can I have **that**?' interrupted Draco, pointing at the withered hand on its cushion*. [English 60]
b. *'Est-ce que je peux avoir **ça**?' coupa Drago, en montrant du doigt la main desséchée posée sur le coussin.* [French 58]
c. *'Kann ich **die** haben?', unterbrach Draco und deutete auf die verwitterte Hand auf dem Kissen*. [German 56]

3.5. Dual-anchored three term systems

In this type, there are three different demonstratives: proximal, medial and distal. Specifically, the medial term is used both to refer to something near the addressee and to something at a medium distance away from the speaker (irrespective of the location of the addressee). From the data from the questionnaire, this type was established for Spanish (proximal *este,* medial *ese*, distal *aquel*) and Basque (proximal *hau,* medial *hori,* distal *hura*). The following examples from the parallel texts show clear contexts in which the intended referent is near the addressee. These contexts are particularly useful to analyze the medial term in three-term systems.

[7] As ARRIVÉ *et al.* (1986: 211) say: "La forme *ça* n'a pas morphologiquement l'aspect d'une forme composée. Toutefois ses emplois sont ceux des formes composées. *Ça* est d'ailleurs historiquement issu de *cela,* peut-être sous l'influence de l'adverbe *çà*. Dans l'usage oral contemporain, *ça* tend à se substituer à *cela*, lui-même plus employé que *ceci*." moreover, PRICE (1971: 127) argues that "as a demonstrative, the simple pronoun *ce* has been almost entirely displaced by the compound form *ceci* (< *ce* + *ci*) and *cela* (< *ce* + *là*). (In speech, *cela* is usually reduced to *ça*, which is tending to go the way of *ce* and be weakened to 'it' […]."

English, which has a two-term system, always uses the distal/unmarked term, whereas Spanish and Basque use the medial term.

(10) *Harry, glancing over, saw Malfoy stoop and snatch up something. Leering, he showed it to Crabbe and Goyle, and Harry realised that he'd got Riddle's diary.*
   a. *'Give **that** back' said Harry quietly.* [English 258]
   b. *'¡Devuélveme **eso**!' – le dijo Harry en voz baja.* [Spanish 204]
   c. *'Itzuli **hori**!' – esan zion Harryk isilka.* [Basque 201]

(11) *Seconds after they had knocked, Hagrid flung it open. They found themselves face to face, with him aiming a crossbow at them. Fang the boarhound barking loudly behind him. […]*
   a. *'What's **that** for?' said Harry, pointing at the crossbow as they stepped inside.* [English 280]
   b. *'¿Para qué es **eso**?'– preguntó Harry, señalando la ballesta al entrar.* [Spanish 221]
   c. *'Zertarako da **hori**?' – galdetu zion Harryk, barrura sartu eta balezta seinalatuz.* [Basque 218]

3.6. Addressee-anchored three type systems

In this system with three demonstratives, the medial term is only used to refer to something near the addressee. In the questionnaire study, such demonstratives systems were found in Sardinian (proximal *custu,* medial *cussu,* distal *cuddu*), Tuscan (proximal *questo,* medial *codesto,* distal *quello*), and in Portuguese (proximal *esto,* medial *esso,* distal *aquel*). I have had no access to translations of Harry Potter in these languages to verify the results from the questionnaire.

3.7. Systems that shows a tendency toward reduction

In Serbo-Croatian, a special variant of a three-term demonstrative system has been attested. In Serbo-Croatian there are three demonstrative terms (proximal *òvāj,* medial *tâj,* distal *ònāj*) but only the proximal and the medial term are regularly used. This might point towards a development from a three-term to a two-term system. I have had no access to a translation of Harry Potter in this language to verify the results from the questionnaire.

3.8. Not prototypically dyad-oriented three term systems

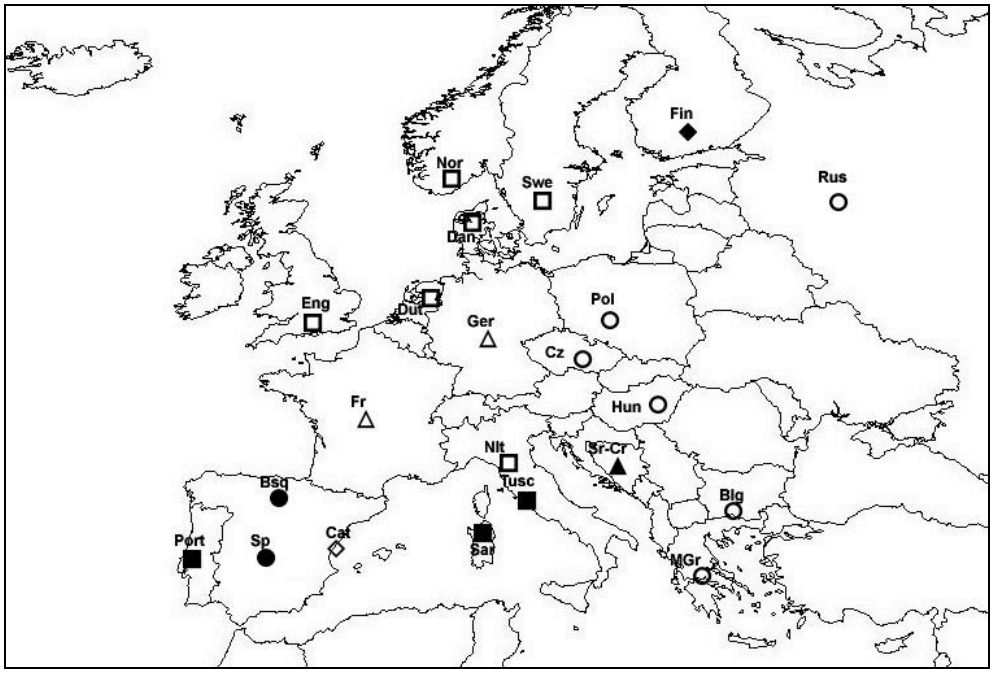
Finnish codifies a contrast between a space shared by the speaker and the hearer and a space outside of this area. *Tämä* is used for inside and *tuo* for the opposite meaning. *Se* refers to "something in the addressee's perceptual sphere" (LAURY 1996: 306). This behavior is typical of a dyad-oriented system, but we have to take into account the fact that Finnish is a non-article language; for this reason, the use

of demonstratives is comparable only to a certain degree and this is the explanation for the label 'not prototypically dyad-oriented' system.

(12) a. *'Tie **that** round the bars,' said Fred, throwing the end of a rope to Harry.* [English 32]
b. *'Sido **tämä** kaltereiden ympäri', Fred sanoi ja heitti köyden pään Harrylle.* [Finnish 33]

(13) *Harry, glancing over, saw Malfoy stoop and snatch up something. Leering, he showed it to Crabbe and Goyle, and Harry realised that he'd got Riddle's diary.*
a. *'Give **that** back' said Harry quietly. [English 258]*
*b. 'Anna **se** tänne', Harry sanoi hiljaa. [Finnish 258]*

Map 1. Areal distribution of the systems of demonstrative pronouns



□ Proximal/unmarked
○ Unmarked/distal
◇ Prototypically dyad-oriented
△ Toward one term
● Dual-anchored type
■ Addressee-anchored type
▲ Toward reduction
◆ Not prototypically dyad-oriented

3.9. Summary

Map 1 summarizes the types discussed in this section. In the sample of the questionnaire, two-way systems (14 cases) are more frequent than three-way distinctions (7 cases). The areal distribution of the two-term systems shows the existence of three areas. The first one, formed by the languages of northwestern Europe (Norwegian, Danish, English, Dutch, and Northern Italian), shows a contrast between a proximal term and an unmarked demonstrative. Second, French and German, which are considered the most prototypical Standard Average European (SAE) languages (VAN DER AUWERA 1998), show a tendency toward the reduction from a two-term system to a one-term system. Third, a further area is formed by the languages of middle-eastern Europe (Polish, Russian, Czech, Hungarian, Bulgarian, and Modern Greek). These are languages that have demonstrative systems that contrast an unmarked term with a distal one. Two of these languages (Bulgarian and Modern Greek) belong to the Balkan Sprachbund (BANFI 1985; 1991).

Three-term systems are less widespread than two-term systems and they are diffused in the Mediterranean area plus Finnish. Of these languages, Spanish and Basque show a dual-anchored type system. Finally, Tuscan and Sardinian have an addressee-anchored system.

4. The systems of demonstrative adverbs

As I have argued in the previous section, the systems of demonstrative pronouns can be classified in two basic types: two-term and three-term systems. The same distinction can also be found for demonstrative adverbs. I will first discuss the two term systems (Section 4.1), followed by the three term systems (Section 4.2). Finally, in Section 4.3, I will discuss the geographical distribution of these types in the languages of Europe.

4.1. Two-term systems

As far as two-term systems of demonstrative adverbs are concerned, various subcategories can be distinguished. First, there are two-term demonstrative systems (with a contrast between a proximal term and a distal one) in which the distal term is unmarked. This has been attested in Norwegian (proximal *her,* unmarked *der*), Danish (proximal *her,* unmarked *der*), English (proximal *here,* unmarked *there*), and Dutch (proximal *hier,* unmarked *daar*).

(14) a. *'HARRY! What d'yeh think yer doin' down* ***there****?'* [English 62]
b. *'HARRY! Wat mot dat* ***daar****?'* [Dutch 44]

Second, there are two-term demonstrative systems in which the proximal demonstrative is unmarked. This has been attested in Polish (unmarked *tu(taj),* distal *tam*), Russian (unmarked *tut,* distal *tam*), Czech (unmarked *tady,* distal *tamhle*), Hungarian (unmarked *itt,* distal *ott*), Bulgarian (unmarked *tuk,* distal *tam*), and Modern Greek (unmarked *edõ,* distal *ekeî*). In these cases, the parallel corpus

seems to confirm the generalizations obtained through the questionnaire. In contexts in which English uses the distal adverb *there*, Polish, Czech and Hungarian use the proximal term, as exemplified in (15)-(17).

(15) *He dreamed that he was on show in a zoo, with a card reading 'Underage Wizard' attached to his cage. People gogged through the bars at him as he lay, starving and weak, on a bad of straw. He saw Dobby's face in the crowd and shouted out, asking for help, but Dobby called,*
   a. *'Harry Potter is safe* ***there,*** *sir!' and vanished*. [English 29]
   b 'Harry Potter jest ***tutaj*** *bezpieczny, sir!", I zniknął.* [Polish 29]
   c. *'****Tady*** *je Harry Potter v bezpeči, pane!' a zmizel*. [Czech 25]
   d. *'Harry Potter* ***itt*** *bitzonságan van, uram!', azzal eltűnt.* [Hungarian 27]

(16) a. *'HARRY! What d'yeh think yer doin' down* ***there****?'* [English 62-63]
   b. *'HARRY! Cholibka, a co ty* ***tutaj*** *robisz?'* [Polish 61]
   c. *'HARRY! Prosím tě, co* ***tady*** *pohledáváš?'* [Czech 51]
   d. *'HARRY! Mi a cickafarkat keresel te* ***itt****?'* [Hungarian 55]

(17) a *'Wait* ***there****', he called to Ron*. [English 327]
   b. *Poczekaj* ***tutaj****!' zawołał do Rona.* [Polish 318-319]
   c. *'Počkej* ***tady****!' křykl na Rona.* [Czech 256]
   d. *'Várj meg* ***itt****' kiáltott át Ronnak.* [Hungarian 283]

Finally, a dyad-oriented system two-term system has been found in Catalan (proximal *aquí,* distal *allà*). In cases where English uses the distal demonstrative, Catalan uses the proximal, just like in Polish, Czech and Hungarian.

(18) a. *'Wait* ***there'****, he called to Ron. 'Wait with Lockhart. I'll go on.* [English 327]
   b. *'Espera't* ***aquí*** *– li va cridar al Ron. Espera m'amb el Decors. Jo continuo.* [Catalan 296]

4.2. Three-term systems

As far as three-term systems are concerned, various different subsystems can be distinguished. First, I distinguish so-called dual-anchor systems. For an explanation of their behavior, see Section 3.5. Dual-anchor systems allow us to improve the traditional and insufficient classification between 'person-oriented' systems and 'distance-oriented' systems. In dual-anchor systems the medial term is used not only referring to a place near the addressee (person-oriented), but also referring to a place at a middle distance away from the speaker (distance-oriented). This is attested in Spanish (proximal *aquí,* medial *ahí,* distal *allí*), Basque (proximal *hemen,* medial *hor,* distal *han*), and Serbo-Croatian (proximal *ovdje,* medial *tu,* distal *tamo*). Example (19) shows a context in which the speaker points very clearly to a space near the addressee. This is the beginning of a letter, implying that the demon-

strative adverb refers to the place where the addressee is. In these contexts, Spanish and Basque use the medial term.

(19) a. *Dear Ron, and Harry if you're* ***there****, ...* [English 53]
b. *Querido Ron, y Harry, si estás* ***ahí****, ...* [Spanish 45]
c. *Ron maitea, eta Harry ere bai,* ***hor*** *baldin badago: ....* [Basque 43]

Second, there are addressee-anchored type systems, as found in Sardinian (proximal *innoi,* medial *inguni,* distal *inguddeni*), and Tuscan (proximal *qui,* medial *costì,* distal *lì-là*). As shown in Section 3.6, in these systems the medial term is used exclusively referring to a space near the addressee (the tradtional 'person-oriented' system). I do not have any examples to verify the results from the questionnaire because I have had no access to any translations of Harry Potter in these languages.

Third, a not prototypically dyad-oriented system is attested in Finnish (proximal *täällä,* medial *siellä,* distal *tuolla*), see Section 3.8.

(20) a. *'HARRY! What d'yeh think yer doin' down* ***there****?'* [English 62]
b. *'HARRY? Mitä sinä* ***täällä*** *hortoot?'* [Finnish 63]

Fourth, German has a system with a contrast among proximal, medial and distal terms (proximal *hier,* medial *da,* distal *dort*). However, the examples (21)-(22) show the widespread use of the adverb *da*, indicating that *da* is becoming the default demonstrative adverb.

