

P6

**Algorithmic corpus-based approaches to the typological
comparison of complex sentences**

**Michael Cysouw
Uwe Quasthoff**

1 Allgemeine Angaben

Neuantrag auf Gewährung einer Sachbeihilfe im Rahmen der DFG-Forschergruppe
"Performanzbasierte Analyse komplexer Sätze in sprachtypologischer Perspektive"

1.1 Antragsteller

Dr. Michael Cysouw

Forschungsgruppenleiter (E15)
geb.: 18.06.1970, niederländisch

Forschungsgruppe 'Quantitativer Sprachvergleich'
Fakultät für Sprach- und Literaturwissenschaften
Ludwig Maximilians Universität München
Geschwister Scholl Platz 1
80539 München

Tel.: +49 89 2180 6123
E-mail: cysouw@lmu.de
Web: <http://web.me.com/cysouw/>

Privatadresse:
Auerfeldstrasse 20
81541 München
Tel.: +49 89 4805 8805

Prof. Dr. Uwe Quasthoff

apl. Professor
geb.: 17.09.1956, deutsch

Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
PF 100920
04009 Leipzig

Tel.: +49 341 97 32233
E-mail: quasthoff@informatik.uni-leipzig.de
Web: http://www.asv.informatik.uni-leipzig.de/staff/Uwe_Quasthoff

Privatadresse:
Hermann-Keller-Straße 35A
04158 Leipzig
Tel.: +49 341 5213885

1.2 Thema

Algorithmic corpus-based approaches to the typological comparison of complex sentences

Algorithmische Korpus-basierte Ansätze zum typologischen Vergleich komplexer Sätze

1.3 Fachgebiet und Arbeitsrichtung

104-03: Sprachwissenschaften/Typologie, Außereuropäische Sprachen

1.4 Voraussichtliche Gesamtdauer

72 Monate (6 Jahre)

1.5 Antragszeitraum

Beantragter Förderungszeitraum: 36 Monate (3 Jahre)
Gewünschter Beginn der Förderung: 01.04.2011

1.6 Englische Zusammenfassung

There is an extensive body of research available that uses corpora to investigate the structure of individual languages. However, there are not many studies on quantitative, corpus-based investigations of a world-wide typological nature. This project will develop quantitative corpus-based methods for large-scale linguistic comparison. To reach this goal, we propose that it suffices to obtain a good approximation of the structure of each individual language using the same algorithmic procedures for all languages alike. The goals of this project are threefold. First, we will prepare corpora of lesser-studied languages for typological comparisons. Because of the limited amount of research on these languages, these corpora will mainly be unannotated corpora. To be able to investigate unannotated corpora, we will also prepare a smaller amount of parallel corpora as a starting point for the automatic analysis. Second, this project will use existing algorithms and develop new algorithms to add (approximate) linguistic annotations and extract relevant statistics from the corpora, allowing for the automatic assessment of typological parameters concerning complex sentences. Finally, the main intrinsic goal of this project (to be pursued in the second phase of the Forschergruppe) is to investigate how much linguistic knowledge of a language is needed to establish a particular typological parameter.

1.7 Deutsche Zusammenfassung

Es gibt eine Vielzahl von Forschungsarbeiten, in denen linguistische Korpora verwendet werden, um die Strukturen einzelner Sprachen zu untersuchen; dagegen gibt es nur sehr wenige Studien, in denen linguistische Strukturen sprachübergreifend auf korpuslinguistischer Grundlage untersucht werden. Dieses Projekt widmet sich der Entwicklung quantitativer und korpusbasierter Methoden zur Analyse sprachlicher Strukturen aus typologischer bzw. sprachvergleichender Perspektive. Dabei gehen wir davon aus, dass sich eine gute Annäherung an die Strukturen einzelner Sprachen mit Hilfe von generellen algorithmischen Verfahren erreichen lässt. Die Ziele des Projekts lassen sich in drei Punkten zusammenfassen: Ersten werden wir Korpora zu wenig erforschten Sprachen mit computerlinguistischen Verfahren so weit aufarbeiten, dass sie für einen typologischen Sprachvergleich zur Verfügung stehen. Da die so aufgearbeiteten Korpora nicht annotiert sind, werden wir ergänzend mit parallelen Korpora arbeiten, die für uns einen Ausgangspunkt bilden, um die nicht-annotierten Korpora mit automatischen Verfahren zu untersuchen. Zweitens wird dieses Projekt bestehende Algorithmen benutzen und neue Algorithmen entwickeln, um die von uns erstellten Korpora zu annotieren und um einschlägige Statistiken für die automatische Bestimmung typologischer Parameter komplexer Sätze aus den Korpora zu extrahieren. Schließlich soll in der zweiten Projektphase der Forschergruppe untersucht werden, wie viel sprachliches Wissen zu einzelnen Sprachen erforderlich ist, um einen typologischen Parameter zu bestimmen.

2 Stand der Forschung, eigene Vorarbeiten

2.1 Stand der Forschung

2.1.1 Corpus-based language comparison

The amount of textual data of the world's languages is currently rising at an incredible rate. There is an extensive body of research available that uses such corpora to investigate the structure of individual languages. However, there are not many quantitative investigations of a world-wide typological nature using corpora. There is some relevant work using texts to compare reference tracking between languages (Givón 1983, Myhill 1992, Bickel 2003) and

some work using basic text counts to assess the morphological typology of a sample of languages (Greenberg 1960, Altmann & Lehfeldt 1973, Cysouw 2007b). Some further studies use direct translational equivalents to compare languages (Fenk-Oczlon 1999, Wälchli 2005). Still, the large majority of corpus research in today's linguistics is monolingual research, with the goal to improve the understanding of the structure of individual languages.

In contrast, this project will develop quantitative corpus-based methods for large-scale world-wide linguistic comparison. In close collaboration with the other projects of the Forschergruppe, we will investigate concrete linguistic questions related to complex sentences using corpora for as many languages as possible. We will also prepare corpora for future typological research in this Forschergruppe and beyond. We expect that the usage of corpora (of various kinds) will revolutionize linguistic comparison as it offers the possibility to go beyond simple categorical classifications of languages into types. The belief that language variation is continuous, and not categorical, is widely accepted in linguistic typology (and linguistics in general), but it has proved very difficult to turn this assumption into practice and produce actual continuous measurements of linguistic variation across a wide range of languages from different parts of the world.

In this project, we will combine various research traditions. First, we will use the web as a corpus (2.1.2), and combine this with inducing annotations by using monolingual unsupervised language analysis (2.1.3). Second, we will use parallel texts to obtain more information about the structure of unanalyzed texts (2.1.4).

