Towards a corpus-based typology of clause linkage: an analytical framework and case studies on non-local dependencies

Balthasar Bickel
Volker Gast

1 Allgemeine Angaben (General information)

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

1.1 Antragsteller (Applicants)

Name:	**Prof. Dr. Balthasar Bickel**	**Prof. Dr. Volker Gast**
Stellung:	Universitätsprofessor	Universitätsprofessor
Geburstdatum:	19.12.1965	14.03.1973
Nationalität:	Schweizer	Deutscher
Institution:	Universität Leipzig Institut für Linguistik	Friedrich-Schiller-Universität Jena Institut für Anglistik und Amerikanistik
Dienstadresse:	Beethovenstrasse 15 04105 Leipzig	Ernst-Abbe-Platz 8 07743 Jena
Telefon:	(0341) 973 7604 (Sekr. 7610)	(03641) 944 546
Telefax:	(0341) 973 7609	(03641) 944 542
E-mail:	bickel@uni-leipzig.de	volker.gast@uni-jena.de
Web:	www.uni-leipzig.de/~bickel	www.uni-jena.de/~mu65qev
Privatadresse:	Hinrichsenstrasse 15 04107 Leipzig Tel. (0341) 998 5666	Hamerlingweg 13 14167 Berlin Tel. (030) 8049 9755

1.2 Thema (Topic)

Towards a corpus-based typology of clause linkage: an analytical framework and case studies on non-local dependencies

Grundlagen einer korpusgestützten Typologie der Satzverknüpfung: ein analytischer Rahmen und Fallstudien zu nicht-lokalen Abhängigkeiten

1.3 Fach- und Arbeitsrichtung (Research area)

104-03: Sprachwissenschaften/Typologie

1.4 Voraussichtliche Gesamtdauer (Anticipated total duration)

72 Monate (6 Jahre)

1.5 Antragszeitraum (Application period)

Beantragte Förderungsdauer	36 Monate (3 Jahre)
Beantragter Förderungsbeginn	01.04.2010

1.6 Englische Zusammenfassung (English summary)

Typology is moving away from seeking absolute, categorical constraints on human language towards research into probabilistic patterns. Such quantitative typology is still mostly based on categorical abstractions made in descriptive grammars, even though there is growing evidence that not only typological generalizations but also individual grammars themselves are probabilistic. What is needed, then, is a way of extracting typological generalizations directly from discourse data, acknowledging such aspects of usage as lexical and contextual biases. This research program

requires cross-linguistic corpora. Since corpus development has proven to be the most efficient means of salvaging linguistic diversity from endangerment, the number of corpora is steadily increasing, and this makes corpus-based typology a realistic enterprise.

This project aims at developing an analytical framework (implemented as a database) and an infrastructure for corpus-based typology, in close cooperation with all other projects of the research unit. It investigates clause linkage because this area of grammar is closely tied to discourse patterns and is highly probabilistic, thus providing an excellent basis for exploring the interface of language use and grammatical patterns. We will sample corpus exemplars of strings of predicate-headed units ('clauses') and richly annotate these strings for their structural properties in context, including not only syntactic and morphological parameters establishing inter-clausal relationships, but also cross-clausal dependencies (e.g. long-distance anaphora, extraction). The data will be analyzed by using statistical data-mining methods in order to detect and explore probabilistic patterns of clause linkage both within and across languages.

1.7 Deutsche Zusammenfassung (German summary)

Die Sprachtypologie wendet sich zunehmend von der Suche nach absoluten, kategorischen Universalien menschlicher Sprache ab, hin zur Untersuchung probabilistischer Verteilungsmuster. Diese Art quantitativer Typologie beruht zumeist noch auf kategorischen Abstraktionen aus deskriptiven Grammatiken, obwohl es immer mehr Evidenz dafür gibt, dass nicht nur typologische Generalisierungen, sondern auch individuelle Grammatiken probabilistisch sind. Folglich sollten typologische Abstraktionen direkt aus Diskursdaten extrahiert werden und Aspekte des Sprachgebrauchs wie lexikalische und kontextuelle Tendenzen berücksichtigen. Die für dieses Forschungsprogramm notwendigen Korpora stehen in zunehmendem Maß zur Verfügung, da sich die Korpuserstellung als Mittel der Bewahrung sprachlicher Vielfalt inzwischen etabliert hat. Korpusgestützte Typologie ist somit möglich geworden.

Im beantragten Projekt wird beabsichtigt, in Zusammenarbeit mit den anderen Projekten der Forschergruppe einen analytischen (als Datenbank implementierten) theoretischen Rahmen sowie eine Infrastruktur für korpusgestützte Typologie zu entwickeln, und zwar am Beispiel der Satzverknüpfung, da dieser Bereich stark diskursgesteuert und somit probabilistisch ist und folglich eine hervorragende Grundlage für die Untersuchung der Schnittstelle zwischen Sprachgebrauch und Grammatik bietet. Wir werden Stichproben von prädikativen Einheiten (‚clauses') erstellen und diese nach strukturellen Kategorien annotieren, unter besonderer Berücksichtigung nichtlokaler Abhängigkeiten. Die Daten werden mit Hilfe statistischer *data mining* Verfahren analysiert, um sowohl innersprachliche als auch übereinzelsprachliche probabilistische Muster der Satzverknüpfung zu identifizieren.