(21) a. *'Oh, Ron, there won't be anyone in* ***there'****, said Hermion.* [English 170]
b. *'Ach Ron,* ***da*** *wird niemand drin sein', sagte Hermine.* [German 162]

(22) a. *There was an ugly sort of wardrobe to his left, full of the teachers' cloaks. 'In* ***here****. Let's hear what it's all about.* [English 315]
b. *Zu seiner Rechten stand ein hässlicher Kleiderschrank voller Lehrerumhänge. '****Da*** *rein. Hören wir erst mal, was eigentlich los ist.* [German 301]

French and Portuguese are traditionally seen as having three-term systems. However, there is a clear tendency to reduce the three terms to two terms (French proximal *ici/là,* distal *là-bas* and Portuguese proximal *aqui/ali* distal *além*). From the data obtained with the French translation of Harry Potter, it is possible to observe the widespread use of the adverb *là*, progressively replacing *ici*. This is a tendency already recognized: "It should be noted also that usage of the proximal and distal demonstratives heavily favours the latter, particularly in speech" (HARRIS 1998: 221). Examples (23) and (24) show contexts in which the places referred to are clearly near the speaker. In these cases, English uses the proximal term *here*. However, French uses the (formerly) distal *là*.

(23) a. *'What're you doing **here**?'* [English 218]
b. *'Qu'est-ce que vous faites **là**?'* [French 215]

(24) a. *'I'm **here**!' came Ron's muffled voice from behind the rockfall.* [English 326]
b. *'Je suis **là**!' répondit la voix étouffée de Ron, derrière l'amas de rocs.* [French 319]

Finally, in Northern Italian a system of demonstrative adverbs is attested that shows a tendency to develop a contrast among three terms (proximal *qui/qua,* medial *lì,* distal *là*).

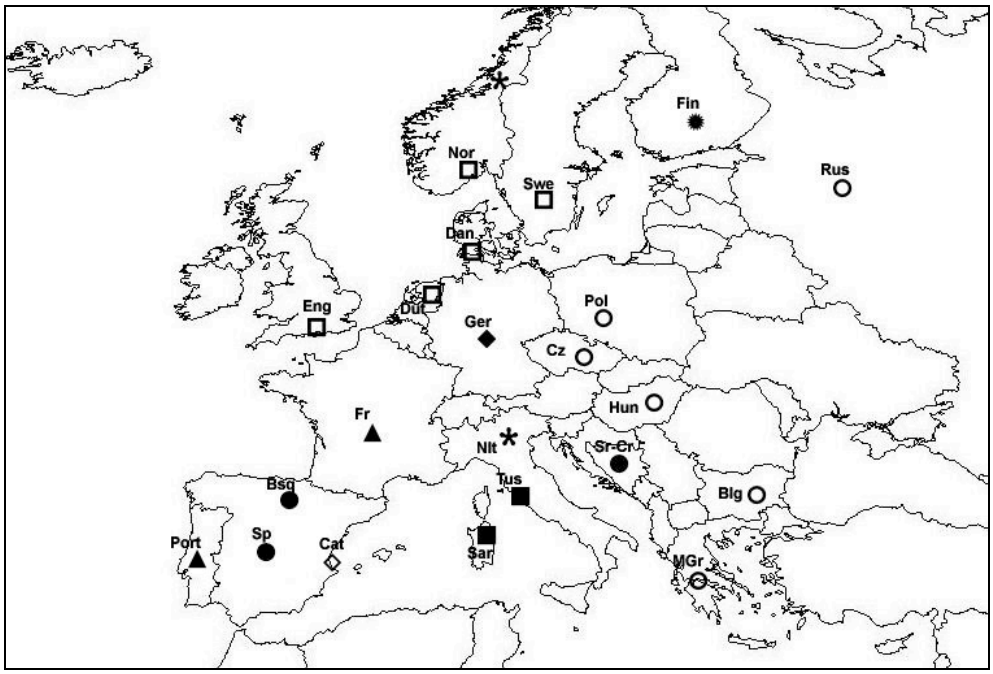
(25) a. *'It's over **there**, it got washed out'. Harry and Ron looked under the sink, where Myrtle was pointing. A small, thin book lay there.* [English 249]
b. *'Eccolo **lì**, si è bagnato tutto!' Harry e Ron guardarono sotto il lavandino, nella direzione indicata da Mirtilla. Per terra c'era un libricino.* [Italian 208]

(26) a. *'Ron – that girl who died. Aragog said she was found in a bathroom', said Harry, ignoring Neville's snuffling snores from the corner. 'What if she never left the bathroom? What if she's still **there**?'* [English 304]
b. *'Ron… la ragazza che è morta. Aragog ha detto che fu trovata in un gabinetto' disse Harry ignorando Neville che russava fragorosamente dall'altra parte della stanza. 'E se non fosse mai uscita dal gabinetto? E se fosse ancora **là**?'* [Italian 253-254]

4.3. Summary

In European languages two-term systems of demonstrative adverbs are widespread, as can be seen on Map 2. A comparison between Map 1 and Map 2 clearly shows the lack of isomorphism between the systems of demonstrative pronouns on the one hand (Map 1) and the systems of demonstrative adverbs on the other (Map 2). The systems of adverbs show a more complex articulation: this conforms to a general typological tendency: "perhaps one can hazard the generalizations that speaker-centered degrees of distance are usually (more) fully represented in the adverbs than the pronominals" (LEVINSON 2004: 43). Moreover, in Map 2 it is possible to individuate a northern area and an eastern area characterized by the prevalence of two-term systems, and a southern area with the majority of three-term systems.

Map 2. Areal distribution of the systems of demonstrative adverbs



☐ Proximal/unmarked
○ Unmarked/distal
◇ Prototypically dyad-oriented
● Dual-anchored type
■ Addressee-anchored type
✹ Not prototypically dyad-oriented
◆ Proximal/neutral/distal
▲ Toward two terms
✱ Toward three terms

5. Conclusions

In this paper, some of the translational equivalents of the English demonstrative pronouns and demonstrative adverbs have been investigated in the languages of Europe. It has to be kept in mind that I have investigated only some places of one text in one translation for each language, which may have led some idiosyncrasies. But, with these caveats, the research has shown that there are no very complex systems of demonstratives in the languages of Europe. Nevertheless, also systems that, at a first glance, seem to be relatively simple can vary in a rather subtle way in their conditions of use, making it difficult to make a typological classification.

It has been possible to identify three sub-groups within the languages considered (DA MILANO 2005). The first one includes approximately the languages of the so-called Charlemagne Sprachbund (VAN DER AUWERA 1998): French, German, (core), and Dutch, English, Danish, Norwegian, Northern Italian (periphery). The second subgroup includes the languages of central-eastern Europe: Russian, Czech, Polish, Hungarian, Bulgarian, and Modern Greek. The third subgroup includes

Mediterranean languages: Basque, Spanish, Portuguese, Tuscan, Sardinian, and Serbo-Croatian, but also Finnish.

The use of parallel texts, with the opportunity to check the contexts in which the demonstratives occur, has made it possible to verify nuances seemingly negligible (and in many descriptions, neglected) in the way in which demonstrative systems are structured. It has turned out to be fruitful to use parallel texts as a control test of data obtained through the questionnaire. The results from the parallel texts mainly confirmed the prior typological generalizations. I would agree with WU (2004: 203) that "[…] handled properly, the use of parallel corpora can produce fruitful results in a comparative/contrastive study".

References

ARRIVE, MICHEL, GADET, FRANÇOISE & GALMICHE, MICHEL (1986) : *La grammaire d'aujourd'hui : guide alphabétique de linguistique française*. Paris : Flammarion.

BANFI, EMANUELE (1985): *Linguistica balcanica*. Bologna: Zanichelli.

BANFI, EMANUELE (1991): *Storia linguistica del sud-est europeo*. Milano: Franco Angeli.

DA MILANO, FEDERICA (2005): *La deissi spaziale nelle lingue d'Europa*. Milano: Franco Angeli.

DIESSEL, HOLGER (1999): Demonstratives. Form, function and grammaticalization. Amsterdam: John Benjamins.

DIXON, RONALD M.W. (2003): Demonstratives. A cross-linguistic typology, in *Studies in Language* 27, 61-112.

ENFIELD, NICK J. (2003): Demonstratives in space and interaction. Data from Lao speakers and implications for semantic analysis, in *Language* 79, 82-117.

GELLERSTAM, MARTIN (1996): Translations as a source for cross-linguistic studies, in: AIJMER, KARIN, ALTENBERG, BENGT & JOHANSSON, MATS (eds.), *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994,* Lund Studies in English 88. Lund: Lund University Press, 53-62.

GREENBERG, JOSEPH (1966): *Universals of Language*. 2nd edition, Cambridge (Mass.): M.I.T. Press.

HARRIS, MICHAEL (1998): French, in HARRIS, MICHAEL & VINCENT, NIGEL (eds.), *The Romance Languages*. London & Sidney: Croom Helm, 209-245.

JUNGBLUTH, KONSTANZE (2001): Deictics in the Dyad of Conversation. Findings in the Romance Languages, habdout 23. Jahrestagung der DGfS Sprache und Kognition, Leipzig 28.02-02.03.

KENNY, DOROTHY (1998): Corpora in translation studies, in: BAKER, MONA (ed.), *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge, 50-53.

LAURY, RITVA (1996): Conversational Use and Basic Meaning of Finnish Demonstratives, in GOLDBERG, ADELE (ed.), *Conceptual structure, discourse and language*. Stanford: CSLI Publications, 303-319

LEVINSON, STEPHEN C. (2004): Deixis and Pragmatics, in HORN, LAURENCE R. & WARD, GREGORY (eds.), *The Handbook of pragmatics.* Oxford: Blackwell, 97-121.

PRICE, GLANVILLE (1971): *The French language: present and past*. London: Edward Arnold.

STOLZ, THOMAS (this issue): Harry Potter meets Le Petit Prince: On the usefulness of parallel literary corpora in crosslinguistic investigations.

VAN DER AUWERA, JOHAN (ed.) (1998): *Adverbial Constructions in the Languages of Europe*. Berlin-New York: Mouton de Gruyter.

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: Verlag Otto Sagner., 201-217.

WU, YI'AN. (2004): Spatial Demonstratives in English and Chinese. Amsterdam: John Benjamins.

Parellel texts

ROWLING, JOANNE K. (1998): *Harry Potter and the Chamber of Secrets*.London: Bloomsbury.
Basque: (2001): *Harry Potter eta sekretuen ganbera*.. Donostia: Ediciones Salamandra.
Catalan: (1999): *Harry Potter i la cambra secreta*. Barcelona: Editorial Empúries.
Czech: (2000): *Harry Potter a Tajemná komnata*. Praha: Albatros.
Dutch: (1999): *Harry Potter en de Geheime Kamer*. Amsterdam: Die Harmonie.
Finnish (*Harry Potter ja salaisuuksien kammio*. Helsinki: Tammi.
French: (1999) : *Harry Potter et la Chambre des secrets*. Paris: Gallimard Jeunesse.
German: (1999): *Harry Potter und die Kammer des Schreckens*. Hamburg: Carlsen Verlag Gmbh.
Hungarian: (2000): *Harry Potter és a titkok kamrája*. Budapest: Animus.
Italian: (1999): *Harry Potter e la Camera dei Segreti*. Milano: Salani Editore.
Polish: (2000): *Harry Potter i Komnata Tajemnic*. Poznán: Media Rodzina.
Spanish:(1999): *Harry Potter y la cámara secreta.* Barcelona: Ediciones Salamandra.

Correspondence address

Federica Da Milano
Dipartimento di Linguistica Teorica e Applicata
Corso Strada Nuova, 65
Pavia
chiccadm@tin.it

LOURENS DE VRIES (Amsterdam)

Some remarks on the use of Bible translations as parallel texts in linguistic research

The use of the Bible in parallel text corpora poses special challenges for researchers. The purpose of this paper is to describe the specific nature of Bible translations that sets them apart from other parallel texts such as translations of Harry Potter or the U.N. Universal Declaration of Human Rights. The special nature of translated Bibles is caused by textual multiplicity, canonical multiplicity and multiplicity of translation types. These three factors reflect one underlying cause, the specific skopos of Bibles: the religious functions of translated Bibles for a wide range of different Jewish and Christian communities.

1. Introduction

Texts that are translated into very many of the world's languages (like the Harry Potter books or the Bible) are an intriguing and important source of data for linguistic typology. Some of these texts, like documents from the United Nations or translations of the Bible, are publicly available (often in electronic format) and can with relative ease be transformed into digital corpora of parallel texts, without too many problems in the area of copyrights.

The use of the Bible in parallel text corpora poses special challenges for researchers. Bible translation is a process rooted in communities that create their own Bibles that conform to religious and hermeneutic notions of "Bible" valid in these communities. This creative and selective process determines to a large extent which Hebrew or Greek base text is chosen in the face of textual multiplicity, which books are included or rejected in the various canons of communities, which readings or interpretations are selected in the face of multiple readings and interpretations, which levels of style and lexis are acceptable, how much or how little interference from source language and text structures is allowed into the translation (foreignization versus naturalization), and so on. In this article, first, the notion of skopos, a central notion in translation studies, is introduced. Then I discuss the problems posed by the textual, canonical and translational multiplicity of Bible translations.

2. Skopos multiplicity

To understand the notion of *skopos* (pl. *skopoi*), it is essential to be aware of the nature of translating as an activity that always involves problems of selectivity and underdetermination. First, a single translation can never show all aspects of a source text. Translators have to decide on one specific wording, and in that process inevitably some aspects of the source are lost (selectivity). In the words of ORTEGA Y GASSET (1937; 2000: 62): 'It is, at least it almost always is, impossible to approximate all the dimensions of the original text at the same time.' Furthermore, although some translations are excluded as wrong by the source text, there remains much choice, since any text always can be translated in more than one way, with the source text legitimating these various ways of rendering the text. Source texts,

irrespectively of how brilliantly they are analysed, underdetermine their possible interpretations and translations. BECKER (1995: 370) refers to these problems of selectivity and underdetermination with the terms "deficiency" and "exuberancy", respectively.