2.1.2 Web as a corpus

Currently, a large body of textual information is becoming available through linguistically-informed collections of texts, both in the context of the extensive efforts to document endangered languages and through the widespread development of richly annotated corpora for major languages. However, even more impressive is the amount of text produced by the steadily rising number of people using the internet in their own native language. More and more linguists are starting to use the 'web as a corpus' (cf. Baroni & Bernardini 2004 as an early example), though most of this research is directed towards the already well-studied top 50 languages in the world. Only a relatively minor amount of research attempts to use the web as source for the other thousands of languages (cf. Scannell 2007 for an example). Scannell's collection already shows that it is relatively easy to compile a reasonable amount of language material for hundreds of languages.

Given such a wealth of data, it seems reasonable to use this resource for more sophisticated linguistic comparison. The reason that such work has not yet become increasingly widespread is that there are obvious difficulties in using web data. For example, the data is available in practical orthographies, and judging from the orthographies of English and French that might seem to be a major problem. Luckily, the practical orthographies of most languages in the world are not as idiosyncratic as those of French and English, so this problem is actually less pressing than sometimes thought (although one of course still has to be careful in treating the orthography of such sources). Also, the texts from the web are all written language, and much of it is quite formal. Ideally, such written-style web corpora would be amended with sizable selections of exchanges in internet-fora, which would represent a much freer style of language use. Currently there are not yet enough such fora in lesser described languages, but that is surely a kind of data that is destined to grow in the near future.

The remaining and most important impediment to the widespread use of these data in language comparison is that all the data are monolingual, and a researcher would traditionally need quite a large amount of in-depth knowledge of each language to be able to use such data. To still be able to use monolingual data in large-scale language comparison without needing intimate personal knowledge of hundreds of languages (which would be practically impossible), we will use various automatic approaches to try and prepare linguistic annotations: first by using purely monolingual unsupervised analysis (2.1.2), and second by using

massively parallel texts combined with text mining approaches to learn more about the structure of individual languages (2.1.3). Both these approaches are rather rough from a purely linguistic perspective, and neither of them will improve on careful manual language-specific analysis by linguists. However, we expect that the rough approximations to the structure of the languages will be good enough to improve typological comparison—which is a rather rough kind of linguistic analysis to begin with.

2.1.3 Monolingual unsupervised language analysis

There is a large field in natural language processing in which structure is automatically extracted from texts. Such approaches are typically monolingual (i.e. they only use an unannotated corpus of texts in one language, without further knowledge about the meaning or structure) and ideally unsupervised (i.e. they work without help or correction of a human). Such approaches reach a relatively good approximation to the structure of a language, though they are far from perfect.

There are various aspects of structure induction that we will use in our project: stemming/morpheme extraction, parts-of-speech tagging, chunk identification and the extraction of constituent structure. First, there is a long tradition in natural language processing to automatically extract the stem of inflected forms ('stemming') and perform automatic morpheme segmentation algorithm (these two goals are of course strongly linked, cf. Porter 1980, Creutz & Lagus 2002, 2005; Bordag 2006, 2007, 2008). Typically, these approaches have difficulty with the precise boundary of the morphemes, and sometimes they will miss one or add one that is not there. However, the rough number of affixes and the identification of the 'same' affix in different situations works rather well.

Second, so-called 'POS-tagging' can add part-of-speech tags to any word in the corpus. At the moment, good POS-taggers are available only for about 30 languages, because most POS-taggers need extensive manual training. To overcome the manual training phase, unsupervised POS-tagging (UnsuPOS, Biemann 2006, 2007) uses clustering methods to identify 'word cluster classes'. These classes are characterized by containing words having similar contexts of high frequency words ('stop words'). The number of classes is usually higher than the traditional number of part-of-speech classes. For instance, German nouns are normally further distinguished in classes with respect to their gender. In the case of proper names, first names and last names are distinguished as well as place names. Currently, UnsuPOS-tagging requires large corpora for training. Here there is a clear need for improvement in the UnsuPOS algorithms.

Finally, and crucially to the question of complex sentences, word classes can be used to induce chunks, and even further possibilities are available to establish constituent structure (cf. Klein & Manning 2005). However, these approaches are still very new and in need of improvement.

2.1.4 Linguistic comparison through parallel texts

To enhance the automatic analysis of unannotated corpora we will use parallel texts. In general, parallel texts are just texts with a translation into another language. Because the same content is expressed in two different linguistic structures, such a combination of text with translation offers the possibility to map knowledge about the structure available in one language onto the structure in the other language. Such a structural mapping is of course never a simple one-to-one mapping. However, there is normally enough parallelism to allow for the recognition of comparable structures across languages. There are many different flavors of parallelism (depending on how strict the parallelism is, and how clearly it is annotated), but we will focus on massively parallel texts, i.e. the same text (mostly on a paragraph-by-paragraph base) translated into very many different languages.

There is some research on the usage of parallel texts for language comparison (cf. Wächli 2007, 2009, 2010; Stolz 2007), but for these studies the analysis of all languages was still performed manually. Although this allows for most control about the annotation of the individual languages, it is quite a laborious approach. Yet, there is a wide field of study in natural language processing that induces more detailed parallelism from a higher level parallelism. Typically, this research aligns words on the basis of a given parallelism on the sentence level, but in general it is useful for any finer grained annotation, once a higher level is available (cf. Och & Ney 2003; Tiedemann 2003). This approach is mainly used for language to language translation, so the main interest of the field is in bilingual parallel corpora as a resource to help improving automatic translation. However, there is also some research on the use of parallel texts for the unsupervised induction of linguistic knowledge (cf. Dahl 2007; Samardžić & Merlo 2010).

In the context of the European Union there is interest in a more widespread application of parallel corpora to help preparing the many translations necessary in this multilingual governmental environment. Consequently, the Europarl corpus (<http://www.statmt.org/europarl/>) is probably the most well-known massively parallel corpus, consisting of translations of the European Parliament in all official EU languages. However, there are also various other sources of massively parallel texts as collected by Jörg Tiedemann (Tiedemann & Nygard 2004; Tiedemann 2009) in his OPUS project (<http://urd.let.rug.nl/tiedeman/OPUS/>). These sources are not used for typological comparison, although they are ready, just waiting to be used. The major problem for typological comparison with the parallel corpora from OPUS is that there are only translations available for well-studied languages. For the remaining few thousand lesser-described languages we will have to rely on rather limited resources as will be compiled in this project.