2 Stand der Forschung, eigene Vorarbeiten (State of the art, earlier work)

2.1 Stand der Forschung (State of the art)

2.1.1 The empirical basis of typological generalizations

While in the past century typology mostly shared the goals of generative grammar and attempted to establish absolute universals as the boundary conditions of the human language faculty, the field has started to move away from this, in favor of research into probabilistic patterns (Dryer 1998, Bickel 2007). This move is motivated by at least three insights: (i) Absolute universals cannot be established on the basis of language samples (since a sample can never guarantee

that the next language outside the sample will not be an exception), but samples allow probabilistic generalizations of what is more vs. less likely to develop (Cysouw 2005, Bickel in press-a). (ii) The worldwide distribution of grammatical structures is not driven by universal principles alone but is deeply affected by areal diffusion patterns and inheritance patterns within families (Nichols 1992 and many others since), none of which is deterministic. (iii) Modern tools of data-mining (e.g. distance-based methods: Cysouw 2007, Croft & Poole 2004), statistical modeling (e.g. generalized linear models: Justeson & Stephens 1990, Bickel 2008a), hypothesis testing (e.g. randomization tests: Janssen et al. 2006), etc. have made it possible to compute distributional patterns from typological databases.

The databases used in this research, however, are mostly based on categorical abstractions made in descriptive grammars, i.e. categorical classifications of entire languages as having features like "SVO" or "de-ranking complement clauses". Even when such classifications are relativized to grammatical subsystems — thus allowing for splits like 'SVO in main clauses, but SOV in dependent clauses' — many such statements still do not sufficiently reflect actual patterns in the language: in many languages, choices in word order or subordination strategies are probabilistically conditional on the discourse context (e.g. specific epistemic functions of complement clause constructions) and the lexical material (e.g. specific matrix verbs in complement clause constructions). Such usage conditions have been shown to play a key role in language acquisition, i.e. in how grammars are transmitted over time (e.g. Tomasello 2003, Diessel 2004).

As discussed in more detail below, many fields of linguistics have recognized this, and analyses of corpora, both quantitative and qualitative, have become a standard. Similarly, formal theories have begun to incorporate probabilistic modeling, as for example in probabilistic versions of Optimality Theory (e.g. Jäger 2007). In typology, however, generalizations have only occasionally been based directly on corpus data (e.g. Greenberg 1959, Fenk-Oczlon & Fenk 1985, Bickel 2003, DuBois et al. 2003, Wälchli 2007). When corpus data enter typological work, they do so mostly as the basis for usage-based motivations of typological generalizations (such as the preference for subject-before-object orders or universalist definitions of 'subordination'). But the generalizations themselves are based on categorical abstractions in descriptive grammars (e.g. Givón 1983, DuBois 1987, Hawkins 1994, Croft 2000, Haspelmath 2008, among many others). From the perspective of usage-based theory, this is a curious situation: if usage-based aspects are essential components of grammatical constructions, it seems odd to blend them out when searching for cross-linguistic generalizations and then feed them back in only when explaining the generalizations.

There is of course also a practical issue involved: until recently not many corpora have been available beyond better-studied European languages. However, this is rapidly changing as the result of large-scale initiatives to document endangered languages in the form of corpora (such as the DoBeS and ELDP initiatives;[1] cf. Ostler 2008 for an overview). And in turn, the fact that endangered languages research has identified corpus development as the most efficient means for salvaging linguistic diversity means that in the future many languages will be in fact better accessible through corpora than through elicited datasets published in descriptive grammars. In the end, typology will need to be based more directly on corpora anyway.

What is urgently needed, then, is an analytical framework and a set of methods that allows one to efficiently detect probabilistic patterns of grammar both within and across corpora.

[1] `http://www.mpi.nl/DOBES/`, `http://www.hrelp.org/`.

2.1.2 Multilingual corpora

While corpora are not standardly used in linguistic typology, other branches of comparative linguistics, e.g. contrastive linguistics (cf. Gast forthcoming, Johansson 1998) and variation linguistics (e.g. Kortmann & Szmrecsanyi 2009) rely heavily on them. One of the main challenges of comparative corpus linguistics is the establishment of comparability, i.e. a way to make sure that quantitative differences between corpora reflect properties of the underlying systems, rather than properties of the texts or text types. This is achieved by compiling appropriately sampled 'multilingual' corpora. Two major types of multilingual corpora are commonly distinguished, 'parallel' and 'comparable' ones. Parallel corpora contain a source text and its translation(s), while comparable corpora are not translational equivalents but have been sampled from similar genres (see for instance McEnery et al. 2006).[2] Parallel and comparable corpora are used for different purposes, the former primarily in applied linguistics (e.g. translation studies) and the latter typically for the investigation of theoretical questions (see Aijmer 2008 and Xiao 2008 for overviews and discussion; cf. also P6 Cysouw/Quasthoff).

2.1.3 Software for corpus annotation

There are several reasons for the neglect of corpus methods in linguistic typology. First, cross-linguistic corpora, while being a steadily growing resource, are still relatively scarce and not easily accessible; second, and more importantly perhaps, appropriate software for the annotation and processing of cross-linguistic corpora is not yet available. Most of the relevant software used by fieldworkers is geared towards the storage of lexical information and interlinear glossing (e.g. Shoebox/Toolbox, a product distributed by the Summer Institute of Linguistics). Structural matters such as prosodic phrasing, constituency information (tree building) and anaphoric dependencies are not standardly provided for by such tools.