Translators solve problems of selectivity and underdetermination by invoking criteria outside of the source text. This is their only option, whether or not they are aware of it. These external criteria emerge from a complex and heterogeneous set of factors collectively referred to in empirical translation studies as the "skopos" of the translation. The term skopos, the Greek word for "purpose", was introduced to translation studies by VERMEER (2000: 1) who analysed translation as an action, and grounded the idea of skopos in the intrinsically purposive nature of all human action. For NORD (1991: 28), another prominent spokesperson of the German skopos school, "translation is the production of a functional target text maintaining a relationship with a given source text that is specified according to the intended or demanded function of the target text (translation skopos)." I will use the notion of skopos to analyse the unavoidable process during to solution of the two problems faced by all translators, namely selectivity and underdetermination.

One can speak of function or skopos in relation to commissioners and translators who have certain skopoi or functional goals for the translation (intended translation function). For example, a missionary may want to translate the Bible to plant a church in a community. However, in the course of time translations may acquire different functions in target communities since once born they have a functional life of their own (acquired functions). For example, some so called "common language" versions of the Bible were meant for external functions, to bring the message of Scriptures close to modern audiences outside the churches, not as liturgical and ecclesiastical Bibles. But many church members of churches that use older, more literal versions in the liturgy, use the common language versions for private or family reading. In some church communities common language versions are used in church services also.

Further, communities may have expectations of translations, they expect to be able to do certain things with the text (expected functions). This is a crucial factor in Bible translations as the various Christian communities such as Catholics, Pentecostals or Orthodox have different theologies of Scripture, essentially different notions of "Bible". Sufficient overlap between the intended skopos (or function) and the expected function is crucial for acceptance of any new version of the Bible in a community. For some communities the translation must reflect the transcendent otherness of God and the translation function mainly in the liturgy where the text is celebrated and the public reading is a sacred ritual; communication of messages is not the aim. Other communities see the Bible as messages of God for humanity, messages that should be communicated as clearly as possible. For example, consider a simple Greek clause like Mark 1:37, as shown in (1). The Dutch *Nieuwe Vertaling* translates this clause as shown in (2).

(1) Classical Greek (Mark 1:37)
Πάντες ζητοῦσίν σε
Pantes zētusin se
all:PL seek:PRS.3PL thou:ACC

(2) Dutch (Mark 1:37, *Nieuwe Vertaling* 1952)
Allen zoeken u
all seek:PRS.3PL thou

While (2) shows one aspect of the source well, namely the syntax of the Greek clause, it does not include the durative aspect that a Greek verb in the present tense expresses. When translators decide to translate the durative aspect, there are various possibilities in Dutch, all equally supported by the source text. For example, the Dutch *Groot Nieuws Bijbel* has (3) with the durative auxiliary *lopen* 'to walk'. The *Nieuwe Bijbel Vertaling* has another, progressive-like construction *op zoek zijn* 'to be seeking', as shown in (4).

(3) Dutch (Mark 1:37, *Groot Nieuws Bijbel,* 1988)
Iedereen loopt u te zoeken
everyone walk:PRS.3SG thou PART seek:INF

(4) Dutch (Mark 1:37, *Nieuwe Bijbel Vertaling,* 2004)
Iedereen is naar u op zoek
everyone be:PRS.3SG to thou PART seek

The versions that reflect the durative aspect cannot at the same time reflect the syntax of the Greek clause. Conveying both the durative aspect and the syntax of the Greek source in one Dutch clause is simply impossible. Translators have to decide which aspect of the source should get priority in the translation (selectivity). At the same time this example shows the problem of underdetermination: the Greek source text legitimates multiple Dutch translations.

Now given the selectivity and underdetermination of translations, how do translators decide whether to translate (1) as (2), (3), or (4)? Considerations about equivalence cannot help since all these translations can claim to be equivalent to some aspects of the source text and none is excluded by the source text. The solution to take skopos considerations into account. The differences between the various Dutch translations follows from their skopos. For example, the Dutch *Groot Nieuws Bijbel* has a common language skopos. It is a translation primarily made for people outside the churches (external function). Accordingly, its translation of Mark 1:37, as shown in (3), conveys what this sentence means in common Dutch, but does not show the form of the Greek syntax. In contrast, the *Nieuwe Vertaling* has a church-internal skopos. It was made to function in church communities with inspiration theologies that want to maintain the inspired nature of the (literal) Word of God in the source. This leads to the translation as shown in (2), which approaches the form of the Holy Scriptures and is also regular Dutch.

Bible translations are different from other translated texts, both in terms of quantity and of quality, because of specific religious functions that the Bible has in the various communities. In terms of quantity, there are very often many translations of the Bible in one language that reflect different skopoi (cf. the numerous English translations of the Bible). No other book is translated in so many ways into the same language. Qualitative differences between Bible translations and translations of other texts derive from the religious functions of the Bible. For example, Bible translations exist in extreme translational types, both extremely foreignizing (high source language interference, Holy Inspiration skopos) and extremely domesticating types (missionary skopos). In between these extremes there are many intermediate translational types reflecting specific religious and secular functions.

The notion of the skopos (or goal) of a Bible translation is often associated with specific functions or with special audiences that Bible translations may have, like study Bible translations, common language translations, liturgical translations, Bibles for children, and so on. Although such specific functional elements belong to the skopos of Bible translations, the core of the skopos of Bible translations is formed by theological and hermeneutic elements that define the notion "Bible" for a given community and that emerge from the specific spirituality of that community. Such complex and sometimes partly implicit notions of "Bible" define the target or goal of every new translation of the Bible. The various Jewish and Christian communities have created their own Bibles in the course of their histories of translation. These creative translation histories involve the selection of textual traditions, of books to be included in the Bible, views on the relationship between the human authors and the Divine Author of the Bible, and different answers to the crucial question of the hermeneutical division of labour between the tradition/Church, the individual believer and the Bible translation. Such basic assumptions about the Bible determine how the Bible functions in the various communities and form the framework to further define notions as "study Bible" or "Church Bible". All these skopos-related factors make the Bible a very different and rather tricky type of parallel texts for linguists to work with as a source of data about the languages of the world.

3. Textual multiplicity

The Bible is a collection of Hebrew, Aramaic and Greek texts from Antiquity, and just like other texts from Antiquity, has a complex history of textual transmission. There is not such a thing as "the source text" of the Bible that forms the basis for all translation: both the Hebrew and the Greek Bible are characterized by textual multiplicity. When studying Hebrew and Greek Bible manuscripts, scholars group these manuscripts into multiple textual traditions. The various religious communities have accepted different textual traditions as the authoritative form of the text in the course of their histories.

As far as the Hebrew Bible is concerned, the Qumran findings have given new insights in the rich textual variety of the biblical text in the Second Temple period and they can be grouped into five groups of texts, including proto-Masoretic texts, pre-Samaritan texts and texts close to the reconstructed Hebrew source of the *Sep-*

tuagint (TOV 1992:117). Notice that the adoption of one tradition of texts as base text by religious communities does not solve the problem of multiplicity since all these textual traditions have a lot of internal variation. For example, the group of texts known as the Masoretic texts of the Hebrew Bible defeated, so to speak, other textual traditions and became the authoritative group of texts for Jewish communities. But since the texts of this group have considerable internal variation, printed editions of the Hebrew Bible based on different Masoretic manuscripts (or combinations of Masoretic manuscripts) differ. And this is reflected in translations. TOV (1992:2) gives the example of Genesis 49:10 where the *King James Version* has "until Shiloh come" but other English versions (*New English Bible, New Revised Standard Version*) have "so long as tribute is brought to him."

For Jewish communities the Masoretic texts as selected in the Rabbinic Bibles became very authoritative, especially the second Rabbinic Bible. The first two Rabbinic Bibles were printed in Venice by Daniel Bomberg in the first half of the 16th century (TOV 1992:78). However, no single source has been found from which the editors of the first two Rabbinic Bibles could have derived their biblical text (TOV 1992:78) and scholars believe the editors used various manuscripts. Modern scholarly editions of the Hebrew Bible are based on single sources such as the *Leningrad Codex* (*Biblia Hebraica Stuttgartiensia*) or the *Aleppo Codex* (*Hebrew University Bible*) complemented by a critical apparatus that contains variants from other manuscripts from the Masoretic text tradition and conjectural emendations. Printed editions of the Hebrew Bible differ not only in terms of the Hebrew base text but also in terms of chapter and verse division, in the sequence of the books of the Hebrew Bible and in the layout of the text (TOV 1992:3-8).

The Greek New Testament has a similar complex history of textual transmission and multiplicity of texts and textual traditions. In the early period of the Christian Church local traditions of textual transmissions developed around major urban centres of Christianity such as Alexandria, Antioch, Constantinople, Carthage and Rome. Scholars commonly discern Alexandrian, Western, Caeasarean and Byzantine text types. To complicate matters we find sometimes mixing of traditions in the manuscript evidence (METZGER 1971). Just like the Masoretic tradition was the historical winner in the case of the Hebrew Bible and ended up in the first printed Hebrew Bibles, the Byzantine text tradition became, after the sixth or seventh century, the authoritative form of the text of the New Testament until the rise of textual criticism in the 19th and 20th century. Modern textual criticism tends to favour the Alexandrian text type found in the famous codices *Vaticanus* and *Sinaiticus*.

The first published printed edition of the Greek New Testament was prepared by ERASMUS OF ROTTERDAM in 1516 and contained a rather corrupt form of the Byzantine text because ERASMUS had only late and inferior manuscripts at his disposal. For Revelation his only manuscript lacked the last 6 verses of the book and ERASMUS then translated these verses into Greek from JEROME's *Vulgate*. Also in other parts of his Greek texts he introduced Greek elements on the basis of the *Vulgate*. ERASMUS' edition soon became very much in demand and formed the basis for both LUTHER's German translation of the New Testament (1522) and TYNDALE's English translation of the New Testament (1525). METZGER (1971: xxiii) concludes: "It was the corrupt Byzantine form of text that provided the basis for al-

most all translations of the New Testament down to the nineteenth century." Nowadays most translations of the New Testament translate from a very different Greek text, namely an eclectic text that heavily leans on the Alexandrian textual tradition but that also includes variants from other traditions based on the application of principles from the field of textual criticism (ALAND & ALAND 1982).

To establish some continuity with past translations and with the translation tradition of the community, well-known verses that are now regarded as less acceptable because of text-critical considerations are often included in modern translations but with some indication of their doubted status. Sometimes the unacceptable verse is placed in a footnote with its verse number and in the text the continuity of verse numbers is broken (see for example the *Good New Bible*). Other translations put the less acceptable verse between square brackets. In this way, the less acceptable verse retains its verse number, creating continuity with older translations (e.g. the Dutch *Nieuwe Vertaling,* 1952). A third solution is that the verse number is placed in the text but the verse itself is deleted giving a blank line as in some French versions. Finally, the verse number may be mentioned with the previous verse but the unacceptable verse is in a footnote.

These textual differences are not trivial. For example, the Lord's Prayer in Matthew 6:13 has a longer ending in the *King James Version* "for thine is the kingdom and the power and the glory for ever. Amen." This longer ending will not be found in most modern English translations. Since translations of the Bible differ considerably depending on the Hebrew and Greek texts selected as base for the translation, their status as parallel texts is more complicated than translations of Harry Potter, where there is one undisputed English base text.

4. Canonical multiplicity

The various religious communities have to come to accept in the course of their histories a wide variety of canons, or lists of holy books considered inspired and authoritative; there are also degrees of canonicity (canonical, deutero-canonical, apocryphal) and various communities have both narrower and wider canons. Traditional sequences of books in the Bible also differ from community to community.

The Ethiopic Orthodox Church has all the books found in the *Septuagint*, including 3 Ezra, 3 Maccabees and Psalm 151, but on top of that the Prayer of Manasseh, 4 Ezra, Jubilees and Enoch. The latter two do not appear elsewhere in the *Vulgate* or *Septuagint* traditions (RÜGER 1991: 155). Bibles in Amharic, therefore, have the most books of all Bible translations.

The Syrian Orthodox Church with its ancient Peshitta translation is also interesting because it is the only community with the Letter of Baruch in its Old Testament canon and also because the Peshitta omits 2 Peter, 2-3 John, Jude and Revelation in the New Testament (RÜGER 1991: 156).

The Roman Catholic Church fixed its canon during the Council of Trent in 1546 favouring the *Vulgate*, the Latin translation that had become the authoritative base text for this community. For the Old Testament the canon included the Pentateuch, Joshua, Judges, Ruth, four books of Kings, two books of Chronicles, 1-2 Ezra, Tobit, Judith, Esther, Job, a Psalter with 150 psalms, Proverbs, Ecclesiastes, Song

of Songs, Wisdom, Sirach, Isaiah, Jeremiah (including Baruch), Ezekiel, Daniel, the twelve minor prophets and 1-2 Maccabees.

Modern Protestant Bibles have the shortest list of books included in the translation because they tend to omit the books that were declared Apocryphal by the Reformers (RÜGER 1991:152). The *Confessio Belgica* of 1561 list the following books as Apocryphal, descibed by LUTHER as "books not of equal value with Holy Scripture, yet useful and good to read": 3-4 Ezra, Tobit, Judith, Wisdom, Jesus Sirach, Baruch with the Letter of Jeremiah, additions to Esther, the Song of the Three Men in the Fiery Furnace, Susanah, Bel and the Dragon, the Prayer of Manasseh, the two books of the Maccabees. Older Protestant translations such as LUTHER's translation, the Dutch *Statenvertaling* of 1637 and the *King James Version* do contain the Apocrypha but all with slight variations of books included (RÜGER 1991: 153).

