Sprachtypol. Univ. Forsch. (STUF), Berlin **60** (2007) 2, 172–181

Östen Dahl (Stockholm)

From questionnaires to parallel corpora in typology *

Abstract

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

1. Introduction

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

2. Questionnaires

If one finds an interesting example in language A, the natural question is to ask "How would this be expressed in languages B, C etc.?" If the answer cannot be found in a grammar, as is often the case, the obvious way of getting the answer is to ask a native speaker. A more systematic approach to this is to use a questionnaire, the most straightforward type of which is a translational questionnaire, containing a set of expressions, sentences or connected texts to be translated by a native speaker into the language under investigation. A well-constructed questionnaire covers a certain area of grammar in such a way that it

[*] I am grateful to the editors of this issue for many valuable comments, and to Jan Olov Persson for consultations on statistical problems.

gives information about the ways in which this part of the grammar is structured in the language in question by yielding a set of translational equivalents between the source language and the target language, and indirectly, between different target languages. The notion of translational equivalent should be understood in an operational and theory-independent sense: an expression α in one language is a translational equivalent of an expression β in another language if α is actually used (more than occasionally) as a translation of β by persons who are competent in both languages. How this behaviour should be interpreted is another matter. An obvious limitation of the questionnaire approach is that the choice of expressions to be translated has to be guided by the questionnaire constructor's understanding of the phenomena being studied – which may quite negatively influence the chances of making new discoveries.

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

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

As for the results of the investigation, I think it can be said that although they do not in any serious way contradict what was generally said in the literature at that time, the investigation contributed to sharpening the picture of what tense and aspect systems in human languages are like, especially in conjunction with the grammar-based typological investigations of verbal categories led by Joan Bybee (Bybee 1985; Bybee et al. 1994). At the same time, in spite of the rapid development of language typology, and although questionnaires are now a standard tool for typologists, I do not know of any investigations that have tried to apply the methodology I used. This probably has to do with the inherent difficulties in the method. A general problem is that a typological questionnaire investigation with a good coverage is quite costly. It takes a considerable time to develop a good questionnaire, and at the point where it is mature, you may already have used up most of your available informants as guinea-pigs. Almost unavoidably, the set of languages investigated will be a convenience sample, that is, the choice will depend more on the availability of bilingual and literate informants than on a principled sampling method. (The sample in Dahl 1985 did contain a fair number of non-European languages but was still quite heavily biased. Thus, 21 languages – that is almost a third of the sample – were Indo-European).

3. An alternative: parallel corpora

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

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

1. The languages into which the Bible has been translated wholly or partially are spread fairly evenly over the globe, making the creation of a relatively unbiased sample seem possible.
2. Many Bible translations are readily available for download from the Internet. (However, the set of freely downloadable Bibles, regrettably, looks rather like your typical convenience sample, with a heavy bias towards translations into European and a few major non-European languages.)
3. The Bible is really a collection of quite heterogeneous texts of different genres, including straightforward narratives and argumentative passages.
4. Even if the Bible (like virtually all parallel corpora) represents written language, there is a considerable amount of natural-sounding direct speech in it.
5. Bible texts are usually well prepared for use in parallel corpora, in that the partitioning into chapters and verses can serve as a substitute for sentence alignment. Strong's Numbers (see Cysouw et al., this issue), for the translations where they exist, can even provide word alignment.
6. At least in the case of the New Testament, versions of the original text (Greek[1]) with complete lexical and morphological markup are freely available.[2]

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

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

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

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

When I worked on the TMA questionnaires, I had the advantage that the verb forms were already marked up by experts on the respective languages. The fundamental problem of parallel corpora studies, that of alignment, thus did not exist. When comparing the distribution of grammatical items (morphemes, constructions etc.) in Bible translations, on the other hand, we do not in general have access to grammatically analyzed texts – with one important exception: the Greek original. We must therefore find a method to match or align the grammatical items across languages. This is not an easy task and it is obvious that before we can do anything similar to what I did with the TMA questionnaires a huge amount of work is needed.