A number of annotation tools are available for specific data formats, and for specific tasks. For example, the *Annotate* tool, which was developed at the Sfb 378 (Saarland) for the annotation of the NEGRA corpus and is also used for the TIGER corpus, provides a graphical user interface for morphological, syntactic and functional (relational) annotations (cf. Plaehn & Brants 2000).[3] The SALSA[4] Annotation tool SALTO ('SALsa TOol', cf. Burchardt et al. 2006), which supports corpora in the SALSA/TIGER XML format (cf. below), can be used to generate lexical semantic annotations in a FrameNet style. MMAX2 is a useful tool for the encoding of anaphoric dependencies within a text (cf. Müller & Strube 2006). EXMARaLDA, developed at the Sfb 538 (Hamburg), allows for a detailed (though handmade) annotation of dialogue data (cf. Schmidt & Wörner 2009). The NITE XML Toolkit (NXT)[5] contains highly configurable software for the transcription, annotation and querying of multi-modal corpora. For matters relating to conversational analysis, ELAN ('EUDICO Linguistic Annotator')[6] and the RST Annotation Tool (RST: 'Rhetorical Structure Theory')[7] are also often used. Valuable though these tools are for the tasks that they are designed for, none of them provides the full range of functionality required by a typological project aiming at rich and fine-grained corpus annotation.

---

[2] As is noted there, the terms 'parallel' and 'comparable' corpus are used differently by some authors.

[3] Cf. `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html`.

[4] 'SAarbrücken Lexical Semantics Annotation and Analysis Project', cf. `http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index-salsa1`.

[5] Cf. `http://groups.inf.ed.ac.uk/nxt/`.

[6] Cf. `http://www.mpi.nl/tools/elan.html`.

[7] Cf. `http://www.isi.edu/licensed-sw/RSTTool/index.html`.

2.1.4 Data models and annotation schemes

While a specifically 'typological' annotation software is not yet available, considerable progress has been made in the development of (standards for) annotation schemes and the underlying data models. Since Bird & Liberman (2001), the concept of 'annotation graph' has established itself as a standard for the logical structure of linguistic annotations (see also Evert et al. 2003 on graph theory in general, and on the NITE Object Model in particular). The 'Linguistic Annotation Framework' (LAF) (cf. Ide & Romary 2006) is under development as an ISO standard (ISO 24612). At HU Berlin, a generic data model called 'Salt' has recently been created which is intended to complement LAF (cf. Zipser & Romary 2010). Irrespective of matters of standardization, the various models are very similar already, and there is a high degree of compatibility between them.

The annotation schemes available (i.e. the technical implementations of underlying data models) show more heterogeneity, but standardization efforts are gaining momentum. Early initiatives like TEI ('Text Encoding Initiative') and the formulation of a 'Corpus Encoding Standard' (CES) have provided the basis for establishing best practices in corpus mark-up. Today, there is wide consensus that XML (rather than the more generic SGML) should be used as a mark-up language (cf. XCES, the XML-version of CES). The family of XML-based formats comprises annotation schemes such as SALSA/TIGER XML (cf. Lezius 2002:Ch.7, Erk & Pado 2004),[8] PAULA ('Potsdamer Austauschformat für Linguistische Annotationen', developed at the Sfb 632 [Berlin/Potsdam]),[9] and, most importantly perhaps, GrAF ('Graph Annotation Format'), an XML-linearization of the LAF data model (also under discussion as an ISO standard; cf. Ide & Suderman 2007). Zipser's Salt model also comes with an XML implementation (Salt-XML), for which a documentation is not yet available, however.

The genericity of the underlying graph models has led to a considerable degree of intercompatibility between the various formats. Ide & Suderman (2007) show how five different formats can be transducted into LAF/GrAF. The TIGERRegistry administration tool supports conversion from several formats into TIGER XML.[10] At HU Berlin, a conversion tool 'Pepper'[11] (based on the Salt data model) has been built which converts several input formats (like those used by the annotation tools mentioned above, e.g. *Annotate*, MMAX2, EXMARaLDA, etc.) into PAULA. PAULA, in turn, feeds into a database system called ANNIS ('ANNotation of Information Structure'), which is maintained by the Sfb 632.[12] ANNIS provides a graphical user interface for the querying and visualization of multi-level annotations. Annotations generated with different tools (and at different levels of analysis) can thus be merged into a single representation.

Even though the strategy of 'distributed multi-level' annotations pursued by the Sfb 632 is certainly a practicable way of dealing with the requirements of fine-grained corpus-based analyses, the development of a single annotation tool remains a major desideratum of cross-linguistic corpus research and is in fact one of the central objectives of our project.