5. Multiplicity of translation types

The religious function of the Bible has important hermeneutic and translational implications that sets Bible translations apart. Whereas the hermeneutic position of the reader of translations of other books from Antiquity, such as the works of Herodote or Homer, is often assumed to be that of someone overhearing a conversation or reading a letter that was not intended for the modern reader, religious communities view the Bible as God's Word addressed to the (community of) readers of the translation. God is the Divine Author of the Bible and the community of believers is the addressee. In the course of time certain communities of believers have stressed the first part of this assumption, namely that it is God that speaks in the Bible, and that therefore the translation should be as literal and foreignising as possible: it is the voice of the Divine Other that should be discerned in the translation. For example, Bible translations that bring the text to the modern readers, by naturalising and domesticating the text, are totally unacceptable for Russian-Orthodox and Greek-Orthodox communities. They want Bible translations reflecting the Otherness of the Divine Author.

Other communities, for example American evangelical communities with a strong missionary drive, likewise subscribe to the assumption that God is the Divine Author of the Bible and the community of believers is the direct addressee but they emphasize the hermeneutic status of the new readers and listeners of Bible translations as the intended addressee of the Bible. Since God spoke in the Bible in order to be understood, readers of translations should be able to understand the Bible as if God had spoken to them in their own languages. This leads to a translation type called communicative translations that are extremely explicative and naturalising.

Quite often communities uses multiple types of translations for multiple (religious) functions, for example rather special philological translations to use as Study Bible, traditional literal translations for liturgical functions (e.g. *King James*), and yet other translation types for external functions (e.g. the "loose" *Good New Bible* for evangelistic campaigns). Because of these various religious functions Bible translations can be extremely free or extremely literal, in some cases down to the

level of morphemes or function words. The classical example here is AQUILA's revision (around 125 CE) of the *Septuagint*, the Greek translation of the Hebrew Bible. AQUILA's notion of 'Bible', derived from his teacher AKIBA, "determined that every letter and word in the Bible is meaningful. Aquila therefore made an attempt to represent accurately every word, particle, and even morpheme in his translation. For example, he translated the every Hebrew *nota accusativi* אֶת separately with συν 'with', apparently on the basis of the other meaning of אֶת, namely 'with' (TOV 1992:146).

Whenever translators worked for communities that saw the Bible as inspired on a word-by-word basis, this Holy Inspiration skopos leads to translations that try to preserve the order and categories of the words as found in the source texts. The monumental Dutch *Statenvertaling* (1637) is an example of a Bible with a Calvinistic Holy Inspiration skopos. Another aspect of this type of translation is the tendency to use the same translation equivalent for each occurrence of a given source word, so called "lexical concordance", irrespective of the lexical patterns and collocations of the target language.

Many Bible translations for minority languages that were made after the Second World War by missionaries and organizations, like Wycliffe Bible Translators and the United Bible Societies, have a missionary skopos (KRONEMAN 2004). They were meant as stand-alone texts. They do not assume pastors, priests or elders to explain the text and the goal is to bring the message of salvation as close as possible to the readers or listeners. This leads to translation of the explicative type. Consider the following example of an SIL translation from Indonesian Papua, with a message-oriented, missionary skopos, the *New Una Version* in its translation of Mark 1:2a-3, first given in Greek (5) and in English (6) in the rather literal *Revised Standard Version* (1952) followed by the *New Una Version* (2004) in (7), with an English backtranslation (8) by KRONEMAN (2004:383):

(5) Greek (Mark 1:2a-3, following the edition of ALAND et al. 1975)
᾿Ιδοὺ ἀποστέλλω τὸν ἄγγελόν μου πρὸ προσώπου σου,
ὃς κατασκευάσει τὴν ὁδόν σου·
φωνὴ βοῶντος ἐν τῇ ἐρήμῳ·
῾Ετοιμάσατε τὴν ὁδὸν κυρίου,
εὐθείας ποιεῖτε τὰς τρίβους αὐτοῦ,

(6) English (Mark 1:2a-3, *Revised Standard Version*, 1952)
Behold, I send my messenger before thy face,
who shall prepare thy way;
the voice of one crying in the wilderness:
Prepare the way of the Lord,
make his paths straight.

(7) Una (Mark 1:2a-3, *New Una Version*, 2004)
Kekebnurum. Nira Imtamnyi biryi ninyi tentok ara ni uram erbinkwandanyi bisi bokdonokwan. Anyi bira kanda ninyi Lembinkwandemnyi bisi menekdiryok, bisik lilibkwankir. –"Ni uram erbinkwandanyi bira ninyi kun kum ai aryi kubdiryok, uram dobkwandi. Erci uram weik doboka ato ebkwandi, "Er Iya Mikibnyi yankwansir ati, sunci sundamnyi kiknibminikdamunci, bisik yabdarur. Er iya Mikibnyi yankwansir bisik asi udikum yabmun cok, ersi kibdobdarur." Ato eboka er Imtamnyi uram erbinkwandanyi biryi uram dobkwandi.

(8) English (Mark 1:2a-3, literal backtranslation from the *New Una Version*, KRONEMAN 2004:383)
Listen. I the heavenly One will send a person who will go in order to tell my words. As for this person, he will go before you who are the one who will rescue people, and he will pave the way for you. – As for the person who will go in order to tell my words, being in the place where people usually don't live, he will shout. Shouting, he will say like this, "The Most Powerful One will come to you, and therefore you must prepare yourselves, and pave the way. You must make straight the way that the most powerful One will come, and welcome him." Saying like this, the person who will go in order to tell the words of the heavenly One will shout.

KRONEMAN (2004:383) mentions some of the explicative elements in the literal English backtranslation of the Una version. With respect to the Greek source there is, for example, participant explicitation (shown here in boldface): "I, **the heavenly One** … and … you **who are the one who will rescue people**." There is also explicitation of a cultural assumption of the source. The element of "welcoming" has been made explicit, since it seems to be central to the idea of preparing the road for the king: "You must make straight the way that the most powerful One will come, and **welcome** him."

To present a further indication of the wide variety of translation types, consider the following translations of Romans 1:16-17 and note how the Greek phrase δικαιοσύνη θεοῦ "righteousness of God" has been translated (italicized in the examples):

(9) Greek (Romans 1:16-17)
Οὐ γὰρ ἐπαισχύνομαι τὸ εὐαγγέλιον, δύναμις γὰρ θεοῦ ἐστιν εἰς σωτηρίαν παντὶ τῷ πιστεύοντι, Ἰουδαίῳ τε πρῶτον καὶ Ἕλληνι. δικαιοσύνη γὰρ θεοῦ ἐν αὐτῷ ἀποκαλύπτεται ἐκ πίστεως εἰς πίστιν, καθὼς γέγραπται, Ὁ δὲ δίκαιος ἐκ πίστεως ζήσεται.

(10) English (Romans 1:16-17, *Revised Standard Version*)
For I am not ashamed of the gospel: it is the power of God for salvation to every one who has faith, to the Jew first and also to the Greek. For in it *the righteousness of God* is revealed through faith for faith; as it is written, "He who through faith is righteous shall live."

(11) English (Romans 1:16-17, *Common English Version*)
I am proud of the good news! It is God's powerful way of saving all people who have faith, whether they are Jews or Gentiles. The good news tells *how God accepts everyone* who has faith, but only those who have faith.[*] It is just as the Scriptures say, "The people God accepts because of their faith will live".

(12) English (Romans 1:16-17, *Good News Bible*)
I have complete confidence in the gospel; it is God's power to save all who believe, first the Jews and also the Gentiles. For the gospel reveals *how God puts people right with himself*: it is through faith from beginning to end. As the scripture says, "The person who is put right with God through faith shall live."

(13) English (Romans 1:16-17, *New International Version*)
I am not ashamed of the gospel, because it is the power of God for the salvation of everyone who believes: first for the Jew, then for the Gentile. For in the gospel *a righteousness from God* is revealed, a righteousness that is by faith from first to last,[c] just as it is written: "The righteous will live by faith."

6. Conclusion

Bible translation like all other translation is a skopos-guided activity but the religious nature of the skopos of Bible translation sets Bibles apart from other types of texts. Both the Hebrew and Greek Bible have a complex history of textual transmission and communities have accepted certain forms of the text as authoritative and rejected others. Some Jewish and Christian communities have accepted the results of the academic field of textual criticism and others have not. Communities that accepted these results also accepted that later translations and revisions put certain verses between brackets or omitted them altogether. Therefore, Bible translations are based on different source texts, and comparing Bible translations is very tricky if you do not know the Hebrew or Greek base texts used. When the Bible translation has no preface or introduction with information on the biblical base texts used, linguists will have to consult specialists in the field of Bible translation for information on Hebrew and Greek base texts that were used.

Another source of complications for the linguist is that different Bibles have different sets of books in them because different communities have different notions of "Bible" (canonical multiplicity), and sometimes combine books that are separate in other translations. Order and titles of books may also differ.

The final source of complications is the wide variety of translational types based on the various religious functions of the Bible: communities do very different things with the Bible and translators produce translations that serve these needs. From translations with a high degree of interference from source languages and source texts that contain a kind of "translationese" to communicative translations that present the Bible as if it was a product of the target culture, adding very many elements to clarify the text for modern readers.

The conclusion is that linguists can use Bibles for linguistic research but only if they are willing to consult specialists in the field of Bible translation to learn about the skopos of these translations and its consequences for base text, canon and translational type.

Abbreviationa

ACC accusative, INF infinitive, PL plural, PRS present, PART particle.

References

ALAND, KURT & BLACK, MATTHEW & MARTINI CARLO & METZGER, BRUCE & WIKGREN, ALLEDN (1975): *The Greek New Testament*. New York: United Bible Societies.

ALAND, KURT & ALAND, BARBARA (1982): *Der Text des Neuen Testaments*. Stuttgart: Deutsche Bibelgesellschaft.

BECKER, ALTON L. (1995). *Beyond Translation: Essays towards a modern Philology*. Ann Arbor, Mich.: University of Michigan Press.

KRONEMAN, DICK (2004): *The LORD is my shepherd. An exploration into the theory and practice of translating biblical meaphor*. Doctoral Dissertation Vrije Universiteit, Amsterdam.

METZGER, BRUCE M. (1971): *A textual commentary on the Greek New Testament*, New York: United Bible Societies.

NORD, C. (1991): *Text analysis in translation: theory, methodology, and didactic application of a model for translation-oriented text analysis*. Amsterdam: Rodopi.

ORTEGA Y GASSET, J. (1937). 'La Miseria y el esplendor de la traducción'. In : *Obras Completas: Tomo V*, Madrid: Revista de Occidente, 427-448.

ORTEGA Y GASSET, J. (2000), 'The misery and splendor of translation', in: L. Venuti (ed), *The translation studies reader*, London: Routledge, 49-64. [Translated by Esther Allen.]

RÜGER, HANS PETER (1991): The extent of the Old Testament canon, in: MEURER, SIEGFRIED (ed.), *The Apocrypha in Ecumenical Perspective*. New York: United Bible Societies.

TOV, EMANUEL (1992): *Textual criticism of the Hebrew Bible*, Minneapolis: Fortress Press.

VERMEER, HANS (2000): Skopos and commission in translational action, in: VENUTI, L. (ed.), *The translation studies reader*. London: Routledge, 221-232.

Correspondence address

Lourens de Vries
Vrije Universiteit Amsterdam,
De Boelelaan 1105
1081 HV Amsterdam
lj.de.vries@let.vu.nl

MICHAEL CYSOUW (Leipzig)
CHRISTIAN BIEMANN (Leipzig)
MATTHIAS ONGYERTH (Leipzig)

Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts[1]

We describe a method for the automatic alignment of parallel texts using co-occurrence statistics. The assumption of this approach is that words which are often found together are linked in some way. We employ this assumption to automatically suggest links between words in different languages, using Bible verses as information units. The result is a word-by-word alignment between between different translations of the Bible. The accuracy of our method is evaluated by using Strong's numbers as a benchmark. Overall, the performance is high, indicating that this approach can be used to give an approximate gloss of Bible verses.

1. Introduction

Using parallel texts for linguistic typology is a highly interesting and potentially fruitful approach. However, currently such work is tedious and highly laborious, as every example sentence from every language in the typological sample has to be interpreted individually by a researcher. In this paper, we will propose a method of automatic alignment[2] between translations that could help the interpretation of sentences in a language not intimately known to a researcher, thus possibly speeding up the process of gathering typological data. We envision a system in which a typological researcher selects particular stretches of text from a language of choice because they are considered potentially interesting for a particular linguistic question. Then the system will return the translational equivalents of these sentences in another language, suggesting also an approximate gloss. Of course, the selection, the full analysis, and the interpretation of the sentences will still be left to the typologist.

As an example, consider the verse John 14:6 from the English King James' Version: "Jesus saith unto him: I am the way, the truth, and the life: no man cometh unto the Father, but by me." The Estonian equivalent of this verse is shown in (1) and the Mandarin Chinese equivalent is shown in (2). The glosses given are the glosses suggested by the automatic procedure as described in this paper (unmatched words are indicated by a dash). Although the glosses are not perfect nor complete, they are helpful for a first analysis of these sentences.[3]

---

[1] We thank BERNHARD WÄLCHLI for useful comments on earlier version of this paper, and we thank BERNHARD COMRIE and GERHARD HEYER for making possible this cooperation between the Max Planck Institute for Evolutionary Anthropology and the University of Leipzig.

[2] Please note that the term 'alignment' is not used here in the linguistic sense (i.e. relating to the marking of arguments), but in the 'normal' meaning of putting things in line.

[3] B. WÄLCHLI (p.c.) informs us that the Estonian gloss does not have any errors. The inclusion of a demoted actor phrase in passive (*minu kaudu*, 'by me') is bad Estonian, but this is a problem with the Bible translation, not with our alignment. H.-J. BIBIKO (p.c.) informs us that the Chinese gloss almost perfect. Only the character glossed as 'but' does not mean *but*.

