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