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

To an astonishing extent, grammatical or functional words can be identified with high-frequency words – at least in the languages I have looked at, and I see no reason why it should not be the case universally. Thus, in the KJV New Testament, the most frequent word which is unequivocally lexical rather than grammatical is *God*, which has rank 23 and frequency 1,372. Now, if one tries to run a word-alignment algorithm on a Bible translation along the lines suggested in Cysouw et al. (this issue), it turns out that high-frequency words create special problems. The three most frequent words in the King James Version of the New Testament are *the* (11,036 occurrences), *and* (10,721 occurrences) and *of* (6,129 occurrences). In the Greek New Testament, one single word-form, *kai* 'and', occurs more often than the following three words on the ranking-list taken together – it is found 9,208 times in the text. If we instead consider the Strong's Numbers, which reflect lexical items rather than word-forms (with a few exceptions), we find that Strong's Number "3588", which represents the Greek definite article in its various forms,[3] occurs no less than 20,317 times, that is approximately 14.5 per cent of the whole text, and on average 2.5 times per Bible verse. Word-alignment procedures discussed in the literature often follow the principle of dividing up the texts into aligned chunks, and then compute the probability that a word w_1 in a source text co-occurs with a word w_2 in the target text in a chunk *c*. As noted above, the verse constitutes a natural unit in Bible texts, and it would seem natural to use it also in word alignment – this is also suggested to be feasible in Cysouw et al. (this issue). However, for high-frequency elements such as definite articles, which tend to occur several times in each verse, this does not seem to be a very good idea – the number of false combinations will simply be too large. This is a problem I shall return to below. But it is not only the high text frequency of grammatical items such as the definite article that creates problems for word-alignment but also their cross-linguistic variability. Thus we know that many languages lack definite articles altogether. If a high-frequency grammatical word in the source text does not correspond to anything at all in the target text and vice versa, this creates a considerable amount of noise (in the technical sense of that word) for the word-alignment procedure. In particular, if the target text contains a grammatical item not found in the source text, there is no way of identifying it from the source text alone. In a multilingual parallel corpus, however, this problem can possibly be solved if we study the cross-linguistic distribution of gram-types, such as definite articles. If we know where in a text grammatical items of different cross-linguistic types are likely to appear, we'll be able to assign high-frequency items to those types before starting to align lexical words. Thus the study of the cross-linguistic patterns in the distribution of grammatical items in parallel corpora is needed for the understanding of cross-linguistic gram-types and for the word-alignment process in general.

As I suggested in the preceding paragraph, the verse may be too large a unit when studying the distribution of grammatical items in Bible texts. I would suggest that the best solution is not to try and divide up verses in smaller chunks on the basis of punctuation or other signals. Rather, one should use a moving "word window", which means that for a given word in text A we consider the words that are at a distance of no more than *n* words from the corresponding position in text B, for some suitable value of *n*. An easy way to define the position of a word in the Bible text is by identifying the verse where it occurs and its position (counted in numbers of words) from the beginning of that verse. When

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

comparing different Bible texts, the problem arises that the length of verses will not always be the same. This can be circumvented by a process of normalization: a verse is treated as if had the same length as in the Greek original and the positions of words in translations are recomputed accordingly. In this way, each word will have a number that identifies the most probable counterpart in the original text. The existence of Strong-numbered translations makes it possible to study how words in translations are distributed relative to the source words. In eight translations representing six European languages (English, Dutch, French, German, Portuguese, Russian), I found that of the words in the Greek texts that were assigned Strong's Numbers in the translations at most a few per cent were found at a (normalized) distance of more than five words from the original. Since the languages where translations with Strong's Numbers are available are a rather bad sample from the typological point of view, I have also performed a similar test on some other languages – including SOV languages such as Basque and West Greenlandic and one VOS language (Western Cakchiquel) by investigating the distribution of the translations of the Greek name *Petrós* 'Peter', as proper names are fairly consistently rendered and easily recognized. As it turns out, even if the recall rate is sometimes significantly lower for these languages (that is, fewer words are identified in the translations), the gain made by widening the window, even to whole verses, is at most slightly above ten per cent of the occurrences found. This suggests that the influence of word order may be less than one would think. In the following examples, I shall be using a word window with a maximum normalized distance of five words in each direction.

4. A first example: the definite article

Let us now see what happens when we start comparing the distribution of grammatical items cross-linguistically between Bible translations, starting out from a simple case: the definite article in NT Greek and English. The reason this case is simple is that since the English definite article is invariable and the word *the* has no other very frequent function,[4] we can simply see to what extent *the* is marked with the Strong's Number "3588", implying that it corresponds to some form of the Greek definite article. As we have already seen, the Greek article has almost twice the frequency of English *the.* The most prominent reason for this difference is probably that NT Greek relatively consistently uses the definite article also before proper names. In spite of this, the extent to which the two languages use definite articles in the same context is quite large; as it turns out, there are 7,719 cases of *the* marked by the Strong's Number "3588" in KJV, that is, 68 per cent of all occurrences of the English *the*.