(1) Estonian (Uralic)

Jeesus	*ütleb*	*temale:*	*Mina*	*olen*	*tee*	*ja*	*tõde*	*ja*	*elu,*
Jesus	saith	him	I	am	way	and	truth	and	life
ükski	*ei*	*saa*	*Isa*	*juure*	*muidu*	*kui*	*Minu*	*kaudu!*	
man	no	–	Father	unto	–	–	I	by	

(2) Mandarin Chinese (Sino-Tibetan)

耶穌	說	我	就	是	道路、	真	理、	生命；			
Jesus	saith	I	–	am	way	truth	–	life			
若	不	藉	著	我，	沒有	人	能	到	父	那	裡	去。
–	but	by	–	I	no	man	–	–	father	–	–	–

This paper is organised as follows. First, there is some general discussion on our approach to automatic alignment. In Section 2, we present a short survey of the problem of automatic word-by-word alignment. In Section 3, the fundamental principle of our approach to this problem is presented, viz. *co-occurrence statistics*, which is based on the assumption that words are linked, when they are often found together in a corpus of a particular language. Then, in Section 4, we discuss how these statistics can be used for alignment between different languages. The basic idea is to count co-occurrences in the same sentences between two difference languages. Such count will be called *trans-co-occurrences*.

The second part of this paper presents an application of this method. Here, we attempt to align different translations of the Bible. In Section 5, we describe how we extracted a sentence-by-sentence alignment from Bible translations, and how we prepared such translations for our analysis. In Section 6, the sentence-by-sentence alignment is turned into a word-by-word alignment using trans-co-occurrences. Finally, in Section 7 the resulting word alignments are evaluated using a concordance-method as used in Bible exegesis: the so-called Strong's Numbers. The results of this evaluation are promising, suggesting that our approach to the alignment of parallel texts is worthwhile, and should be pursued further.

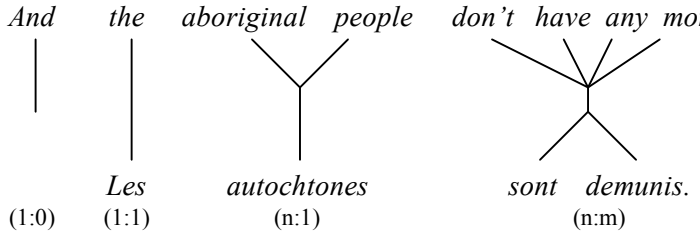
2. Word alignment

The task of word alignment is to link wordforms in a text to its correspondences in the translated text in another language, in such a way that the connected words supply the same contents. Computational proposals for this problem have been made starting in the late 1980's (cf. VÉRONIS 2000 for a survey). For most parallel texts, the problem already starts with the alignment of sentences. Given a text and its integral translation, which sentence in language *B* is to be considered the translation of a sentence in langage *A*? Much of the literature on automatic alignment deals with this problem. However, for our current task of aligning Bible translations, the sentence alignment is to a great extend already provided in the form of verse numbering, which is included in all Bible translations (cf. Section 5 for more details). The task thus is reduced to producing word-by-word linkage on the basis of given sentence-by-sentence (or better verse-by-verse) alignment.

The kinds of linkage attested varies depending on the typological structure of the languages and on the freedom of the translation. An example of word-by-word alignment is presented in Figure 1, following the examples and analysis by BROWN *et al.* (1990; 1993). A commonly attested type of linkage is a *1:1* association, exemplified here with the link between *The* and *Les*. In this case we can assume that the meaning of both wordforms are roughly equal. In *1:0* linkage, the equivalent of a particular wordform is not present in the translation, as shown for *And* in Figure 1. Often, words have to be associated with multiple words in the other language. This are so-called *1:n* or *n:1* associations, regularly found with compounds or fixed constructions (cf. *autochtones* in the figure). Figure 1 also highlights the most complicated case: a general *n:m* alignment, where on both sides multiple words are linked together. Although it is possible to divide these multi-word constructions into smaller parts in both languages separately, this cannot be done simultaneously in both languages in a compatible way. Such general *n:m* alignments will occur with high frequency when two rather strongly agglutinating or polysynthetic languages are aligned.

In this paper, we will approach the problem of word alignment using co-occurrence statistics. This method has, to our knowledge, not been attempted for the alignment of parallel texts. The research reported on here is only a first attempt at using this method for this goal, and there are various improvements possible. However, even with the rather basic implementation used here, we are already getting fairly good results, suggesting that this approach is worthwhile pursuing.

Figure 1. An alignment between an English sentence and a French translational equivalent showing different kinds of linkage.



3. Using co-occurrence statistics

The goal of co-occurrence statistics is to extract pairs of words that are associated from a corpus. The underlying assumption is that while generating text, people are complying to syntactic and semantic restrictions of their (natural) language in order to produce correct sentences. When analyzing a large quantity of text (a text corpus), words that tend to appear together will reflect these linguistic restrictions. While it is generally possible to produce sentences containing arbitrary pairs of

words, in most of the cases the words appearing together will have something to do with each other and statistics will be able cut out the noise.

The joint occurrence of words within a well-defined unit of information, for example the sentence, a whole document, or a word window,[4] is called a co-occurrence. The simplest co-occurrence statistics would be to count how often two words co-occurr within all units of information in the corpus. However, because more frequent words have higher probabilities in appearing together with any word, just because they are frequent, this will not give meaningful associations. Therefore, a significance measure is applied that takes the single word frequencies as well as their joint frequency into account. In our experiments, we use a log-likelihood measure that, intuitively speaking, measures the amount of surprise to see two words co-occurring together as often as they do, compared to the statistical expected number of co-occurrences if we assume independence of occurrence. Here, the significance values for the co-occurrence of two words A and B are calculated according to the formula as shown in (3), cf. BIEMANN *et al.* (2004a).

$$(3) \qquad sig(A,B) = \frac{x - k\log x + \log k!}{\log n}$$

n = number of units of information in the corpus
k = number of joint occurrences of A and B within a unit of information
$x = ab/n$
a = number of occurrences of A in the corpus
b = number of occurrences of B in the corpus

The significances are computed for every pair of words in the corpus. The significance values give us the possibility to rank the co-occurrences of a given word, as higher significance values denote a higher degrees of association. Normally, such statistics are applied on monolingual corpora, and the results are semantic nets. Semantically related words tend to show a high degree of association.[5]

4. Trans-co-occurrences

When applying co-occurrence analysis to multi-lingual parallel texts, we are interested in the association between pairs of wordforms, each from a different language. In that usage, co-occurrence statistics can automatically extract translational equivalents of wordforms, given a sentence-aligned bilingual corpus. Given a sentence translation pair we merely calculate significant co-occurrences between wordforms from different languages and call them *trans-co-occurrences*. If a wordform A in the first language is always translated into wordform B in the second

[4] A *word window* is a stretch of text defined relative to a central word X within a given window size S. The word window around X consists of all words occuring next to X up to maximally S words away. For example, the window of size three around the word 'text' as occurring in the first line of this footnote, consists of the words {a, stretch, of, defined, relative, to}.

[5] This property can be used to create semantic networks for short texts or spoken language streams as discussed in BIEMANN *et al.* (2004b)

language, then *B* will be the highest ranked trans-co-occurrence of *A*. In contrast, often *A* will have various high ranked trans-co-occurrences, normally all with clearly smaller significance values, which represent alternatively possible translations. In this general case, there are several possibilities to translate a wordform from one language into another. In this situation, the most prominent translation will be ranked highest, followed by less prominent translations and finally noise.

Given the data obtained by trans-co-occurrence statistics, it is possible to construct dictionaries from parallel texts in a fully automatic way.[6] All trans-co-occurrences above some significance threshold will be entered in the dictionary. The quality, as compared to manually compiled dictionaries, can be estimated at 60%–80% correctness (SAHLGREN 2004, BIEMANN & QUASTHOFF forthcoming). However, here we are currently not interested in building up dictionaries, including all possible meanings of a particular word, but in word-by-word alignment between two given translational equivalents; in our case of the Bible.

5. Preparations: sentence alignment and markup

For our research, we used Bible translations from the SWORD PROJECT as parallel corpora.[7] To calculate the trans-co-occurrences, two Bibles were merged to a new bilingual bible. Using the Bible's verse numbering as anchors, we combined corresponding sentences to a new longer sentence through concatenation. In principle, we could have simply concatenated whole verses, but we decided to try to restrict the information unit because we were afraid that verses would be too long to yield significant co-occurrences.[8] We tried to restrict the information unit to, roughly, the size of a sentence. To achieve this, we first splitted verses into smaller parts, using full stops and semicolons as separators. If the number of parts obtained is identical for the two languages, then we splitted the verse. However, if the number of parts is not identical, we kept to the complete verse. For example, consider the verse Genesis 1:2 in the English King James Version (KJV) and the German Luther translation as shown in (4).

(4) a. And the earth was without form, and void;
and darkness was upon the face of the deep.
And the Spirit of God moved upon the face of the waters.

b. Und die Erde war wüst und leer, und es war finster auf der Tiefe;
und der Geist Gottes schwebte auf dem Wasser.

---

[6] Note that such an automatically generated dictionary would be a dictionary of wordforms, and not the classical type of linguistic dictionaries only listing lexemes.

[7] http://www.crosswire.org/sword/index.jsp

[8] With hindsight, seeing the results of our investigation, we now think that this step was not necessary. The algorithm that we have used seems to be robust enough to cope with longer information units, like whole verses of the Bible. However, it is to be expected that using larger information units requires more instances (i.e. parallel units) to get reliabale statistics. Of course, by taking larger units, we end up with less units, and would probably get worse results.

As can be seen from this example, after splitting the verse the number of obtained parts differs between the two languages. The English version (4a) consists of three parts, but the German translation (4b) only consists of two parts. So in this case, we are unable to restrain the information unit. The whole verses are simply concatenated into a bilingual sentence, as shown in (5). For the automatic distinction of the languages, each word was marked with language-identifying tags, like '@en' for English or '@de' for German, as shown in (6).

(5) And the earth was without form, and void; and darkness was upon the face of the deep. And the Spirit of God moved upon the face of the waters. Und die Erde war wüst und leer, und es war finster auf der Tiefe; und der Geist Gottes schwebte auf dem Wasser.

(6) And@en the@en earth@en was@en without@en form@en and@en void@en and@en darkness@en was@en upon@en the@en face@en of@en the@en deep@en And@en the@en Spirit@en of@en God@en moved@en upon@en the@en face@en of@en the@en waters@en Und@de die@de Erde@de war@de wüst@de und@de leer@de und@de es@de war@de finster@de auf@de der@de Tiefe@de und@de der@de Geist@de Gottes@de schwebte@de auf@de dem@de Wasser@de

Following this approach, two Bible translations can be combined into one language-tagged bilingual Bible. This bilingual text can then be used to compute the trans-co-occurrences for each word.[9]

6. Algorithm for word alignment

Using the trans-co-occurrence statistics, any wordform in a particular sentence from the Bible will now be linked to a wordform in the other language (we used the occurrence of spaces in the text as wordform delimiters). To demonstrate our approach to such word alignment, consider the verse Luke 11:4, as shown in (7) – the English KJV translation in (7a) and the German Luther version in (7b).

(7) a. And lead us not into temptation; but deliver us from evil.
b. Und führe uns nicht in Versuchung, sondern erlöse uns von dem Übel.

From this verse, we have selected the English words *temptation* and *deliver* as exemplars. The German trans-co-occurrences of these English words are tabulated in Table 1 and Table 2, respectively, ordered by the co-occurence significance. The highest ranked words are thus considered to the best overal translational equivalent. However, these tables are based on the whole Bible, so all kind of words do occur, irrespective of the actual words that are found in the German version of the verse Luke 11:4. (The words that occur in this verse are printed in boldface in the tables.) If we would simply take the highest ranked word also present in the Ger-

[9] The procedure to compute the (trans-)co-occernces is described in detail in BIEMANN *et al.* (2004a).

man sentence as the best match, then the English *temptation* is correctly linked to the German *Versuchung*. However, as can be seen from Table 2, the English *deliver* is then wrongly linked to the German *nicht*. The pair (*deliver*, *nicht)* has a higher significance value than the correct pair *(deliver*, *erlöse*). This error sometimes occurs with highly frequent words like *nicht* or *in*.

The basic idea to alleviate this problem is to combine the ranks of the significance statistics looking from English to German with the statistics when looking from German to English. For example, the English *deliver* suggested *nicht* as the best match (on rank 15). However, when we look at the trans-co-occurrence statistics for the German word *nicht*, the English word *deliver* is only ranked as match number 44. In contrast, for the German word *erlöse*, the English word *deliver* ends up as the highest ranked trans-co-occurrant, though it was only ranked on number 19 in Table 2. The pair (*deliver*, *nicht*) has thus ranks 15 and 44, which seems intuitively worse than the pair (*deliver*, *erlöse*) with ranks 19 and 1. We formalized this intuition by defining a MATCH VALUE *m* for a pair of English-German words as shown in (8), based on the multiplication of the two rank-numbers.[10] On the basis of this value we get the right match, because the match value *m*(*deliver*, *erlöse*) is 0.229, which is clearly higher than the match value *m*(*deliver*, *nicht*), which is 0.039.

$$(8) \qquad m(e,g) = \frac{1}{\sqrt{rank_e(g) \cdot rank_g(e)}}$$

Table 1. Ranked German trans-co-occurrences of the English word *temptation*. A selection of words from the German version of Luke 11:4 are printed in boldface.

rank	word	overall corpus frequency	number of co-occurrences	co-occurrence significance
1	**Versuchung**	10	9	59
2	fallet	6	4	26
3	Anfechtung	8	4	25
4	verstocket	4	2	13
5	betet	39	3	13
…				
7	**erlöse**	12	2	11
10	**Übel**	61	2	8
12	**nicht**	7541	11	7

[10] The square root in this formula prevents the *m* values from becoming small very quickly, which might lead to many, possibly confusing, decimal zeros. However, this use of the square root is basically irrelevant, as we are only interested at the relative ordering of the resulting *m* values, and not at their absolute magnitude.