2.1.5 Language comparison through treebanks

Just recently some work has been started to use treebanks for typological comparison. Liu (2010) for example compares available treebanks to establish word order typologies from corpora. This is a fascinating prospect, but such work depends on available treebanks, and it is quite laborious to build treebanks from scratch. This implies that the usage of treebanks is not yet feasible for very large typological samples of languages, though it remains a tempting prospect on the horizon. From recent work by Tiedemann & Kotzé (2009) it appears that this future is not very far off, as they show that treebanks can to some extent be induced from parallel corpora. It is precisely the goal of this project to compile parallel corpora for many minority languages, so this line of research can be further developed in the future.

2.2 Eigene Vorarbeiten

2.2.1 Michael Cysouw

Michael Cysouw has worked extensively on the quantitative typological comparison of the world's languages, both dealing with concrete linguistic diversity (Cysouw 2003b, 2007a, i.a.) and with a strongly methodological interest (Cysouw 2002, 2003a, 2005, 2007c, i.a.). His research activities include various projects in which parallel texts, corpora and translational equivalents in general were used to compare languages.

In collaboration with Bernhard Wächli he has been investigating the potential of parallel texts for linguistic typology. A research paper on the typological structure of motion verbs in parallel texts showed the practical application of this approach by dealing with a large number of examples from many languages (Wächli & Cysouw 2010). This collaboration has also resulted in an edited volume including practical examples of such research, combined with the discussion of methodological problems surrounding the use of parallel texts (Cysouw & Wächli 2007). The concrete language comparisons reported on in that volume (concerning

prepositions, multi-verb constructions and demonstratives) were analyzed in traditional linguistic manner, involving much manual labor. In reaction, Cysouw, in collaboration with Quasthoff's students from the department of natural language processing (Cysouw, Biemann & Ongyerth 2007) developed a method to automatically pre-process parallel texts to facilitate linguistic analysis. This is a practical implementation of language-independent methods to align parallel texts, making it easier to quickly compare languages without assuming any knowledge about their structure. Further, the combination of a detailed corpus study of an individual language with a broad typological survey was used in the investigation of content interrogatives (Cysouw 2007b).

Translational equivalents across different languages were also the underlying data for the investigation of case functions in Tsezic languages (Cysouw & Forker 2009). Although for that paper we did not use the same kind of parallel texts as we are planning to use in the current project, it highlighted the potential that translational equivalents offer for the comparative investigation of languages. In particular, we were able to derive a historical reconstruction of the Tsezic languages from the detailed variation of the usage of case markers that can be extracted from cross-linguistic parallelism.

2.2.2 Uwe Quasthoff

Uwe Quasthoff has coordinated the corpus project „Projekt Deutscher Wortschatz“ for more than 15 years. During these years, many aspects of corpora production were explored and built into a corpus production process. This includes web crawling, text conversion, esp. HTML-stripping, language identification, sentence separation, pattern-based text cleaning, co-occurrence analysis in very large corpora, multiword detection, named entity recognition, unsupervised morphological analysis, unsupervised parsing, and several further aspects of corpus statistics (Biemann et al. 2004, Quasthoff et al. 2006, Biemann et al. 2008, Hänig, Bordag & Quasthoff 2008). These techniques were found to be (nearly) language independent and were therefore developed further, resulting in corpora of currently about 50 languages (searchable at <http://corpora.informatik.uni-leipzig.de/>, Biemann et al. 2007, Quasthoff 2010). The research resulted in 3 dissertations and about 20 diploma or master theses. The current corpus processing pipeline will be used and greatly extended in the context of the Forschergruppe to produce more corpora. Using this rich data source, Quasthoff has published widely on analytical aspects like text mining (Heyer, Quasthoff & Wittig 2008), collocations (Quasthoff & Schmidt 2010), and similarity measures for corpus data (Biemann & Quasthoff 2007).

3 Ziele und Arbeitsprogramm

3.1 Ziele

3.1.1 The preparation of corpora for typological comparison

There are many corpora available for the world's major languages (even massively parallel, like Europarl). However, to be able to also investigate less well-studied languages, this project will collect and prepare textual resources in many of the world's lesser-studied languages. The inclusion of such lesser-studied languages is essential to prevent Eurocentric biases in cross-linguistic research. In this project, we will prepare (i) unannotated monolingual corpora using data from the internet, and (ii) a massively multilingual parallel corpus. Through the combination of these two kinds of corpora we will be able to automatically extract typological characteristics of languages—and in the context of the Forschergruppe specifically linguistic structures of complex sentences.

More specifically, the large monolingual corpora will be used to extract highly detailed statistics about the structure of the various languages. However, because of their monolingual nature it will be difficult to obtain results about linguistically more intricate characteristics, like many aspects of the structure of complex sentences. For that reason we will use the massively multilingual parallel corpus. By using this corpus, we will be able to obtain a much better insight into the relevant constructions as used in each language. This knowledge can then be used to bootstrap the extraction of information from the larger unannotated corpora. All corpora prepared in this project will be made publicly available at the end of the project, though copyright considerations—which we will do our best to solve or circumvent—might reduce public access.

3.1.2 The development of algorithms to detect linguistic structure

In cooperation with the other projects of the Forschergruppe, this project will develop algorithms to extract relevant statistics from the corpora, allowing for the automatic assessment of typological parameters. The algorithms and resulting statistics to be developed will be used to investigate the following aspects (among others):

- Recognition and classification of complex sentences, cf. P7 (Bickel/Gast)
- Position and length of embedded clauses, cf. P5 (Diessel) and P7 (Bickel/Gast)
- Headedness and symmetricity of embedded clauses, cf. P5 (Diessel)
- Argument-explicitness in embedded clauses, cf. P1 (Haspelmath/Michaelis)
- ‘Finiteness’ of predicates in embedded clauses, cf. P1 (Haspelmath/Michaelis), P3 (Gast/Schäfer) and P7 (Bickel/Gast)
- Information structure of complex sentences, cf. P4 (Lühr/Zeifelder)
- Polyfunctionality of complementation/subordination markers, cf. P1 (Haspelmath/Michaelis)

Summarizing our approach, we will pursue the following levels of analysis to extract information out of corpora:

- Using unsupervised monolingual analysis, we will infer a first rough approximation to the structure of the corpora (e.g. tagging of constituents, parts of speech, stems/morphemes, co-occurrences).
- we will use our own massively parallel corpus (augmented with manual annotation of exemplars) to identify relevant language-specific constructions in the world’s languages.
- The identified language-specific constructions will then be used to further investigate the monolingual corpora using data-mining techniques to find more examples of the same kind.
- Given some basic structural analysis of a collection of relevant complex sentences, we will extract continuous typological parameters for language comparison.

3.2 Arbeitsprogramm

3.2.1 The preparation of corpora for typological comparison

We will collect two different kinds of corpora: (i) a massively parallel corpus and (ii) a large collection of unannotated monolingual corpora.