2.1.5 Typological databases and corpora

While an infrastructure for typological corpus research is not yet available, great progress has recently been made in the development of typological databases (cf. Everaert et al. 2009). Typological databases are typically built around three types of 'entities' – i.e. objects that a database

[8] See also `http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html`.
[9] Cf. `http://www.sfb632.uni-potsdam.de/~d1/paula/doc/`.
[10] See (`http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TIGERRegistry.html`).
[11] Cf. `http://korpling.german.hu-berlin.de/trac/saltnpepper`.
[12] Cf. `http://www.sfb632.uni-potsdam.de/d1/annis/`.

provides information about (cf. also Dimitriadis 2009): (i) 'languages', understood as any type of (idealized) linguistic system (variety, idiolect, etc.), (ii) 'examples' (tokens of language instantiating the system in question), and (iii) 'constructions' (alternatively 'strategies', 'patterns', 'markers', 'structures', or simply the 'categories' of a language, i.e. the units of linguistic analysis; we will use the term 'construction' in the following). The entities of a database are those objects that we can ask questions about. Questions and answers are represented as pairs of 'attributes' (or 'variables', 'parameters', 'features', etc.) and 'values' (also called 'features', 'properties', etc.; there is a broad range of corresponding dichotomies). A construction can then be seen as a set of attribute-value pairs, e.g. converbial subordination could be defined by pairs such as <LAYER OF EMBEDDING: adverbial>, <OPERATOR SCOPE: disjunct>, <SUBJECT REFERENCE: conjunct> etc.

These data models are not too different from those underlying annotated corpora. In particular, most typological databases – like corpora – contain tokens of annotated language data (though unlike in corpora, examples in databases are typically elicited or, when taken from textual sources, presented out of context). The main difference between typological databases and corpora thus is that the former contain information about abstract constructions while the latter are limited to information about concrete exemplars. However, an annotated corpus can also be regarded as a an 'emergent database', i.e. a database that is generated on the fly by carrying out a query of specific sets of attribute-value pairs. Unlike a (classical) typological database, it is highly flexible with respect to the sets ('constructions') that it provides information about, as these sets are defined in each query anew.

2.2 Eigene Vorarbeiten (Earlier Relevent Work by the Applicants)

2.2.1 B. Bickel

Syntactic typology of clause linkage: database and methods Bickel (in press-b) proposes a fine-grained system of variables for measuring the degree to which clause linkage constructions differ from each other, both within and between languages (also cf. Schackow et al. in press). Table 1 illustrates the variables for some constructions in Belhare (Sino-Tibetan) and Chechen (Nakh-Daghestanian).[13]

Language	*Marker*	*Illoc. scope*	*Tense scope*	*Finiteness*	*Symmetry*	*Illoc. marking*	*Tense marking*	*Questions*	*NP extraction*	*Focusability*	*Position*
Belhare	*ki*	flexible	extensible	finite	asymm.	harmonic	✓	✓	✓	✓	flex.
Belhare	*naa*	disjunct	extensible	finite	flexible	*	*	✓	*	✓	flex.
Chech.	*na*	extens.	conjunct	nonfin.	asymm.	*	*	✓	*	*	flex.
Chech.	*nach*	disjunct	local	nonfin.	asymm.	*	✓	✓	*	✓	flex

Table 1: Sample entries from Bickel (in press-b)

[13] 'Flexible' refers to the absence of constraints, 'disjunct' means that the scope of an operator can be on either the dependent or the main clause but never on both at once, 'conjunct' that the scope must include both clauses, 'local' that the scope is limited to the main clause, 'extens[ible]' that the scope can be on either the main clause alone or on both the main and the dependent clause, and 'harmonic' that both clauses must express the same category choice.

The (dis)similarity between any two constructions can be measured by computing the proportion of identical values (the 'relative Hamming distance'). In Table 1, this suggests that Belhare *ki*-constructions are more similar to Chechen *nach*-constructions than to Chechen *na*-constructions (sharing 5 out of 11 properties as opposed to 3 out of 11). All pairwise comparisons are then aggregated into a single representation capturing the overall similarities between all constructions. Bickel (in press-b) applies the split-graph 'NeighborNet' technique (Bryant & Moulton 2004, Huson & Bryant 2006) to a sample set of 69 constructions from 24 languages.[14] The result of this allows discovering that the Chechen *na*-construction is typologically positioned between detached participle constructions common in Europe and chaining-like constructions typical of New Guinea and Africa, while the Chechen *nach*-construction is placed in a cluster (a cross-linguistic 'prototype') that collects *and*-like coordinations (including Belhare *ki*-constructions, labeled 'chain') on the one hand and topic-forming ad-sentential clauses on the other hand.

Traditional notions like 'subordination' are defined by sets of mutually entailed properties. For example, 'adverbial subordination' is traditionally expected to impose disjunct scope of illocutionary operators and a ban on question word formation in and extraction from dependent clauses. However, instead of rigid entailments ('absolute universals'), we more commonly find probabilistic associations, and Bickel (in press-b) uses entropy-based methods to detect these algorithmically. One of the findings is an association between disjunct scope of illocutionary operators and a ban on question word formation in dependent clauses. This is true of most European cases of 'adverbial subordination', and it seems to be indeed the most frequent pattern in the data surveyed, but as Table 1 above shows, it is only a probabilistic, not a categorical association: Belhare *naa*-clauses tolerate genuine questions inside the dependent clause even though such clauses share many other properties with 'adverbial subordination', including disjunct scope of illocutionary operators.

The statistical techniques used on databases of constructions can be easily ported to databases of corpus exemplars: instead of constructions, the database rows are filled by exemplars. From these, one can estimate patterns of similarities and associations between variables in the same way as Bickel (in press-b) has done for constructions. The result will again be probabilistic generalizations within and between languages, but now directly based on language use.