Table 2. Ranked German trans-co-occurrences of the English word *deliver*. A selection of words from the German version of Luke 11:4 are printed in boldface.

rank	word	overall corpus frequency	number of co-occurrences	co-occurrence significance
1	erretten	79	71	260
2	errette	37	34	126
3	Hand	1052	79	109
4	Hände	408	45	78
5	geben	592	47	68
…				
15	**nicht**	7541	117	27
19	**erlöse**	12	7	24
22	**uns**	1525	39	22
59	**führe**	42	5	10
70	**Versuchung**	10	3	9

In this way, the best translational equivalent for a particular word can be found with rather great precision (see the next section for an evaluation of this approach).[11] However, the match value is even more informative because the height gives an indication of how good is the best match that is found. The best possible result is achieved when the matched words are both the higest ranked trans-co-occurrences. Both ranks are then one, and the resulting match value *m* is 1.00. If the matched pair is less directly equivalent, the match value will be lower (cf. *m*(*deliver*, *erlöse*) = 0.229 as discussed above). The height of this value can be used to select only the best translations. Allowing also lower valued matches, more words are actually linked to a translation. However, there will also be some more errors included. This trade-off is investigated in the next section.

7. Using Strong's Numbers as a benchmark

To evaluate the results of our algorithm, we used the so-called 'Strong's Numbers' that are available for some Bible translations. These numbers are annotations added to a Bible text following a system devised by JAMES STRONG in the 19th century. JAMES STRONG (1822-1894) was professor of exegetical theology at Drew Theological Seminary (Madison, New Jersey). Under his guidance, an exhaustive concordance between the King James Version (KJV) of the Bible and the Hebrew

[11] The resulting tables of trans-co-occurrences are a higly valuable resource for other research as well. Note, for example, that it is also possible to use the trans-co-occurrence statistics, as obtained by analysis of the Bible, for the translation of other, yet untranslated texts. However, we can not use the bidirectional match value in that case, but only the ranking as implicit in the trans-co-occurrence statistics.

Old Testament (i.e. the Masoretic Text, called *Tanakh* in Hebrew) and the Greek New Testament (i.e. the *Textus Receptus*) was compiled, apparently with the help of more than a hundred unnamed colleagues. This concordance first appeared in 1890. It is based on a dictionary of all words occurring in the Hebrew and Greek Bibles, which are numbered along their alphabetical order. These numbers are then inserted in the English text of the KJV. Following this example, the same numbers were later also added to various other translations of the Bible.

As an example, consider the verse Revelation of John 1:8 from the New Testament in the KJV translation, as shown in (9). The Greek letter *A*, translated into English as 'Alpha', is the first entry in the Greek alphabetical listing. Accordingly, the word 'Alpha' in the KJV translation is marked with the number <1> behind it. The main difference between these Strong's Numbers and a modern XML-style mark-up is that the Strong Numbers only mark the end of the entry and not the start. This leads to some problems for automatic processing, because it is not clear exactly to which part a Strong's Number refers. For example, the words *is to come* in (9) are not individually marked by a Strong's Number, but only as a group. In most cases, the Strong's Number appears to be placed immediately following the main lexical equivalent of the word in the Greek or Hebrew text. We decided to include only this last word before a Strong's Number for the evaluation of our algorithm. Also note that in some cases there are multiple Strong's Numbers associated with one part of the English translation (e.g. the same phrase *is to come*, associated with the numbers 2064 and 3801). This situation arises because in some cases there are multiple words in the Greek or Hebrew texts which are translated as just one word or phrase into English. We included both numbers for testing the results of our algorithm.

(9) I <1473> am < 1510> **Alpha <1>** and <2532> Omega <5598>, the beginning <746> and <2532> the ending <5056>, saith <3004> the Lord <2962>, which <3588> is <5607, 3801>, and <2532> which <3588> was <2258, 3801>, and <2532> which <3588> **is to come <2064, 3801>**, the Almighty <3841>. (KJV, Rev. 1:8)

When two translations of the Bible are both marked with Strong's Numbers, then these numbers can be used to evaluate an automatically generated alignment. There are four different situations that can occur when comparing the automatic alignment with the Strong's Numbers:

- **Correct:** the aligned words are both followed by a Strong's Number, and these numbers are identical (in case there is only one number) or show an overlap (in case there are multiple numbers)
- **Error:** the aligned words are both followed by a Strong's Numbers, but these numbers are different (in case there is only one number) or do not show any overlap (in case there are multiple numbers)
- **One-sided miss:** only one of the aligned words is followed by a Strong's Number, but the other is not.
- **Uninformative:** both aligned words are not followed by a Strong's Number.

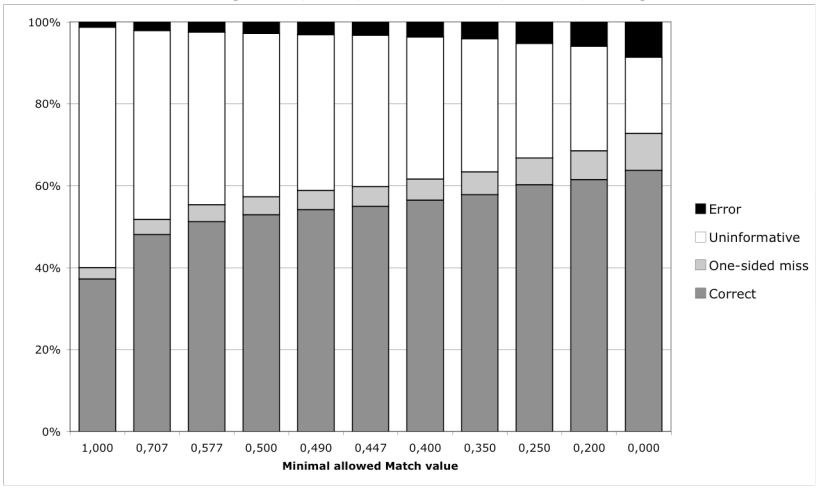
As an example, compare the KJV translation in (9) with the German translation by Luther in (10). When the automatic alignment aligns the English *I* to the German *Ich*, this is counted as correct because both are followed by the same Strong's Number <1473>. However, if the algorithm aligns *am* with *ich*, this would clearly be an error, as the words are followed by different Strong's Numbers. One-sided misses occur for example when the English *Lord* is (correctly) aligned with *Gott*. Although this is correct, this alignment cannot be validated because there is no Strong's Number directly following the German *Gott*. From some random inspection of such cases, we suspect that the far majority of such one-sided misses are actually correct alignments that are obscured by the specific placement of the Strong's Numbers in the text. Finally, there are alignments that cannot be interpreted because both words are not followed by a Strong's Number. For example, neither the article *the* nor *der* are followed by a Strong's Number, and are thus uninformative for the evaluation.

(10) Ich <1473> bin <1510> das A <1> und <2532> das O <5598> , der Anfang <746> und <2532> das Ende <5056> , spricht <3004> Gott der HERR <2962> , der <3588> da ist <3801> und <2532> der <3588> da war <2258> <3801> und <2532> der <3588> da kommt <2064> <3801> , der Allmächtige <3841> . (Luther, Rev. 1:8)

The actual number of errors and correct alignments depends on the MATCH VALUE *m(e,g)*, as defined in the previous section. The match value gives an indication how good the algorithm evaluates a particular alignment of two words between the translations. An alignment with the highest possible match value of 1.00 means that the algorithm rates this as a good match; a lower match values indicates less confidence. In Figure 2, we show the evaluation of the English (KJV) - German (Luther) alignment, depending on the allowed match values. In the first column, only the alignments with a match value of 1.00 are shown. As can be seen in Figure 2, more than 50% of these alignments are uninformative. If lower match values are allowed, this portion becomes smaller, but also the number of errors increases. To show this trade-off between accuracy and overall performance, we defined measures for precision and recall on the basis of these validations, as shown in (11). These values for precision and recall are rather conservative and thus very probably lower than the actual performance of the automatic alignment. We expect that most of the one-sided misses and many of the uninformative cases are actually correct alignments. However, we have no way to assess that more precisely at this point.

(11) Precision = correct / correct + error + one-sided miss
Recall = correct / all alignments

Figure 2. Evaluation of English (KJV) - German (Luther) Alignment



We computed the precision and recall for every match value (i.e. for every column in Figure 2). The resulting values are plotted in Figure 3, connected by a line There are two lines shown in this figure because we performed the alignment directionally. One line in the figure represents the precision and recall for the direction where we started with the English translation and then tried to find the best match in the German translation. The other line represents the inverted procedure. Interestingly, the precision from English to German is better than the other way around, although the recall roughly remains the same. This is probably caused by the fact that German has more mophology than English, and consequently the German translation has less words. The resulting major difference is that the number of one-sided misses is clearly higher for the direction German to English.

Figure 3. Trade-off between precision and recall for the English-German alignment

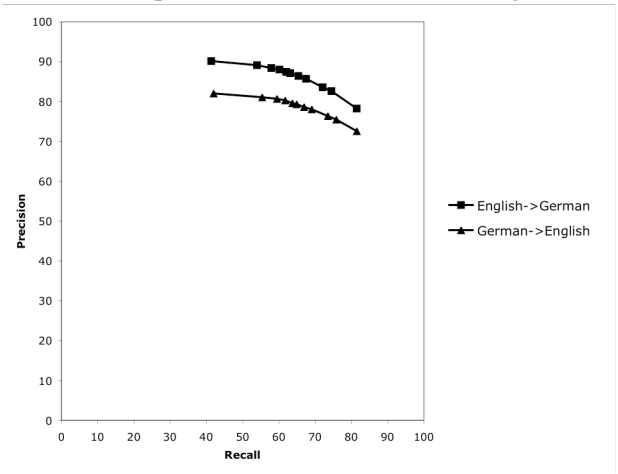
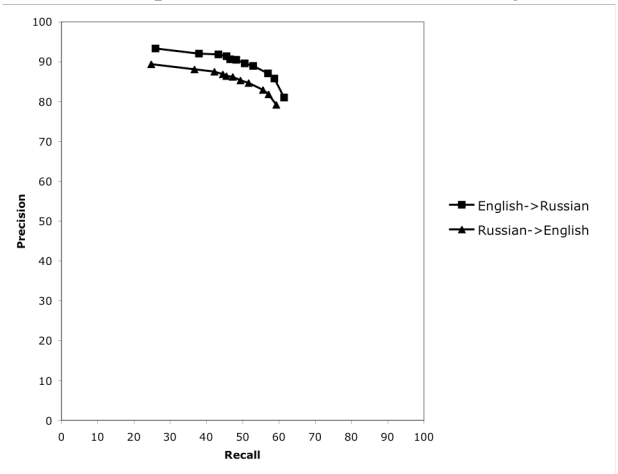


Figure 4. Trade-off between precision and recall for the English-Russian alignment

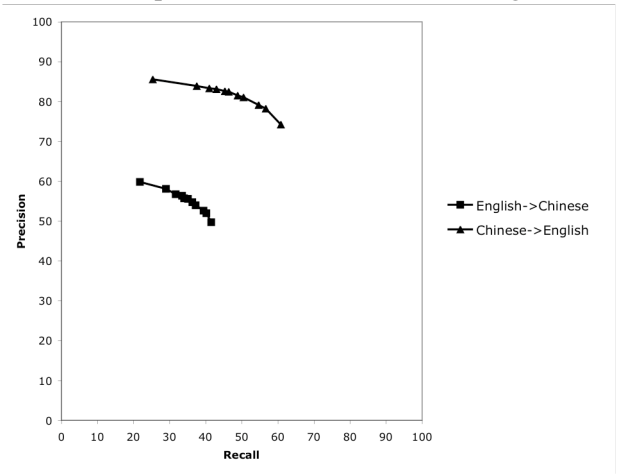


We performed the same evaluation for the alignment of the English KJV translation with the Russian 'Synodal' translation from 1876. The results of the evaluation using the Strong's Numbers is shown in Figure 4. The precision is comparable to the English-German alignment, but the recall is much worse. This is the result of a much higher fraction of uninformative alignments. This is probably caused by the multitude of textual variants of the 'original' texts. The Synodal text of the Russian orthodox Church is probably based on a different original as the English KJV translation (see DE VRIES, this issue).

Finally, we also evaluated an English-Chinese automatic alignment by using the Chinese 'Traditional Union' translation, which has also been annotated by Strong's Numbers. The results are shown in Figure 5. The first aspect to take note of is the large discrepancy between the two directions of alignment. The alignment from English to Chinese is much worse than the alignment from Chinese to English, although for the alignments with German and Russian this direction even performed slightly better. The reason for this large discrepancy is that we did not parse the Chinese text for words.[12] The algorithm simply looked for the best match between any Chinese character and any word in the English text. However, most lexical words in the English text is translated by multiple Chinese characters. Now, for the evaluation of our algorithm we also took the first Chinese character before any Strong's Number. If we start from an English word followed by a Strong's Number, the best match will very often not be the Chinese character directly in front of the Strong's Number, but one of the other characters that also are part of the translation. As a result, we get a very high proportion of one-sided misses for the direction English to Chinese, which diminishes the precision. In contrast, for the direction from Chinese to English, the precision is roughly on the same level as for the alignment from German to English. The recall is worse because of a much higher proportion of uninformative matches.

[12] Of course, this could have been done, e.g. by http://www.mandarintools.com/segmenter.html.