3.2.1.1 Preparing the massively parallel corpus

We will prepare a small massively multilingual parallel corpus using the Universal Declaration of Human Rights and various religious texts. The only reason to choose these texts is that

translations of these same texts are relatively easy to obtain in very many languages, including such languages for which normally only very few resources are available. We will use the following sources:

- Universal Declaration of Human Rights: electronically available translated into almost 400 languages (<http://unicode.org/udhr/>)
- Watchtower: Christian pamphlets translated into about 350 languages (not counting sign languages (<http://watchtower.org/languages.htm>))
- Bibles: we currently know about 100 free electronically available translations of the bible, from a variety of sources. We expect more to be available on further investigation. (Because there are various colleagues outside of this Forschergruppe that work on preparing Bible translations for automatic access, we plan to organize a workshop to coordinate these efforts, cf. Section 4.3 and 5.2)

Although these sources are already rather strictly parallel in their underlying structure, some sizable work is needed to clean up the sources and prepare the basic parallelism as detailed as possible without assuming detailed linguistic knowledge. The problem is that all sources specify their parallelism on the paragraph level, which has to be broken down to a sentence, or even clause level to be really useful for linguistic comparison (e.g. using UPLUG <http://sourceforge.net/projects/uplug/>). The basic parallelism will be purely annotated on the basis of the delimiters as used in the original sources, i.e. the text is chunked by spaces and punctuation marks, and these chunks are used for aligning the texts. The basic alignment consists of an explicit statement concerning the chunks that are likely to be equivalent in the different languages. More detailed attempts at syntactic parsing and morphological segmentation, and the parallel alignment of these fine-grained chunks, will be part of this project (see below), but will not be assumed for the establishment of the basic parallelism. All parallelisms will be annotated using the PAULA format (<http://www.sfb632.uni-potsdam.de/~d1/paula/doc>) or the slightly more widespread Corpus Encoding Standard (<http://www.cs.vassar.edu/CES/>). Both are stand-off annotation schemes which allow for flexible annotations of parallelism (and other detailed linguistic information). See also P7 (Gast/Bickel, section 3.2.2) for a more detailed discussion of stand-off annotation.

It should be clear that this massively parallel corpus is far from an ideal linguistic resource. All sources are rather strict translations, possibly inducing significant influences of written style and ‘translationese’ into the texts. However, there are various reasons that we still want to use these sources for linguistic comparison. First, note that we do not plan to use these sources to obtain a better understanding of the structure of each individual language. We will be using these sources for typological language comparison, i.e. to get an approximate impression of the relative similarity between languages. For this in itself rather coarse-grained endeavor we think also less than ideal resources might be sufficient. Secondly, the religious texts—especially the Watchtower pamphlet, but also most of the Bible translations—are prepared with the explicit goal to convince people. These texts are written to be understood and accepted by native speakers, and thus are not as artificial as traditional European bible translations. The situation is clearly different with the Universal Declaration of Human Rights, which has more of an official status. Thirdly, we will use the parallel corpus only as a head start into the investigation of the monolingual unannotated corpora. We expect these to be much less influenced by translationese (though there is of course still a written style bias). Finally, these massively parallel corpora do provide such an enormous amount of information about the world’s languages that it would be a waste to not at least try to use them for comparative linguistic purposes.

In general, there is no reason to restrict ourselves to massively parallel corpora (i.e. the same text available in very many translations). For the lesser-studied languages which are the core objective of our project there are also many—currently underused—pairwise parallel corpora (i.e. original texts with translations). These are almost as informative as massively

parallel texts. The main advantage of massively parallel texts is that it is assured that every example is available in all languages, simplifying comparison. Original texts with translations need extra effort to establish comparable contexts across languages. It is possible, though tedious, to establish such parallelism by using the translations as mediator. Further complicating matters, such pairwise parallel texts are mostly available in printed form, necessitating full-fledged digitization (and consequently higher funding). We will only pursue this approach (possible in the second phase of the Forschergruppe) when the current project has successfully shown that such a task is worthwhile for the further advancement of language comparison.

3.2.1.2 Preparing unannotated monolingual corpora

To collect unannotated monolingual corpora we will use the texts from our own parallel corpus, newspapers, and Wikipedia as the basis for language-identification, and then use web-crawling based on these data to expand the size and variability of the resources. The potential number of languages for these different text genres is as follows:

- Newspaper texts: About 120 languages (<http://www.abyznewslinks.com/>)
- Wikipedia texts: There are Wikipedias with more than 100 articles for about 200 languages.
- Web text: About 300 languages using a method like bootcat (<http://bootcat.sslmit.unibo.it/>)

The Wikipedia texts are available as dumps in a XML format identical for all languages; all other texts are collected using a web crawler. The processing of these texts consists of the following steps, which are already implemented in a processing pipeline:

- Character set conversion: All texts are converted into Unicode (UTF8)
- HTML/XML-stripping: This procedure extracts plain text out of the web pages
- Sentence separation: Text is separated into parts ('sentences') using source-specific delimiters (mostly punctuation marks and line breaks)
- Language cleaning: Sentences not of the designated language are removed
- Pattern-based cleaning: Only regular language expressions remain in the corpus, i.e. relicts from tables, lists, source code etc. are removed

The result of this process is a list of well-formed sentences. To allow for the public distribution of the data at the end of the project, this list of sentences is mixed into random order. With this approach we overcome copyright restrictions and can thus freely distribute the corpora. For internal usage within the Forschergruppe it will of course be possible to access the context of each sentence, though we will very probably not be able to make this information publicly available.

Even though these monolingual corpora suffer much less of 'translationese' compared to the parallel texts, it should be noted that the resulting corpora still have a strong written bias and might also very well represent a special "internet" lect of the language in question, and will thus not be representative of the speech of day-to-day interaction on the street, or of the style of telling traditional stories. However, in the study of language diversity, each of these styles or lects is interesting in its own respect. Also, we expect this 'written internet' lect to be structurally similar enough to other lects of each individual language to make typological classifications possible, i.e. the typological type of the 'written internet' lect will be normally closer to other lects of the same language than to written internet lects of different languages. This hypothesis has of course to be verified by a few case studies in which we have access to other corpora. Finally, for all languages studied here there will of course be individual cases of 'translated' constructions in our corpora, but we do not expect such influences to predominantly permeate our corpora.

