

Advantages and disadvantages of using parallel texts in typological investigations¹

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—NEWMAYER'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 ...mightier **than** I...; 4:31. ...less **than** all the seeds...; 4:32 ...greater **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 **easier** for a camel to go...**than** for a rich man to enter...; 12:31 ...**greater 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, NEWMAYER 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 (*ἐξ* '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 NEWMAYER 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-ěr***.
 child-POSS3-DAT/ACC I.GEN to-POSS1SG-DAT take-CONV come-IMP2PL

- (7) Choctaw (Muskogean)

Isht *hus* *som* ***ɬla.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 (Tangkhu), Ossetic, Rhaeto-Romance, Romani (Kaldersh), Rumanian, Russian, Saami (Northern), Santali, Spanish, Swedish, Tabassaran, Tadjik, 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 (Guaymí), Paumari, Piro, Quechua (Imbabura), Shipibo, Yaneshá'	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	<i>they brought unto him all that were diseased</i>	5:6	<i>he ran and worshipped him</i>
2:03	<i>bringing one sick of the palsy</i>	6:33	<i>and ran afoot thither out of all cities</i>
6:27	<i>and commanded his head to be brought</i>	6:55	<i>And ran through that whole region round about</i>
6:28	<i>And brought his head in a charger</i>	9:15	<i>and running to him saluted him</i>
7:32	<i>And they bring unto him one that was deaf</i>	10:17	<i>there came one running, and kneeled to him</i>
8:22	<i>and they bring a blind man unto him</i>	15:36	<i>And one ran and filled a sponge full of vinegar</i>
9:17	<i>I have brought unto thee my son</i>		
9:19	<i>bring him unto me</i>		
9:20	<i>And they brought him unto him</i>		
10:13	<i>And they brought young children to him</i>		
11:02	<i>and bring him</i>		
11:07	<i>And they brought the colt to Jesus,</i>		
12:15	<i>bring me a penny</i>		
12:16	<i>And they brought it</i>		
15:01	<i>and carried him away</i>		
15:16	<i>And the soldiers led him away into the hall</i>		
15:20	<i>and led him out to crucify him</i>		
15:22	<i>And they bring him unto the place Golgotha</i>		

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.

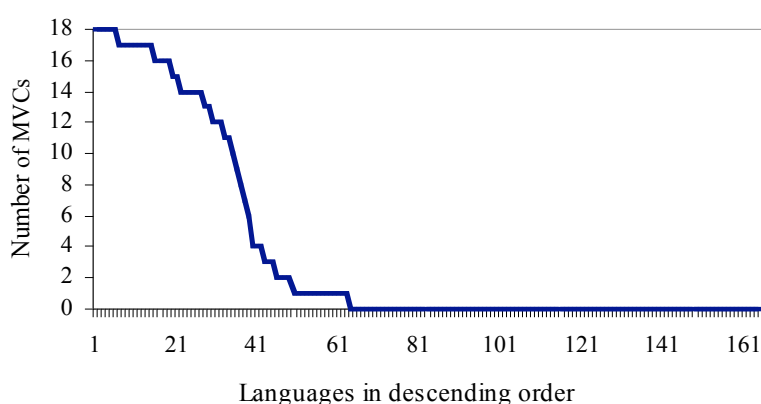
Table 4. Availability of MVC in the two domains.

		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).

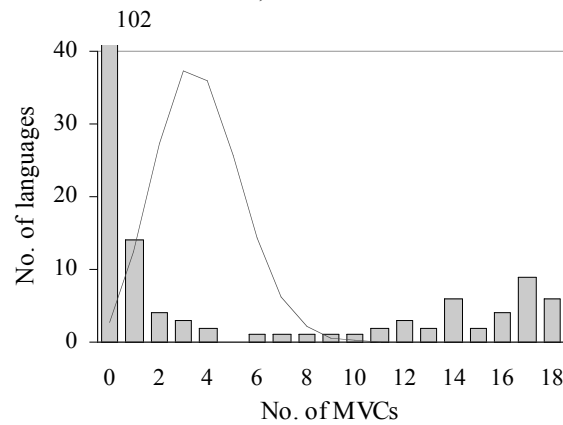


Table 5 gives the number of languages for each value. In brackets the differences to expected values are given. The correlation between MVCs in the two considered domains emerges more clearly when only the extremes are considered (the number of Intermediate cases are all close to the statistical expectation 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 under-represented (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 iiy-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-īm 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-nīl sare kval-īm-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 misionero”. 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 world's 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.
- NEWMAYER, 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.
- STEWART, OSAMUYEMEN 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