Corpus development and exploitation Over the past six years, Bickel has been leading a multinational research team developing one of the largest glossed corpora of an endangered language, focusing on two languages of the Kiranti branch of Sino-Tibetan in Nepal: Puma (Southern Kiranti) and Chintang (Eastern Kiranti). For Puma we now have a total of 152,000 words transcribed, of which about 90% are glossed. The Chintang corpus is considerably larger and continues to grow thanks to three follow-up projects: (i) a Dilthey grant on language acquisition research from the Volkswagen Foundation to Sabine Stoll (MPI for Evolutionary Anthropology, Leipzig), (ii) a EuroBABEL grant on differential object treatment from the DFG to Balthasar Bickel, and (iii) a Ph.D. project by Tyko Dirksmeyer on conversational structures at the MPI for Psycholinguistics in Nijmegen. Currently, the Chintang corpus includes a total of 520,000 words (transcribed and translated), of which about 75% are fully glossed (and the rest is in the process of being glossed). Two thirds of this come from a longitudinal study of language acquisition, but these data also contain a large amount of spontaneous conversation among adults. The corpus is enhanced by collections

[14] Stored as a database in the AUTOTYP system, cf. `http://www.uni-leipzig.de/~autotyp`.

of morphological paradigms and ethnographic notes. All data are available through the DoBeS portal via a specific license and cooperation agreement.[15]

Bickel has designed a flexible tool for converting and importing corpora into the statistical environment R (R Development Core Team 2010)[16] and has begun to contribute quantitative analyses of these corpora to joint work with other team members: (i) Stoll et al. (2009) examine the development of the noun-to-verb ratio among Chintang children and find that this approaches adult levels only when children have mastered the inflectional morphology of the language, as measured by paradigm entropy, i.e. by the extent to which the choice of individual verb forms at any given time becomes unpredictable to the same extent as with adults. (ii) Gaenszle et al. (2010) show that Puma ritual language is characterized by special emphasis on nominal referents and places, and this receives statistical support through a systematic analyses of noun and verb distributions across indigenous genres and styles ('shamanic' vs. 'priestly' ways of chanting).

Together with Sabine Stoll, Bickel has co-directed a number of M.A. theses applying quantitative methods to the Chintang corpus (Taras Zakharko: "Identification of syntactic patterns in large corpora and aspects of structure in Chintang child-surrounding speech"; Sebastian Sauppe: "Der Erwerb der Morphosyntax von Lokaldeixis im Chintang"; Claudia Polkau: "Aspekt im Chintang und im Italienischen: Grammatik und Erwerb"). Further such work is in progress under Bickel's (co-) supervision (specifically one M.A. thesis on Chintang clause linkage by Felix Klein and one on the acquisition of relative clauses by Kristina Kuhn; and a Ph.D. thesis on differential verb agreement by Robert Schikowski).

2.2.2 V. Gast

Gast has been involved in a number of typological database projects. The *Typological Database of Intensifiers and Reflexives* (TDIR, published in 2002; cf. Gast 2009)[17] was one of the first typological online databases of its generation. It emerged from a typological project on intensifiers (funded by the DFG and directed by E. König/FU Berlin) and was created by Gast, D. Hole, P. Siemund and S. Töpper. The technical implementation on the basis of PHP-pages and a MySQL-database was done by Gast. The database formed part of the *Linguistic Typology Research Centre* initiative, which was sponsored by the EU (6th Framework) and coordinated by M. Everaert (Utrecht). Later, the 'Typological Database of Intensifiers and Reflexives' was integrated into the 'Typological Database System' (TDS).[18] Together with A. Dimitriadis and the programmer F. van Vugt (Utrecht), Gast was moreover responsible for the development of the *Berlin Utrecht Reciprocals Survey* (BURS), a highly flexible database system geared towards the requirements of typological research. BURS was developed in a bilateral cooperation project jointly funded by DFG and NWO and directed by E. König and M. Everaert (2005-2008). While the project investigated a specific typological problem (the encoding of reciprocals), the database system was intended to be usable for any other domain as well. Today, the system of BURS is used for a number of databases worldwide, a prominent one being the 'African Anaphora Project' directed by Ken Safir (Rutgers).[19]

In 2006, Gast edited a special issue of the *Zeitschrift für Anglistik und Amerikanistik* with the title 'Empiricism in English Linguistics: The Scope and Limits of Corpus Linguistics' (Gast 2006d).

[15] Cf. http://www.mpi.nl/DOBES; http://www.uni-leipzig.de/~ff/cpdp/frameset_AccessRights.html

[16] Implemented by Taras Zakharko and available at http://www.uni-leipzig.de/~bickel/research/software.html

[17] See http://www.tdir.org.

[18] See http://languagelink.let.uu.nl/tds/.

[19] See http://africananaphora.rutgers.edu.

This issue emerged from a workshop (organized by the editor) where recent developments in English corpus linguistics were discussed.