3.2.2 The development of algorithms to detect linguistic structure

This part is the central part of the current project. Basically it consists of four steps of algorithmic development:

1. The unsupervised annotation of the corpora using purely monolingual algorithms
2. The unsupervised annotation of the corpora using the implicit knowledge encoded in the translations as available in the parallel texts
3. The extraction of language-specific characteristics of specific types of complex sentences. Basically, we will use manually selected examples in the parallel texts and identify possible encoding structures in each language through pattern matching, then use this knowledge to identify further cases in the monolingual corpora
4. Given a set of automatically identified exemplars of a language-specific construction, these will be statistically investigated to establish typological parameters

3.2.2.1 Unsupervised monolingual annotation

There is ample research on unsupervised monolingual annotation of corpora. Various such automatic procedures will be used to annotate the monolingual corpora collected in the project with some basic grammatical structure. It is important to realize that we do not claim that such a quantitative approximation will be identical to traditional linguistic concepts. Statistical algorithms tend to produce slightly different insights into the structure of language. However, such statistical insights are often correlated with linguistic notions, so, in a sense, this part of the project is doing language typology “by proxy”. Further, using the same (mathematical) method for structure induction on all corpora will produce a very consistent measure of morphemic structure, if only because it guarantees that the same criteria are used for all languages—something that is difficult to obtain in the traditional typological approach of perusing reference grammars. Such quantitative indices will be of great help for language comparison, and they enhance the linguistic understanding of structural variability. An intentional side-effect of the project is that insights about the world-wide variability of languages should serve as useful feedback to the development of unsupervised algorithms for structure discovery.

It is straightforward to use an existing unsupervised stemming procedure and morpheme segmentation algorithm and compare languages based on the morphemic structure attested, allowing for quantitative approximations to notions like finiteness, a central concept in the analysis of complex sentences.

Using language-independent algorithms for classifying word forms into groups enables comparisons of the word class structure of the corpora. Unsupervised POS-tagging uses clustering methods to identify clusters of similarly distributed word. The number of classes found through such an approach is usually higher than the number of traditional part-of-speech classes. However, this finer classification can actually improve the pattern identification to be used (see 3.2.2.3). On the basis of POS-tagging, some basic comparisons that are possible are the relative importance of open and closed classes of word forms, or the distribution of the classes in the sentences. This allows, for example, for some basic word order estimates.

Crucially to the question of complex sentences, word classes can be used to establish large coherent parts of language structure (‘clauses’), and even further approaches are available to establish constituent structure. We will attempt to use (and further develop) such high-level methods to annotate our corpora for complex sentences, but it is clear that by this approach alone it will be very difficult to reach good annotations. For that reason we will also take a multilingual route to amend the basic monolingual unsupervised annotation.

3.2.2.2 Unsupervised multilingual annotation

A relatively new line of research to be performed in this project is to use the multilingual parallel corpora to induce language-specific annotation without supervision. In the field of natural language processing, most time and work is spent on the development of purely monolingual approaches, because this is the most pressing problem for practical applications (translations are normally only available to a limited extent). However, for the scientific field of linguistics it is just as interesting to induce structure from available translations. There is a lot of implicit knowledge available in the multilingual translations and it should be possible to use this knowledge to enhance unsupervised annotation (cf. Ongyerth 2007). However, this is very much an open line of research (cf. Pádo & Lapate 2009 on semantic role annotation) and much explorative research is needed in this realm. All the basic aspects of unsupervised annotation mentioned above (stemming, morphology, parts of speech, chunks, constituent structure) will also be approached from this perspective. To take a high-level example related to complex sentences, consider the following. We can easily find adverbial clauses for various European languages based (among other factors) on the adverbs/prepositions that introduce them (e.g. English ‘when’, ‘in order to’, ‘after’ etc.). Using the parallel texts, we can then search for patterns and regularities in the translations of these clauses in other languages, and in this way identify constructions that seem to function similarly to the European notion of adverbial clauses. This approach then directly leads to the following aspect of this project.

3.2.2.3 Extraction of language-specific characteristics

The basic approach to obtain information about complex sentences will be a bootstrap process based on the massively parallel texts. The basic idea is to identify contexts of interest in the parallel texts in which complex sentences are likely to occur. The identification will be performed through the few European languages for which we have personal knowledge and through information collected by other projects in the Forschergruppe. Based on such a set of relevant contexts (e.g. irrealis-complement contexts as identified by specific matrix predicates, cf. P1 Haspelmath/Michaelis), the research of our project will try to (i) identify language-specific structures in the translations of these contexts and (ii) find similar constructions in the monolingual corpora.

Concerning (i), the relevant structures could be described by a human specialist, but this project will assume that the relevant structures are only implicitly given by the examples presented (except for a few test cases in which we will use manual annotation for evaluation). Hence, an automatic feature extraction approach to be developed in this project has to identify typical structures in the training set of examples (e.g. using specific closed-class words, special morphemes, special part-of-speech patterns, punctuation).

Concerning (ii), the typical problem can be stated as follows: A (usually small) set of sentences is given. The problem is to find more sentences with the same feature in the corpus. Usually, sentence similarity is treated as string similarity. In this case, sentences are similar if they have many words in common. Here, in contrast, we are interested in sentences having the same (or a similar) linguistic structure. This kind of similarity will be called structural sentence similarity in the following. The words in these sentences, which are not related to this structure, may be totally different. Hence, new methods for structural sentence similarity will be developed in this project.

For both the algorithms of (i) and (ii) we will use text-mining and machine-learning approaches. An evaluation on the basis of a few hand-annotated cases will be used to analyze the quality of the results. Such evaluation will be used to improve both the feature extraction algorithm and search pattern generation. At a later stage when the automatic methods are trained using many different languages, the evaluation step will give better results and will have less need for improvement.

3.2.2.4 Quantitative establishment of typological parameters

Two types of typological parameters can be extracted from the data collected: (i) absolute characterizations of each language individually according to particular parameters, and (ii) relative characterizations in the form of the distribution of a particular language-specific form over a set of parallel sentences.

First, given a sizable collection of examples of a particular structure for all languages (e.g. all complex sentences with a matrix verb translated as 'to think'), the induced structure can tell us approximately about the language-particular distribution of characteristics like length (simple count of words or segments), headedness (position and identification of unchangeable part), rough order of constituents (through POS tagging and constituent induction), finiteness of embedded verb (through stemming), or even something about information structure (by using statistical information measurements). All such typological parameters will not just be categorical ('yes or no'), but give continuous quantitative evaluations of the variability of each language.

The second option to establish typological parameters is most well-known from case alignment typology: the case alignment of a language does not depend on one independent characteristic, but on the relation between the encoding of different functions (viz. $A = S \neq P$ vs. $A \neq S = P$). This approach can also be used to characterize language-particular constructions. For example, using religious pamphlets from <http://watchtower.org>, we can compare the German word 'aber' with the (apparent) Faroese translation 'men' (we chose Faroese because we do not know this language, though we can interpret the text with some effort to check results). The pamphlets are completely parallel, so we can compare the precise occasions in which both 'aber' and 'men' occur, and compare them with the contexts where only one of them occurs. The contexts in which these words occur can be interpreted as a rough 'finger print' of their usage, allowing for a direct multilingual comparison.