Most translations that a typologist is interested in do not come equipped with Strong's Numbers and represent languages that the researcher does not have any proficiency in. Is it still possible to compare the distribution of grammatical items? The natural first choice is to try the word-alignment methods that have already been proposed in the literature on

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

parallel corpora. Notice, however, that the goal here is slightly different: the main goal of word-alignment is to find out which word in one text is the most likely translation of a word in another. Here, we do not only want to say that the Greek definite article is the most likely counterpart to *the* in English; we also want to obtain a measure of how similar they are in their distribution and ultimately, in what respects they differ. Ideally, then, an automated analysis program should be able to tell us that the Greek and English definite articles differ in that the former is used before proper names and the latter is not.

What we can learn from the literature on word alignment (Varma 2002, Tiedemann 2003) is that there is no single ideal algorithm for matching words in parallel texts. For the time being, I have chosen to use a measure referred to as "T-score" (Fung & Church 1994), which has the advantage of being relatively simple from a computational point of view. Basically, what a T-score is a measure of the association between two items – that is, a very high T-score means that it is highly unlikely that the items should show up in the way they do just by chance. The T-score is computed as shown in (1), where *A* and *B* are two types of corpus events, and *prob(A,B)* means 'the probability of joint occurrence of A and B' and *K* is the number of chunks into which the texts are divided. In my investigation, *K* is the total number of words in the English text, which is identical to the number of "word windows" investigated.

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

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

English	T-score
the	35.94
and	21.67
of	21.33

Table 1: Best results of comparisons between the Greek definite article (Strong's Number "3588") and words in the English King James' Version

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

Table 2: Best results of comparisons between the Greek definite article (Strong’s Number “3588”) and words in the Afrikaans 1953 Bible translation

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

Table 3: Best results of comparisons between the Greek definite article (Strong’s Number “3588”) and words in the French Louis Segond translation

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

5. A second example: future tense

Let us take another example of a grammatical phenomenon: grammatical markers of future time reference. The fact that New Testament Greek had an inflectional future might be expected to make it difficult to compare the future in Greek with a language such as English, where future time reference is only marked by periphrastic means – by auxiliaries such as *shall* and *will*. However, if we try to run a similar test as described in the preceding paragraph on the Greek future tense, that is, look for what words tend to show up most often in the same environments, the English auxiliaries *shall* and *will* come up consistently as the best candidates in most English Bible translations. Moreover, we can observe the historical development of these auxiliaries in their role as future markers, as shown in Table 4. In the earliest English Bible text available to me, the Wycliffe translation from the 14th century, the highest T-scores all belong to forms of the auxiliary *shall,* while *will* is not common enough to be visible in the statistics (it was still only used in its original sense 'want'). In KJV and its more recent clones, *shall* is still dominant, but *will* is on its way up, with values that are about a third of those of *shall*. In those recent Bible translations that try to emulate contemporary English, *will* has taken over and *shall* has been reduced to a very insignificant position.

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

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

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

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

6. Conclusion

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

References

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

Bybee, Joan L. & Dahl, Östen (1989): The creation of tense and aspect systems in the languages of the world, in: *Studies in Language* 13.1, 51–103.

Bybee, Joan L., Perkins, Revere D. & Pagliuca, William (1994): *The evolution of grammar: tense, aspect, and modality in the languages of the world*. Chicago: Univ. of Chicago Press.

Cysouw, Michael, Biemann, Christian & Ongyerth, Matthias (this issue): Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts.

Dahl, Östen (1985): *Tense and aspect systems*. Oxford: Blackwell.

Fung, Pascale & Church, Kenneth Ward (1994): K-vec: a new approach for aligning parallel texts, in: *Proceedings of the 15th conference on Computational linguistics. Volume 2, Kyoto, Japan*.

Tiedemann, Jörg (2003): Combining clues for word alignment, in: *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics. Volume 1, Budapest, Hungary*.

Varma, Nitin (2002): Identifying Word translations in parallel corpora using measures of association. M.Sc. thesis, University of Minnesota.

Vries, Lourens de (this issue): Some remarks on the use of Bible translations as parallel texts in linguistic research.

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