Gast has moreover published on the topics to be investigated in the case studies (cf. Sect. 3.2.3). Gast (2004) deals with the interpretation of non-local *self*-forms in English (cf. also Koenig & Gast 2002). A typology of pronominal expressions based on their locality behaviour is presented in Gast (2006c:Ch.7). Some of Gast's publications deal with scope-bearing elements, in particular focus particles. Gast (2006a) contains a corpus-based investigation of the English particles *also* and *too*. A general survey of focus particles is provided in Gast (2006b). More recently, Gast has worked on 'scalar additive operators', i.e. elements like Engl. *even*, Germ. *sogar*, etc. A typological overview of scalar additive operators is provided in Gast & van der Auwera (2010). In Gast & van der Auwera (forthcoming), a typology is proposed in which scalar additive operators are classified with respect to their scope properties.

3 Ziele und Arbeitsprogramm (Goals and work plan)

3.1 Ziele (Goals)

The overall goal of the project is to develop a system of variables for annotating tokens of clause linkage in corpora in order to make it possible to statistically compute generalizations both within individual languages and as cross-linguistic trends. This will strengthen the empirical grounding of language-specific and typological generalizations and allow insights into the relationship between such generalizations and the individual exemplars that ultimately constitute what speakers process and through which children acquire language.

We aim to reach this overall goal in two phases. Phase I corresponds to the first three years of funding and is dedicated to the development of the system of variables and the annotation of 'model corpora'. The metalanguage and its representation in the form of corpus annotations will be accompanied by case studies of topics that require rich and particularly challenging annotations, viz. various types of non-local dependencies. Phase II corresponds to the second three years of funding and will focus on the annotation of larger-scale corpora, including those provided by other projects of the research unit. Moreover, we will build sets of comparable corpora (cf. Sect. 2.1.2), classifying the relevant texts according to genre. The corpora will then be analyzed by applying statistical data mining techniques for discovering generalizations both within individual and across languages. For this, we will methodologically draw on the work of Bickel (in press-b), as summarized in Sect. 2.2.1.

Based as it is on fine-grained annotations, our project is intended to stand in a complementary relation to P6 (Cysouw/Quasthoff), which aims at the automatic extraction of information from large, unannotated corpora. We will thus be able to compare the results arrived at by applying two diametrically opposed methods to samples of identical data sets. Moreover, our annotated corpora can serve as training sets for the data-mining algorithms of P6 (Cysouw/Quasthoff). In Phase II we intend to integrate the two methodologies, optimizing the relation between data input (annotation) and output (generalizations).

With this background, the immediate goals for the current funding period (i.e. Phase I) are the following:

1. Develop a **system of variables** for annotating exemplars of clause linkage, apply it to a test suite yielding **model corpora**, and critically evaluate the results in comparison

   (a) to variables developed for typological surveys of clause linkage constructions,

(b) and to the metalanguages available in formal theories, especially constructional theories.

2. Develop a **technical infrastructure** for the annotation, storage and analysis of cross-linguistic corpora.

3. Work out **case studies** on challenging aspects of annotating clause linkage exemplars, viz. exemplars that involve all kinds of non-local dependencies, such as extractions, long-distance anaphora or reflexivization, switch-reference or operator scope on non-adjacent dependent clauses.

Goals 1 and 3 form the topic of dissertation projects, while Goal 2 rests in the responsibility of the entire project team.

3.2 Arbeitsprogramm (Work Plan)

3.2.1 Development of a system of variables for annotating corpus exemplars

Variables In classical typological databases, constructions are annotated for sets of typological variables (parameters, features) and the values of these variables are based on elicitation and/or descriptions in grammars. For example, using the system of variables developed in Bickel (in press-b), one could characterize the (converbal) *saŋa*-construction of Chintang as shown in the column 'Construction' in Table 2. Such information typically contains statements about the range of what is possible or impossible in the language — information that can of course only be determined on the basis of systematically elicited sets of grammatical and ungrammatical sentences. A single exemplar, such as the one in (1) only supports 'token' annotations of what is or is not the case, shown here under the column header 'Exemplar' in Table 2.

(1) *im-saŋa=ta gol khoŋs-a-c-e, aŋ?* [CLLDCh3R08S01.0215]
sleep-CONVERB=FOCUS ball play-PAST-DUAL-PAST what
'You were asleep while playing football, right?' (i.e. 'you didn't pay attention!')

Variable	*Construction*	*Exemplar*
Question scope	disjunct	on DEP (question)
Tense scope	conjunct	conjunct
Finiteness of DEP	nonfinite	nonfinite
Illocutionary marking in DEP	not allowed	not marked
Tense marking in DEP	not allowed	not marked
Symmetry of categories expressed	asymmetric	asymmetric
WH and focus formation in DEP	allowed	none
Extraction from DEP	allowed	none
Focus on DEP	allowed	marked
Position of DEP	fixed:pre	pre
Layer of attachment	ad-V or ad-S	ad-S
Coreferential arguments	{S, A}	S
Argument realization	non-{S, A}	none

Table 2: Properties of a V-*saŋa* V sequence.[20]

[20] *Abbreviations:* DEP: 'dependent clause', conjunct: 'scope over both dependent and main clause', disjunct: 'scope over one clause only', S: 'sole argument', A: 'most agentive argument'

'Token variables' only allow inference on what is possible: for example, if question scope falls onto the dependent clause, as it does in (1) ('I know you played, but you were asleep while doing so, right?'), it follows that dependent-clause questioning is possible in the construction. However, large sample sets of annotations allow statistical estimates of general patterns: if one finds that in, say, 90% of V-*saŋa*-V exemplars, question scope falls on either the dependent (here, 'you were asleep, right?') or the main clause ('did you play?'), and there are only few cases where the scope falls on both clauses at the same time, we can infer that constructions with *-saŋa* have a probabilistic preference for disjunct question scope. As noted in Sect. 2.2.1, constructional properties are traditionally assumed to stand in systematic implicational relations to each other: for example, if a dependent clause is adjoined to a sentence ('ad-S') and has disjunct question scope, one expects it not to allow question formation and/or extraction. Again, given sufficiently large collections of corpus exemplars, one can estimate the probability to which such an association holds between variables, and this probability can be used to characterize the construction in a language.