Figure 5. Trade-off between precision and recall for the English-Chinese alignment



This directional difference with the Chinese-English alignment suggests an interesting consequence for the alignment between English and morphologically more complex languages (and that is why we did not parse the Chinese text for words). English could be considered a much more synthetic language compared to the Chinese *script*, as most English words map onto multiple Chinese characters. Of course, such a comparison does not make any sense linguistically. However, this way to look at it argues that the results from our alignment between English words and Chinese characters might be interpreted as showing what would happen if we would try to align English to a more synthetic language. Starting from the morphologically more complex language is difficult for our algorithm (cf. the direction English to Chinese). However, using the alignment from the more isolating language to the more synthetic language seems to give relatively good results (cf. from Chinese to English), even though the structure of the languages are very different. Of course, it would be better to check this claim by actually trying to align the English text to a language with a more complex morphology. Our algorithm does have no problem providing an alignment between English and, say, the Swahili New Testament (which is also available electronically as open source), but we have no way to automatically check such an alignment because there are no Strong's Number added to the Swahili translation (nor to any other translations of morphologically more complex languages).

8. Conclusions

The usage of trans-co-occurrences is a highly promising method to establish translational equivalents in parallel texts. Even in the simple and straighforward version that we used in this paper, the results are already fairly good. At least good enough to provide typologists with an approximate gloss of a stretch of text, which can then subsequently be analysed in more detail by hand.

An important characteristic of our algorithm, which makes it even more interesting for typology, is that there is no knowledge needed about the languages that are to be combined. The algorithm is completely language-independent. The only information that is assumed is an aligned information unit (in our case, the Bible verses) and a word-separator (we simply used the occurrence of spaces). However, one could easily improve this method by adding information—also possibly extracted automatically. For example, instead of a word-by-word alignment, a morpheme-by-morpheme alignment can be attempted, presupposing that we know about the morpheme separation of both languages. In the other direction, another possisble enhancement would be to mark frequent collocations in each language, and not align the individual words, but whole chunks of possible idiomatic expressions.

In contrast, instead of adding information beforehand, it is also possible to use the trans-co-occurrences (as, for example, extracted from Bible translations) for further linguistic analysis. For example, it turns out that (inflectional) morphological variations of the same root often occur together in the trans-co-occurrences (cf. *erretten/errette* and *Hand/Hände* in Table 2). This suggest that trans-co-occurrence statistics might also be used to investigate the inflectional structure of a language. However, all these suggestions are left for further research.

References

BIEMANN, CHIRISTIAN AND QUASTHOFF, UWE (forthcoming): Dictionary acquisition using parallel text and co-occurrence statistics, in: *Proceedings of NODALIDA-05, Joensuu, Finland.*

BIEMANN, CHRISTIAN; BORDAG, STEFAN; HEYER, GERHARD; QUASTHOFF, UWE & WOLFF, CHRISTIAN (2004a): Language-independent methods for compiling monolingual lexical data, in: *Proceedings of CicLING 2004, Seoul Korea.* [LNCS 2945]. Berlin: Springer, 215–228.

BIEMANN, CHRISTIAN.; BÖHM, K., HEYER, GERHARD, MELZ, R. (2004b): Automatically building concept structures and displaying concept trails for the use in brainstorming sessions and content management systems, in: *Proceedings of I2CS, Guadalajara, Mexico.* [LNCS 3473]. Berlin: Springer, ppp–ppp.

BROWN, PETER F.; COCKE, JOHN; DELLA PIETRA, STEPHEN A.; DELLA PIETRA, VINCENT J.; JELINEK, FREDRICK; LAFFERTY, JOHN D. ; MERCER, ROBERT L. & ROOSSIN, PAUL S. (1990): A statistical approach to machine translation, in: Computational Linguistics 16/2, 79–85.

BROWN, PETER E.; DELLA PIETRA, VINCENT J.; DELLA PIETRA, STEPHEN A. & MERCER, ROBERT L. (1993): The mathematics of statistical machine translation parameter estimation, in: *Computational Linguistics* 19/2, 263–311.

SAHLGREN, M. (2004): Automatic bilingual lexicon acquisition using random indexing of aligned bilingual data, in: *Proceedings of LREC-2004*, *Lisboa, Portugal*, 1289–1292.

VÉRONIS, JEAN (2000): From the Rosetta stone to the information society: A survey of parallel text processing, in: VÉRONIS, JEAN (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora.* [Text, Speech and Language Technology 13]. Kluwer: Dordrecht, 1–24.

Correspondence address

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

ÖSTEN DAHL (Stockholm)

From questionnaires to parallel corpora in typology[1]

This rather programmatic paper discusses the use of parallel corpora in the typological study of grammatical categories. In the author's earlier work, tense-aspect categories were studied by means of a translational questionnaire, and cross-linguistic gram-types were identified through their distribution in the questionnaire. It is proposed that a similar methodology could be applied to multilingual parallel corpora. The possibility of identifying grammatical markers by word-alignment methods is demonstrated with examples from Bible texts.

1. Introduction

Research in language typology is heavily constrained by the difficulties in creating adequate data sets. Even in the case of comparatively well-described languages, which constitute a small minority, the information found in reference grammars and more specialized publications tends to be insufficient and often misleading. This is in particular the case for grammatical categories such as tense, mood, aspect, number, definiteness, case etc., which depend on a mixture of syntactic, semantic, and pragmatic factors, many of which are only poorly understood. For the description of such categories descriptive grammarians often rely on traditional definitions and stock examples. Without personal knowledge of a language, a typologist can only make limited use of texts and even if glossed texts are available, the low text frequency of many interesting phenomena makes it difficult to find more than a few examples, and those are often hard to interpret.

2. Questionnaires

If one finds an interesting example in language A, the natural question is to ask "How would this be expressed in languages B, C etc.?" If the answer cannot be found in a grammar, as is often the case, the obvious way of getting the answer is to ask a native speaker. A more systematic approach to this is to use a questionnaire, the most straightforward type of which is a translational questionnaire, containing a set of expressions, sentences or connected texts to be translated by a native speaker into the language under investigation. A well-constructed questionnaire covers a certain area of grammar in such a way that it gives information about the ways in which this part of the grammar is structured in the language in question by yielding a set of translational equivalents between the source language and the target language, and indirectly, between different target languages. The notion of translational equivalent should be understood in an operational and theory-independent sense: an expression α in one lan-

[1] I am grateful to the editors of this issue for many valuable comments, and to JAN OLOV PERSSON for consultations on statistical problems.

guage is a translational equivalent of an expression β in another language if α is actually used (more than occasionally) as a translation of β by persons who are competent in both languages. How this behaviour should be interpreted is another matter. An obvious limitation of the questionnaire approach is that the choice of expressions to be translated has to be guided by the questionnaire constructor's understanding of the phenomena being studied – which may quite negatively influence the chances of making new discoveries.

About a quarter of a century ago, I initiated a questionnaire investigation of tense and aspect systems (originally also including mood) which formed the empirical basis for DAHL (1985). The questionnaire consisted of about 200 sentences in context and applied to a sample of 64 languages. The first step in the analysis of the questionnaires was to mark up every verb with a code for its tense-mood-aspect features. Obviously, this required knowledge of the structure of the language, so in many cases the help of experts on the individual languages was invaluable. The second step was to look for clusters of categories with similar distribution across languages. That is, the goal was to find forms or constructions from different languages that showed up in the same, or roughly the same, places in the questionnaires. As it turned out, for most of the cases where a form or construction had a reasonably large number of occurrences in the questionnaire, it was possible to assign it to such a cross-linguistic cluster, which could then be assumed to represent what JOAN BYBEE and I later named "cross-linguistic gram-types" (BYBEE & DAHL 1989). Examples of such gram-types would be the Past, the Future, the Perfective, the Imperfective, the Progressive, the Perfect, the Experiential, and so on.

How does one find such clusters in the first place? It would in principle be possible to run through all the questionnaires and find correlations between all the grams coded in the analyses. However, given the relatively limited capabilities of computers in the beginning of the eighties, this did not appear to be practically feasible, and I instead used the following kind of heuristics: departing from a known gram G in some language, I looked for grams that seemed to have a similar distribution to G, by computing their correlation to G. The distribution of these grams in the questionnaire was then taken as the first approximation to the "ideal distribution" of the purported gram type, after which the list of individual candidate grams was adjusted to this approximation. When I had performed this operation a number of times, I had defined a cluster of grams that would be reasonably independent of the gram I had started from.

As for the results of the investigation, I think it can be said that although they do not in any serious way contradict what was generally said in the literature at that time, the investigation contributed to sharpening the picture of what tense and aspect systems in human languages are like, especially in conjunction with the grammar-based typological investigations of verbal categories led by JOAN BYBEE (BYBEE 1985; BYBEE et al. 1994). At the same time, in spite of the rapid development of language typology, and although questionnaires are now a standard tool for typologists, I do not know of any investigations that have tried to apply the methodology I used. This probably has to do

with the inherent difficulties in the method. A general problem is that a typological questionnaire investigation with a good coverage is quite costly. It takes a considerable time to develop a good questionnaire, and at the point where it is mature, you may already have used up most of your available informants as guinea-pigs. Almost unavoidably, the set of languages investigated will be a convenience sample, that is, the choice will depend more on the availability of bilingual and literate informants than on a principled sampling method. (The sample in DAHL 1985 did contain a fair number of non-European languages but was still quite heavily biased. Thus, 21 languages – that is almost a third of the sample – were Indo-European).

3. An alternative: parallel corpora

The question is now if there is any possibility of overcoming the limitations of the questionnaire method without losing its advantages. An obvious alternative when looking for translational equivalents is to use parallel corpora (which hardly existed around 1980, at least not in an easily accessible form). Even if most existing parallel corpora are not suitable as bases for typological investigations in that they normally contain texts in a very limited number of languages, typically European ones, the technological developments of recent years have now made parallel corpora a practical possibility for typologists, as is amply demonstrated in the papers in this issue. The text that has been translated into the largest number of languages is the Bible, and since Bible translations are often the main source of knowledge for extinct languages such as Gothic and Old Church Slavonic, the use of Biblical texts as a basis for language description has an old tradition. (In the case of modern Bible translations, the relationship between translation and grammar description is usually the opposite, in that the latter is a prerequisite for the former.)

Bible translations have a number of features that make them attractive as a basis for parallel corpora in typological research:

1. The languages into which the Bible has been translated wholly or partially are spread fairly evenly over the globe, making the creation of a relatively unbiased sample seem possible.
2. Many Bible translations are readily available for download from the Internet. (However, the set of freely downloadable Bibles, regrettably, looks rather like your typical convenience sample, with a heavy bias towards translations into European and a few major non-European languages.)
3. The Bible is really a collection of quite heterogeneous texts of different genres, including straightforward narratives and argumentative passages.
4. Even if the Bible (like virtually all parallel corpora) represents written language, there is a considerable amount of natural-sounding direct speech in it.
5. Bible texts are usually well prepared for use in parallel corpora, in that the partitioning into chapters and verses can serve as a substitute for sentence alignment.

Strong's Numbers (see CYSOUW *et al.*, this issue), for the translations where they exist, can even provide word alignment.

6. At least in the case of the New Testament, versions of the original text (Greek[2]) with complete lexical and morphological markup are freely available.[3]

It goes without saying that there are also problems and drawbacks. The complex relationship between translations and originals and between different versions of the original texts is discussed elsewhere (DE VRIES, this issue). From the present perspective, it can be noted that there is a trade-off between "alignability" and empirical relevance, in that a more literal translation is easier to align with the original but may tell us less about the target language, whereas a translation that aims at transmitting the message in a natural way rather than rendering the original literally will potentially tell us more about the language as it is spoken but will be more difficult to align and parse. Apparently, one cannot have it both ways (and sometimes one gets neither). A dimension that, strictly speaking, is separate from that of the literalness of the translation is the degree to which Bible translations tend to become a genre in themselves, even developing into a separate language variety. Thus, in English, "KJV-ese", as the language of King James' Version might be called, is used both in many modern editions of the Bible as well as in other documents such as Mormon's Book. In many cases, it may be safest not to see Bible translations as representative of anything but themselves, but as samples of written language they are not worse than any other texts.

The total length of the King James Version of the Bible is (approximately) 800,000 words; of these, about 180,000 make up the New Testament. The Greek text contains only about 140,000 words. The variation here is great – the West Greenlandic New Testament is merely 60,000 words long. There are a number of reasons for restricting a parallel Bible corpus to the New Testament, at least initially. Most importantly, a large part of existing translations, in particular for non-European languages, comprise the New Testament only. It is also easier if one has to deal with one source language only, and, as I have already mentioned, fully marked up versions exist only for the New Testament. Furthermore, the sheer length of the Tanakh/Old Testament may make it difficult to handle it computationally, although on the other hand, statistical analyses will yield more reliable results with a more extensive corpus. Consequently, I will in the following be speaking of a corpus that consists of a set of translations of the New Testament.

When I worked on the TMA questionnaires, I had the advantage that the verb forms were already marked up by experts on the respective languages. The fundamental problem of parallel corpora studies, that of alignment, thus did not exist. When comparing the distribution of grammatical items (morphemes, constructions etc.) in Bible transla-

[2] In the following, "Greek" will refer to the Hellenistic or Koine Greek in which the New Testament was written.

[3] See for instance http://users.mstar2.net/broman/editions.html.

tions, on the other hand, we do not in general have access to grammatically analyzed texts – with one important exception: the Greek original. We must therefore find a method to match or align the grammatical items across languages. This is not an easy task and it is obvious that before we can do anything similar to what I did with the TMA questionnaires a huge amount of work is needed.

Work on alignment of parallel texts below the sentence level has (to the extent that I am acquainted with it, at least) been mainly concerned with the alignment of words, and less with the alignment of grammatical structure and grammatical morphemes. The general principle, however, has to be the same for lexical and grammatical meaning: we identify items by assuming that items that have similar distributions are also likely to play the same role in the texts. In fact, this global method is the same as the one I applied to TMA questionnaires in DAHL (1985). That is, the search for cross-linguistic categories and the analysis which has to be done for a parallel corpus to be useful takes the same form. Moreover, it seems to me that the alignment process is helped by an adequate division of labour between the lexical and grammatical analyses.