In a small pilot study, we quickly analyzed one of the pamphlets. The results argue that 'aber' and 'men' are indeed quite similar, but also point to some differences: both 'aber' and 'men' occurs in 19 sentences; 'aber' but not 'men' occurs in 3 sentences; not 'aber' but 'men' occurs in 8 sentences. Such a comparison gives a direct quantitative measure of similarity between languages, which can be used to establish a typology. A wide range of such comparisons can also directly be converted into a semantic map (cf. P1 Haspelmath/Michaelis).

3.2.3 Deliverables

The project will produce the following results:

- Large unannotated corpora for about 200-300 languages
- One massively parallel corpus for about 200-300 languages
- Approximate annotation of the sources through language-independent structure discovery, both monolingual and multilingual
- APIs (Application Programming Interfaces) to access these sources automatically over the internet
- Algorithms to extract typological information from these sources
- Scientific articles about the applicability of corpora for typological research
- Scientific articles about the typological comparison of complex sentences on the basis of our corpora

3.2.4 Work packages

The work of the project will be allocated into ten work packages, as described in the table below. The work packages are roughly sequentially ordered, though some partial overlap will occur (note that A and B will of course both start together in parallel). The numbers indicate the approximate number of months to be spent on the work packages by the envisioned members of this project.

Work package	Cysouw	Mayer	Quasthoff	N.N.
A) Preparation of unannotated corpora			0.5	10
B) Preparation and basic alignment of parallel texts	0.5	4		
C) Unsupervised monolingual tagging			0.5	8
D) Unsupervised tagging using multilingual translations	0.5	10		
E) Evaluation of unsupervised tagging		3		
F) Linking annotations from (C,D) and mutually improve them		4	0.5	4
G) APIs for online access of corpora, copyright clearance	0.5	1	0.5	2
H) Establishment and algorithmic extraction of parameters	1	6	1	6
I) Exemplary evaluation of parameters	0.5	2		
J) Preparing Publications and Dissertation	3	6	3	6
Total amount of Months over 3 Year Period	6	36	6	36

3.2.5 Outlook: difficulty assessment of identifying typological parameters

One of the main goals, planned for the second phase of this project, is to investigate how much explicit knowledge of each language is needed to investigate a typological parameter. Given a particular typological aspect of the structure of complex sentences (e.g. the relative order of main and subordinate clauses, finiteness in the subordinate clause) in the world's languages, we will first try to use only monolingual statistics, then add the insights from the multilingual parallel corpus, and finally add language-specific information in collaboration with other projects in this Forschergruppe (e.g. using available annotated corpora, reference grammars or personal knowledge of language specialists). We expect that some compara-

tive questions can already be answered without much linguistic knowledge, but others will need more manual input. However, exactly which questions will turn out to be easily answered through pure statistics, and which are in need of fine-grained linguistic consideration, is a central question to be answered by this project.

There are two kinds of information for which we plan to test how important they are for the typological assessment of the type of a language: automatically induced knowledge and manually added linguistic information. There are at least four different levels of automatically induced knowledge that might influence the typology of any given parameter:

- Purely monolingual unannotated corpus
- Unsupervised language-independent annotation
- Simplistically parallelized corpus (i.e. only using written word and sentence boundaries)
- Unsupervised annotation of monolingual corpora based on the parallel corpus

As for manually added linguistic information, we expect to be able to compare the following levels in order of increasing knowledge on the typology of any given parameter:

- Manually chosen contexts from parallel texts
- Manual indication of selected crucial link structures of complex sentences
- Manually added detailed interlinear annotation of exemplary sentences for each language

The influence of these different levels of knowledge will be compared to each other, but also to an independent standard as prepared manually by linguists. To some extent the available typological parameters in the World Atlas of Language Structures (WALS) can be used for this, but more crucial will be the evaluation of typological parameters as established in the other projects of the Forschergruppe. Given that this presupposes that the other projects have already obtained some results, we plan to perform these tests in the second phase of the project.

Literatur (Arbeiten der Antragsteller)

- Biemann, Christian, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. 2004. Language-independent methods for compiling monolingual lexical data. *Computational Linguistics and Intelligent Text Processing* 217-228.
- Biemann, Christian, Stefan Bordag, Uwe Quasthoff, and C. Wolff. 2004. Web Services for Language Resources and Language Technology Applications. *Proceedings Fourth International Conference on Language Resources and Evaluation: Lissabon*.
- Biemann, Christian, Gerhard Heyer, Uwe Quasthoff, and M Richter. 2007. The Leipzig Corpora Collection: Monolingual corpora of standard size. *Proceedings of Corpus Linguistic: Birmingham, UK*.
- Biemann, Christian, and Uwe Quasthoff. 2007. Similarity of Documents and Document Collections using Attributes with Low Noise. *Proceedings of WEBIST-07: Barcelona, Spain*.
- Biemann, Christian, Uwe Quasthoff, Gerhard Heyer, and Florian Holz. 2008. ASV Toolbox: A Modular Collection of Language Exploration Tools. *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*.
- Cysouw, Michael. 2002. Interpreting Typological Clusters. *Linguistic Typology* 6.69-93.
- Cysouw, Michael. 2003a. Against implicational universals. *Linguistic Typology* 7.89-101.
- Cysouw, Michael. 2003b. *The Paradigmatic Structure of Person Marking*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.
- Cysouw, Michael. 2005. Quantitative methods in typology. *Quantitative Linguistics: An International Handbook*, ed. by Gabriel Altmann, Reinhard Köhler, and R. Piotrowski, *Handbücher zur Sprach- und Kommunikationswissenschaft*: 27, 554-578. Berlin: Mouton de Gruyter.