Such preferences are of course different from elicitation-based categorical rules, and even if a percentage reaches 100% in a given sample, this does not entail that the preference is in fact a categorical rule since the next utterance outside the sample may still violate it (as per Cromwell's Rule). However, preferences approximate the stochastic principles that determine how speakers produce language and that help children acquire the language. Moreover, they allow direct computation of typological generalizations: if one finds the same probabilistic association between disjunct operator scope and ad-S attachment in many languages, and if this can be shown to be independent of time and space, this suggests a probabilistic universal of language.

In the development of the set of variables (metalanguage) we will take earlier work — specifically, but not exclusively, Lehmann (1988), Cristofaro (2003), Bickel (in press-b) — as a seed and will cooperate closely with other projects of the research unit, in particular P1 (Haspelmath/Michaelis, encoding of cross-clausal argument sharing), P2 (Comrie/Grawunder, prosodic annotations), P3 (Gast/Schäfer, semantic inter-clausal links), and P4 (Lühr/Zeilfelder, information structural annotations).

Model corpora We aim at a system of variables that is empirically well-motivated and fully operationalizable. This makes it imperative that we work through a substantial set of exemplars in corpora of different languages. To this end, we will compile a test suite of exemplars from corpora of languages that we are familiar with and that are listed in Table 3: from each corpus, we will extract random samples of 2000 strings, each containing two predicates and some evidence for connectedness (a specific marker, an orthographic comma (in written corpora), or verb-verb juxtaposition that is otherwise unusual for the language). The resulting test suite will be used for developing and testing the operationalizability of all variables — especially those capturing long-distance dependencies — and will thereby lead to the construction of fully annotated 'model corpora'. In this work we will cooperate closely with P3 (Gast/Schäfer, for English) and P4 (Lühr/Zeilfelder, for Latin).

Also, we will subject selected individual exemplars to more extensive analysis, formulate these analyses in terms of constructional theories of grammar (specifically, Construction Grammar, Role and Reference Grammar, Head-Driven Phrase Structure Grammar, Lexical Functional Grammar) and compare the attribute-value structures provided by our set of variables to the attribute-value structure derived from these theories. This comparison will help us sharpen and improve our system and feed back into the development of these theories (in the form of spin-off articles).

Language	Family	Size	Source
English	Indo-European	1,000,000 words	ICE-GB treebank
Latin	Indo-European	approx. 102,000 words (parsed)	Perseus treebank
Chintang	Sino-Tibetan	520,000 words transcribed and translated, of which ca. 400,000 glossed and translated	Bickel et al. (2010)
Nepali	Indo-European	260,000 words spoken Nepali transcribed, plus 805'000 words written Nepali	Bhāṣā Sañcār[21]

Table 3: Sample corpora to be used in Phase I

3.2.2 Development of a technical infrastructure

Corpus annotation in standoff XML format The metalanguage for the annotation variables will be implemented on the basis of the generic data model 'Salt' (cf. Zipser & Romary 2010), which underlies the 'Pepper' conversion tool used at the Sfb 632 (cf. Sect. 2.1.4). Salt comes with two XML-linearizations, Salt-XML and LAF/GrAF. It is based on a general graph model and thus highly compatible with other models and their implementations. We are in contact with the developer of Salt (Florian Zipser/HU Berlin) and have agreed to cooperate closely in the development of our corpus tools (cf. below).

As recommended in the XCES ('Corpus Encoding Standard for XML'), we will use a 'standoff format' (also called 'standalone') for our annotations. LAF/GrAF, Salt-XML and (the standoff version of) PAULA are examples of such formats. We will use GrAF as the native format of our corpora, but compatibility with both PAULA and Salt-XML will be ensured. In a 'standoff' architecture, the raw text and the annotations are stored in separate files. The corpus itself simply contains the raw data along with some metadata. In a first step, tokens (the minimal 'markables') are identified. Given that our corpora require fine-grained annotations at the morphological level, the corpora will be tokenized into morphs. The tokenizer software maps spans of characters in the primary text to named (numbered) tokens in the 'tokenization file'. This is illustrated in (2).

(2)

H	o	w		b	e	e	r		i	s		b	r	e	w	-	e	d
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
		↓				↓			↓					↓			↓	
		tok_1				tok_2			tok_3					tok_4			tok_5	

The markables thus identified ('tok_1', 'tok_2', etc.) can now be assigned annotations. Annotations can either be 'structural' (indicating constituency and operator scope) or 'classificatory' (classifying the markables in terms of some linguistic category). In structural annotations, constituents are defined as spans or sets of markables. In classificatory annotations, attribute-value pairs are assigned to (simple or complex) markables. The annotations are distributed over several files, comparable to tables in a relational database. They are linked via the appropriate functions of the XML Pointer Language (XPointer). Such cross-file references are indicated by arrows in Figure 1, where a standoff architecture is represented in a simplified form.