To an astonishing extent, grammatical or functional words can be identified with high-frequency words – at least in the languages I have looked at, and I see no reason why it should not be the case universally. Thus, in the KJV New Testament, the most frequent word which is unequivocally lexical rather than grammatical is *God*, which has rank 23 and frequency 1,372. Now, if one tries to run a word-alignment algorithm on a Bible translation along the lines suggested in CYSOUW *et al.* (this issue), it turns out that high-frequency words create special problems. The three most frequent words in the King James Version of the New Testament are *the* (11,036 occurrences), *and* (10,721 occurrences) and *of* (6,129 occurrences). In the Greek New Testament, one single word-form, *kai* 'and', occurs more often than the following three words on the ranking-list taken together – it is found 9,208 times in the text. If we instead consider the Strong numbers, which reflect lexical items rather than word-forms (with a few exceptions), we find that Strong's Number 3,588, which represents the Greek definite article in its various forms,[4] occurs no less than 20,317 times, that is approximately 14.5 per cent of the whole text, and on average 2.5 times per Bible verse. Word-alignment procedures discussed in the literature often follow the principle of dividing up the texts into aligned chunks, and then compute the probability that a word w_1 in a source text co-occurs with a word w_2 in the target text in a chunk c. As noted above, the verse constitutes a natural unit in Bible texts, and it would seem natural to use it also in word alignment—this is also suggested to be feasible in CYSOUW *et al.* (this issue). However, for high-frequency elements such as definite articles, which tend to occur several times in each verse, this does not seem to be a very good idea – the number of false combinations will simply be too large. This is a problem I shall return to below. But it is not only the high text frequency of grammatical items such as the defi-

[4] Some Bible translations annotated with Strong's Numbers do not provide them for function words, presumably because these are considered less essential for the content.

nite article that creates problems for word-alignment but also their cross-linguistic variability. Thus we know that many languages lack definite articles altogether. If a high-frequency grammatical word in the source text does not correspond to anything at all in the target text and vice versa, this creates a considerable amount of noise (in the technical sense of that word) for the word-alignment procedure. In particular, if the target text contains a grammatical item not found in the source text, there is no way of identifying it from the source text alone. In a multilingual parallel corpus, however, this problem can possibly be solved if we study the cross-linguistic distribution of gram-types, such as definite articles. If we know where in a text grammatical items of different cross-linguistic types are likely to appear, we'll be able to assign high-frequency items to those types before starting to align lexical words. Thus the study of the cross-linguistic patterns in the distribution of grammatical items in parallel corpora is needed for the understanding of cross-linguistic gram-types and for the word-alignment process in general.

As I suggested in the preceding paragraph, the verse may be too large a unit when studying the distribution of grammatical items in Bible texts. I would suggest that the best solution is not to try and divide up verses in smaller chunks on the basis of punctuation or other signals. Rather, one should use a moving "word window", which means that for a given word in text A we consider the words that are at a distance of no more than *n* words from the corresponding position in text B, for some suitable value of *n*. An easy way to define the position of a word in the Bible text is by identifying the verse where it occurs and its position (counted in numbers of words) from the beginning of that verse. When comparing different Bible texts, the problem arises that the length of verses will not always be the same. This can be circumvented by a process of normalization: a verse is treated as if had the same length as in the Greek original and the positions of words in translations are recomputed accordingly. In this way, each word will have a number that identifies the most probable counterpart in the original text. The existence of Strong-numbered translations makes it possible to study how words in translations are distributed relative to the source words. In eight translations representing six European languages (English, Dutch, French, German, Portuguese, Russian), I found that of the words in the Greek texts that were assigned Strong numbers in the translations at most a few per cent were found at a (normalized) distance of more than five words from the original. Since the languages where translations with Strong numbers are available are a rather bad sample from the typological point of view, I have also performed a similar test on some other languages – including SOV languages such as Basque and West Greenlandic and one VOS language (Western Cakchiquel) by investigating the distribution of the translations of the Greek name *Petrós* 'Peter', as proper names are fairly consistently rendered and easily recognized. As it turns out, even if the recall rate is sometimes significantly lower for these languages (that is, fewer words are identified in the translations), the gain made by widening the window, even to whole verses, is at most slightly above ten per cent of the occurrences found. This suggests that the influence of word order may be less than one

would think. In the following examples, I shall be using a word window with a maximum normalized distance of five words in each direction.

4. A first example: the definite article

Let us now see what happens when we start comparing the distribution of grammatical items cross-linguistically between Bible translations, starting out from a simple case: the definite article in NT Greek and English. The reason this case is simple is that since the English definite article is invariable and the word *the* has no other very frequent function,[5] we can simply see to what extent *the* is marked with the Strong number "3588", implying that it corresponds to some form of the Greek definite article. As we have already seen, the Greek article has almost twice the frequency of English *the.* The most prominent reason for this difference is probably that NT Greek relatively consistently uses the definite article also before proper names. In spite of this, the extent to which the two languages use definite articles in the same context is quite large; as it turns out, there are 7,719 cases of *the* marked by the Strong number "3588" in KJV, that is, 68 per cent of all occurrences of the English *the*.

Most translations that a typologist is interested in do not come equipped with Strong's Numbers and represent languages that the researcher does not have any proficiency in. Is it still possible to compare the distribution of grammatical items? The natural first choice is to try the word-alignment methods that have already been proposed in the literature on parallel corpora. Notice, however, that the goal here is slightly different: the main goal of word-alignment is to find out which word in one text is the most likely translation of a word in another. Here, we do not only want to say that the Greek definite article is the most likely counterpart to *the* in English; we also want to obtain a measure of how similar they are in their distribution and ultimately, in what respects they differ. Ideally, then, an automated analysis program should be able to tell us that the Greek and English definite articles differ in that the former is used before proper names and the latter is not.

What we can learn from the literature on word alignment (VARMA 2002, TIEDEMANN 2003) is that there is no single ideal algorithm for matching words in parallel texts. For the time being, I have chosen to use a measure referred to as "T-score" (FUNG & CHURCH 1994), which has the advantage of being relatively simple from a computational point of view. Basically, what a T-score is a measure of the association between two items – that is, a very high T-score means that it is highly unlikely that the items should show up in the way they do just by chance. The T-score is computed

[5] It was suggested to me that the construction exemplified by *the bigger the better* might be an exception. Indeed, the pattern "the more * the" gives back 49.5 million hits on Google, which may seem a lot, but typing in the word *the* by itself yields 9.4 billion hits, so the *the…the* construction is actually quite marginal. It is a bit complicated to find the exact frequency of the English construction in the Bible, but it may be of some interest to know that the corresponding German construction *je…desto* occurs exactly once in Luther's translation of the New Testament (Mark 7:36).

as shown in (1), where *A* and *B* are two types of corpus events, and *prob(A,B)* means 'the probability of joint occurrence of A and B' and *K* is the number of chunks into which the texts are divided. In my investigation, *K* is the total number of words in the English text, which is identical to the number of "word windows" investigated.

(1) $$T = \frac{prob(A,B) - prob(A) * prob(B)}{\sqrt{\frac{1}{K} * prob(A,B)}}$$

Suppose that we are comparing the definite articles in Greek and English. For each word *w* in the Greek text, *A* means that *w* is a definite article, *B* means that the English definite article occurs at least once in the "word window" of *w*, that is, the set of words in the English text whose normalized distance is less than the maximum we have determined. It is likely that in the end, we will want to combine T-score with other measures. In particular, the T-score of a combination of items does not tell us how often the items occur together, it just says something about the likelihood that their distribution is due to chance. It is obvious that the method is easiest to apply when the expression we are looking at has an invariant form. The English definite article happens to fulfil this condition. Table 1 shows the words in KJV that have the highest T-scores when compared to the Greek definite article (identified by its Strong number). A similar result can be obtained e.g. for the Afrikaans definite article *die* as shown in Table 2.

Table 1. Best results of comparisons between the Greek definite article (Strong number 3588) and words in the English King James' Version.

English	T-score
the	35.94
and	21.67
of	21.33

Table 2. Best results of comparisons between the Greek definite article (Strong number 3588) and words in the Afrikaans 1953 Bible translation.

Afrikaans	T-score
die 'the'	24.90
van 'of'	18.14
sy 'his'	11.03

Table 3. Best results of comparisons between the Greek definite article (Strong number 3588) and words in the French Louis Segond translation.

French	T-score
la	16.86
le	16.79
de	15.97
qui	15.16
les	13.58

In languages such as French, where the definite article has several different forms (*le*, *la*, *les*), depending on gender and number, it is not possible, by just comparing T-scores, to single out those forms from other common words such as the relative and interrogative pronoun *qui* or the preposition/possessive marker *de*. As can be seen from Table 3, at least one form of the definite article has a lower T-score than *qui* and *de*. Thus, for someone who does not know anything about French beforehand, it is not possible to identify definite articles by this simple method. One way out is to look for co-occurrences within one language. Different forms of a definite article are likely to be in complementary distribution with each other: we would not expect to find them closely together. A definite article and a relative pronoun, on the other hand, will often show up in the same noun phrase. If we thus want to know which of *le* and *qui* that belongs together with *la*, it is quite informative to know that *la* and *qui* in fact have a weak positive T-score (0.6) while the T-score for the co-occurrence of *la* and *le* is clearly on the negative side (–6.58).

5. A second example: future tense

Let us take another example of a grammatical phenomenon: grammatical markers of future time reference. The fact that New Testament Greek had an inflectional future might be expected to make it difficult to compare the future in Greek with a language such as English, where future time reference is only marked by periphrastic means – by auxiliaries such as *shall* and *will*. However, if we try to run a similar test as described in the preceding paragraph on the Greek future tense, that is, look for what words tend to show up most often in the same environments, the English auxiliaries *shall* and *will* come up consistently as the best candidates in most English Bible translations. Moreover, we can observe the historical development of these auxiliaries in their role as future markers, as shown in Table 4. In the earliest English Bible text available to me, the Wycliffe translation from the 14th century, the highest T-scores all belong to forms of the auxiliary *shall,* while *will* is not common enough to be visible in the statistics (it

was still only used in its original sense 'want'). In KJV and its more recent clones, *shall* is still dominant, but *will* is on its way up, with values that are about a third of those of *shall*. In those recent Bible translations that try to emulate contemporary English, *will* has taken over and *shall* has been reduced to a very insignificant position.

Table 4. T-scores for *shall* and *will* in representative English Bible translations, from comparison to Greek future tenses.

Wycliffe (14th century)		Tyndale (1525)		King James' Version (1611)		World English Bible (2000+)	
schal	24.89	*shall*	23.67	*shall*	29.18	*shall*	4.03
schalt	8.63	*shalbe*	13.79	*shalt*	8.95		
schulen	18.98	*shalt*	8.10				
shal	6.77						
shalt	2.38						
will	–	*will*	13.86	*will*	16.62	*will*	25.65

Couldn't we study this development just by looking at the frequencies of *shall* and *will* in the texts? Not quite, since the frequencies do not tell us anything about the functions of the auxiliaries. What we are looking at here is not how often *shall* and *will* are used but how often they are used as counterparts of the inflectional future tense in Greek, which we take as an example of a highly grammaticalized way of marking the future.

I have performed the same test on a number of languages and results are in accordance with expectations. Thus, the future marking auxiliaries in German and Scandinavian, which are less grammaticalized than the English ones, also show lower values. Periphrastic future markers identified in DAHL (1985) such as Afrikaans *sal*, Bulgarian *šte,* Indonesian *akan,* are readily picked out by the comparison with the Greek future. Obviously, this is not to say that Greek morphological categories have any fundamental role to play in the analysis, but they can be used to "bootstrap" the process. What this means is that once we have a preliminary identification of a number of future markers, we can go on to create a "map" of their common distribution, which will serve as a basis for the further search, in the same way as I did in the earlier investigation, using questionnaire data. So far I have just been exploring the possibilities – the results look promising but will have to be reported at a later point in time.

6. Conclusion

In this paper, I have discussed the possibility of using parallel corpora for cross-linguistic studies of grammatical categories. My own exploration of the potential of a parallel corpus based on Bible translations is yet in an initial stage, which explains the programmatic character of this paper. The examples I have chosen were intended as illustrations and do not yield any new insights about the categories in question. What I hope to have shown is that techniques similar to those used for word alignment of parallel corpora are also useful for comparing the distribution of grammatical phenomena across languages. Much remains to be done – the greatest challenge is to include morphological categories in the investigation. It remains to be seen how much can be done by an automatic analysis, and how much that will still necessitate manual analysis of a more traditional kind. But it is my hope that the methodology outlined here will prove fruitful and usable also for parallel corpora based on other texts than the Bible.

References

BYBEE, JOAN L. (1985): *Morphology: a study of the relation between meaning and form*. Amsterdam, Philadelphia: John Benjamins.

BYBEE, JOAN L. & DAHL, ÖSTEN (1989): The Creation of Tense and Aspect Systems in the Languages of the World. *Studies in Language* 13.1, 51-103.

BYBEE, JOAN L., PERKINS, REVERE D. & PAGLIUCA, WILLIAM (1994): *The evolution of grammar: tense, aspect, and modality in the languages of the world*. Chicago: Univ. of Chicago Press.

CYSOUW, MICHAEL, BIEMANN, CHRISTIAN & ONGYERTH, MATTHIAS (this issue): Using Strong's Numbers in the Bible to test an Automatic Alignment of Parallel texts.

DAHL, ÖSTEN (1985): *Tense and aspect systems*. Oxford: Blackwell.

FUNG, PASCALE & CHURCH, KENNETH WARD (1994): K-vec: a new approach for aligning parallel texts, in: *Proceedings of the 15th conference on Computational linguistics - Volume 2, Kyoto, Japan*.

TIEDEMANN, JÖRG (2003): Combining clues for word alignment, in: *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1, Budapest, Hungary*.

VARMA, NITIN (2002): *Identifying Word Translations in Parallel Corpora Using Measures of Association*. M.Sc. thesis, University of Minnesota.

VRIES, LOURENS DE (this issue): Some remarks on the use of Bible translations as parallel texts in linguistic research

Correspondence address

Östen Dahl
Department of Linguistics
Stockholm University
106 91 Stockholm
oesten@ling.su.se