- Cysouw, Michael. 2007a. Building semantic maps: the case of person marking. *New Challenges in Typology*, ed. by Bernhard Wälchli, and Matti Miestamo, *Trends in Linguistics: Studies and Monographs*: 189, 225-248. Berlin: Mouton de Gruyter.
- Cysouw, Michael. 2007b. Content interrogatives in Pichis Ashéninka: corpus study and typological comparison. *International Journal of American Linguistics* 73.133-163.
- Cysouw, Michael. 2007c. New approaches to cluster analysis of typological indices. *Exact Methods in the Study of Language and Text*, ed. by Reinhard Köhler, and Peter Grzbeq, *Quantitative Linguistics*: 62, 61-76. Berlin: Mouton de Gruyter.
- Cysouw, Michael, Christian Biemann, and Matthias Ongyerth. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *Sprachtypologie und Universalienforschung* 60.158-171.
- Cysouw, Michael, and Diana Forker. 2009. Reconstruction of morphosyntactic function: Non-spatial usage of spatial case marking in Tsezic. *Language* 85.588-617.
- Hänig, Christian, Stefan Bordag, and Uwe Quasthoff. 2008. UnsuParse: Unsupervised Parsing with unsupervised Part of Speech tagging. *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*.
- Heyer, Gerhard, Uwe Quasthoff, and Thomas Wittig. 2008. *Text Mining: Wissensrohstoff Text -- Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag.
- Quasthoff, Uwe. 2010. Automatisierte Rohdatengewinnung für die Lexikographie. *Lexicographica*, ed. by Ulrich Heid, Stefan Schierholz, Wolfgang Schweickard, Herbert Ernst Wiegand, and Werner Wolski, Berlin: De Gruyter Mouton.
- Quasthoff, Uwe, M. Richter, and Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. *Proceedings of LREC: Genoa, Italy*.
- Quasthoff, Uwe, and Fabian Schmidt. 2010. Die korpusbasierte Identifikation fester Wortverbindungen. *Lexicographica Series Maior*. Berlin: De Gruyter Mouton.
- Wälchli, Bernhard, and Michael Cysouw. 2010. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*.

Literatur (andere Arbeiten)

- Altmann, Gabriel, and Werner Lehfeldt. 1973. *Allemeine Sprachtypologie: Prinzipien und Meßverfahren*. Uni-Taschenbücher. Munich: Fink.
- Baroni, M, and S Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*: 1313-1316.
- Bickel, Balthasar. 2003. Referential density in discourse and syntactic typology. *Language* 79.708-739.
- Biemann, Christian. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. *Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*: 7-12. Association for Computational Linguistics.
- Biemann, Christian. 2007. *Unsupervised and Knowledge Free Natural Language Processing in the Structure Discovery Paradigm*. Ph.D. Thesis University of Leipzig.
- Bordag, Stefan. 2006. Two-step approach to unsupervised morpheme segmentation. *Proceedings of 2nd Pascal Challenges Workshop*: 25-29.
- Bordag, Stefan. 2007. *Elements of Knowledge-free and Unsupervised lexical acquisition*. Ph.D. Thesis University of Leipzig.
- Bordag, Stefan. 2008. Unsupervised and knowledge-free morpheme segmentation and analysis. *Advances in Multilingual and Multimodal Information Retrieval* 881-891.
- Creutz, Mathias, and Krista Lagus. 2002. Unsupervised discovery of morphemes. *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*: 21-30. Association for Computational Linguistics.

- Creutz, Mathias, and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Publications in Computer and Information Science, Report A 81
- Dahl, Östen. 2007. From questionnaires to parallel corpora in typology. *STUF-Sprachtypologie und Universalienforschung* 60.172-181.
- Fenk-Oczlon, Gertraud, and August Fenk. 1999. Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology* 3.151-177.
- Givón, T., ed. 1983. *Topic Continuity in Discourse: A Quantative Cross-language Study*. Amsterdam: Benjamins.
- Greenberg, Joseph H. 1960. A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics* 26.178-194.
- Klein, Dan, and Christopher D. Manning. 2005. Natural language grammar induction with a generative constituent-context model. *Pattern Recognition* 38.1407–1419.
- Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 210.1567-1578.
- Myhill, John. 1992. *Typological Discourse Analysis*. Oxford: Blackwell.
- Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29.19-51.
- Ongyerth, Matthias. 2007. *Automatische Erstellung zweisprachiger Wörterbücher aus Paralleltexten: Ein sprachunabhängiger Ansatz*. Leipzig: Diplomarbeit Universität Leipzig.
- Padó, Sebastian. 2007. *Cross-lingual annotation projection models for role-semantic information*. Saarbrücken dissertations in computational linguistics and language technology. Saarbrücken: DFKI.
- Padó, Sebastian, and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research* 36.307-340.
- Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14.130–137.
- Samardžić, Tanja, and Paola Merlo. 2010. Cross-lingual variation of light verb constructions: using parallel corpora and automatic alignment for linguistic research. *ACL Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING)*. Uppsala.
- Scannell, Kevin P. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Building and exploring web corpora: proceedings of the 3rd Web as Corpus Workshop*, ed. by C. Fairon, H. Naets, A. Kilgarriff, and G-M. de Schryver, *Cahiers du Central*: 4, 5-15. Louvain: Presses Universitaires de Louvain.
- Stolz, Thomas. 2007. Harry Potter meets Le petit prince-On the usefulness of parallel corpora in cross-linguistic investigations. *STUF* 60.100-117.
- Tiedemann, Jörg. 2003. Combining clues for word alignment. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*: 339-346. Association for Computational Linguistics.
- Tiedemann, Jörg. 2009. News from OPUS—A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing V: Selected Papers from Ranlp 2007* 237.
- Tiedemann, Jörg, and Gideon Kotzé. 2009. Building a large machine-aligned parallel treebank. *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories*: 197–208.
- Tiedemann, Jörg, and Lars Nygaard. 2004. The OPUS corpus-parallel & free. *Proceedings of the Fourth International Conference on Language resources and evaluation (LREC'04)*: Lisboa.
- Wälchli, Bernhard. 2005. *Co-compounds and Natural Coordination*. Oxford: Oxford University Press.
- Wälchli, Bernhard. 2007. Advantages and disadvantages of using parallel texts in typological investigations. *STUF* 60.118-134.
- Wälchli, Bernhard. 2009. *Motion Events in Parallel Texts: A study in primary-data typology*. Bern: Habilitationsschrift Universität Bern.
- Wälchli, Bernhard. 2010. Similarity semantics and building probabilistic semantic maps from parallel texts. *Linguistic Discovery*.

4 Beantragte Mittel

4.1 Personalbedarf und Personalkosten

Für das Projekt werden folgende Stellen beantragt:

A) Eine Post-Doktoranden-Stelle

TV-L E13, 100%
für drei Jahre, anzusiedeln in München

Aufgaben:

- Erstellung der parallelen Korpora
- Entwicklung und Anwendung von Annotationsalgorithmen auf Basis der multilingualen Daten
- Linguistische Evaluation aller in diesem Projekt produzierten Annotationen
- Mitarbeit an der Entwicklung der automatischen typologischen Parametererstellung
- Exemplarische Evaluation der automatisch erstellten typologischen Parameter
- Mitarbeit an der Zusammenführung der verschiedenen Korpora und Annotationen innerhalb dieses Projektes
- Zusammenführung der Daten dieses Projektes mit den Daten von P7 (Gast/Bickel)

Diese Stelle würden wir bevorzugt mit Thomas Mayer (Konstanz) besetzen. Er hat den idealen Hintergrund sowohl in der Linguistik als auch in der Informatik, und beschäftigt sich schon jetzt mit der Benutzung paralleler Texte für die komparative Linguistik. Seine Dissertation ist in der letzten Phase und wird voraussichtlich vor Ende 2010 eingereicht.

B) Eine Doktoranden-Stelle

TV-L E13, 65%
für drei Jahre, anzusiedeln in Leipzig

Um diese Stelle für einen Doktoranden der Informatik auch finanziell einigermaßen attraktiv zu machen, beantragen wir einen erhöhten Teilzeitanteil von 65%.

Aufgaben:

- Erstellung der monolingualen Korpora
- Anwendung und Weiterentwicklung von Annotationsalgorithmen für monolinguale Korpora
- Mitarbeit an der Entwicklung der automatischen typologischen Parametererstellung
- Mitarbeit an der Zusammenführung der verschiedenen Korpora und Annotationen innerhalb dieses Projektes
- Erstellung und Dokumentation der APIs für die Abfrage der Korpora und Annotationen

C) Eine Studentische Hilfskraft

10 Stunden/Woche
zur Unterstützung der Projektmitarbeiter bei der Korpusarbeit

4.2 Reisen

Reisekosten für die Projektmitglieder werden gemäß den allgemeinen Vorgaben für Forschergruppen beantragt.

4.3 Workshop

Für die Koordination der verschiedenen Kollegen, die an der Aufarbeitung von Bibelübersetzungen für die komparative Linguistik arbeiten, wollen wir einen Workshop zur Abstimmung der praktischen Aspekte dieses Parallelkorpus abhalten. Dazu ist geplant, folgende Kollegen einzuladen: Jörg Tiedemann (Uppsala), Bernhard Wälchli (Bern), Sergio Meira (Leiden/Nijmegen), Östen Dahl (Stockholm) und Dan Haug (Oslo). Für diesen Workshop beantragen wir 2500,- EUR (= fünf mal 500,- EUR für Reisen innerhalb Europas).

5 Voraussetzungen für die Durchführung des Vorhabens

5.1 Zusammensetzung der Projektgruppe

Die Projektgruppe besteht aus den Antragstellern Dr. Michael Cysouw und Prof. Dr. Uwe Quasthoff, dem voraussichtlichen Mitarbeiter Thomas Mayer und einem Doktorand aus dem Bereich der Informatik/Automatische Sprachverarbeitung, sowie der studentischen Hilfskraft. Der Münchner Teil des Projektes wird angebunden werden an die Forschungseinheit „Quantitativer Sprachvergleich“ an der LMU München, wo es eine enge Zusammenarbeit mit dem ERC-Projekt „QuantHistLing“ gibt, in dem u.a. auch an der Annotation von Texten gearbeitet wird. Der Leipziger Teil des Projektes wird angebunden werden in an das Sprachdatenressourcen-Projekt der Automatischen Sprachverarbeitung, wo schon seit vielen Jahren Webkorpora aufgearbeitet werden.

5.2 Zusammenarbeit mit anderen Wissenschaftlern

Innerhalb der Forschergruppe gibt es enge Verbindungen zu allen anderen Projekten, weil wir uns damit auseinandersetzen wollen, wie die in diesem Projekt vorgeschlagene Aufarbeitung der Korpora für verschiedene linguistischen Ansätze vorteilhaft eingesetzt werden kann. Konkret gibt es eine enge Zusammenarbeit zur P7 (Bickel/Gast), weil die Korpora und die Annotationen aus unserem Projekt mit der in P7 (Bickel/Gast) geplanten Datenbank und den Korpusannotationen eng verzahnt werden sollen. Praktischen Koordinationsbedarf gibt es mit den typologischen Anteilen der Projekten P1 (Haspelmath/Michaelis), P3 (Gast/Schäfer) und P5 (Diessel), weil die Interessen dieser typologischen Studien als Testfälle für die in unserem Projekt erarbeitete Methodik dienen sollen (für weitere inhaltliche Berührungspunkte zu den anderen Projekten der Forschergruppe, siehe die Auflistung unter Abschnitt 3.1.2).

Außerhalb der Forschgruppe gibt es enge Kontakte zu vielen Kollegen, die zu den Themen unseres Projektes arbeiten. Im Bereich der Sammlung und Auswertung von Paralleltextrn stehen wir in Kontakt mit Jörg Tiedemann (Uppsala) und Bernhard Wälchli (Bern). Bei der Sammlung von Bibeltexten werden wir mit mehreren Kollegen zusammenarbeiten, die sich auch aktiv mit die Benutzung solcher Daten für die Linguistik beschäftigen, wie Östen Dahl (Stockholm), Sergio Meira (Leiden/Nijmegen), Bernhard Wälchli (Bern) oder Dan Haug (Oslo). Für die Sammlung monolingualer Korpora arbeiten wir zusammen mit Kevin Scannell (Saint Louis, MO). Weiterhin stehen wir im engen Kontakt zu Christian Biemann (San Francisco, CA/Powerset) und Harald Hammarström (Nijmegen/Leipzig), die sich auch mit der automatischen Verarbeitung und Annotation von Korpora beschäftigen. Für die Übersetzungen der allgemeinen Erklärung der Menschenrechte stehen wir in Kontakt mit Eric Muller (San José, CA/Adobe).

6 Erklärungen

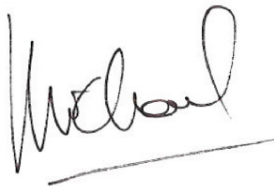
Ein Antrag auf Finanzierung dieses Vorhabens wurde bei keiner anderen Stelle eingereicht. Wenn wir einen solchen Antrag stellen, werden wir die Deutsche Forschungsgemeinschaft unverzüglich benachrichtigen.

Wir verpflichten uns, mit der Einreichung des Antrags auf Bewilligung einer Sachbeihilfe bei der DFG die Regeln guter wissenschaftlicher Praxis einzuhalten.


Wir haben bei der Antragstellung die Regelungen zu den Publikationsverzeichnissen (Leitfaden I.8.) und zum Literaturverzeichnis (Leitfaden II.2.) beachtet.

Die Vertrauensdozenten der Ludwigs Maximilians Universität München und der Universität Leipzig sind von der Antragstellung unterrichtet worden.

7 Unterschriften



Michael Cysouw
München, den 11. August 2010



Uwe Quasthoff
Leipzig, den 13. August 2010