As pointed out above, the annotation schemes used for different corpora are highly intercompatible, as the underlying data models are largely parallel. We will thus be able to convert

[21] `http://www.bhashasanchar.org`. We already have user licence agreements for this and all other corpora listed here.

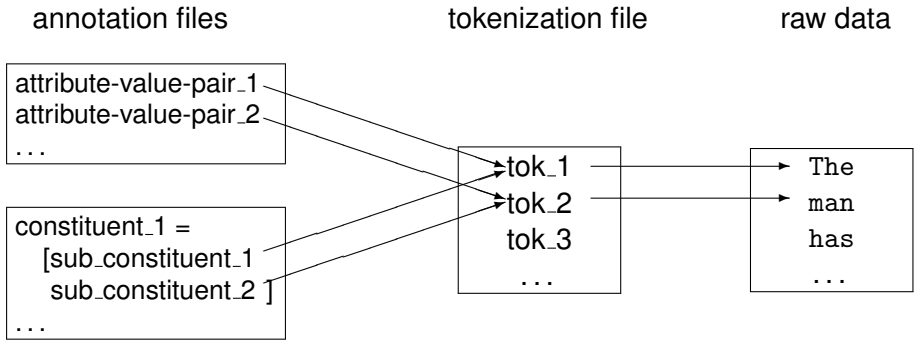

Figure 1: Architecture of standoff annotation (simplified)

existing corpus resources into our native format (LAF/GrAF). Powerful conversion tools are available already (e.g. Pepper, TIGERRegsitry, components of the NITE XML Toolkit), and new routines for corpus conversion can be developed with reasonable effort (cf. e.g. Ide & Suderman 2007 and the experience we have gained ourselves in the conversion of the Chintang corpus mentioned in Sect. 2.2.1).

Some challenges of linguistic annotation Our metalanguage requires a degree of expressivity which goes beyond the assignment of category labels to markables. For example, it is part of our research program to investigate the scopal behavior of operators such as negators within complex sentences, and we have to reckon with underspecified or ambiguous instances. While the scope of the (intonational) question operator in Chintang is unambiguous in the discourse context of (1) above, the following example is ambiguous between a question about the walking or about the talking:

(3) *ko-saŋa hand-a-ciy-e?* [CLLDCh1R03S04.021]
roam-CONVERB chat-PAST-by.oneself-PAST
'Did she chat by herself while roaming about?' *or* 'Did she roam about when chatting by herself?'

There are two general methods of representing underspecified structures: (i) an additional (virtual) node dominating the alternative structures is introduced (cf. Ide & Suderman 2007), or (ii) an additional type of annotation is added which specifies the options of attachment between pairs of nodes. The latter solution has been proposed by Kountz et al. (2008), who introduce an extension to the LAF/GrAF scheme. Structural alternatives are represented in the form of 'constraint lists', in analogy to other types of annotations. To what extent this approach provides a workable solution for scope ambiguities remains to be determined in our project work.

Another challenge of corpus annotation concerns the encoding of non-local dependencies such as extractions out of finite clauses (cf. also Sect. 3.2.3), as in the following:

(4) What$_i$ do you think [$_S$ that he said t$_i$]?

There are again two ways of dealing with non-local dependencies in corpora. First, additional ('virtual') markables may be inserted in the position of the 'trace'. This has been done in the

Penn Treebank. Traces can then be co-indexed with the corresponding 'fillers' via a common co-reference link. A second way to deal with such non-local dependencies is to allow intersecting edges in constituent structure, as is done in TIGER XML. In this case, (4) can be represented without a trace by simply assuming that *said* and *what* form a (discontinuous) VP. While this is an intriguing solution, it runs into problems in cases where one filler corresponds to more than one gap, as in the case of 'parasitic gaps'. In our case studies (cf. Sect. 3.2.3), we will explore these alternative ways of capturing non-local dependencies in corpus annotation, i.e. the evaluation of the (dis)advantages of each method is part of our research program.

Development of annotation tools We will develop software that can be used for the creation and annotation of corpora in a format as specified above. Comprehensive annotation guidelines will be provided to ensure consistency. The software should not only provide a user-friendly graphical interface facilitating structural and classificatory annotations, but also allow for the input of meta-annotations such as confidence rates, as we will standardly conduct reliability tests in the annotation process. The software will be developed in close cooperation with colleagues from the Sfb 632 (esp. Anke Lüdeling and Manfred Stede) as well as with Florian Zipser, the developer of Salt. It will be made freely available as soon as it is functional. As we will use the ANNIS database developed at the Sfb 632 for data retrieval and visualization, we do not plan to develop another query tool in Phase I. Depending on our experiences with ANNIS in Phase I, we may envisage this however for Phase II. Given that ANNIS is distributed under the Apache Public license, it is also conceivable that we enhance ANNIS, if possible in cooperation with our colleagues at the Sfb 632.

Parts of the annotation will be created interactively (cf. Brants & Plaehn 2000). In 'interactive corpus annotation', annotations are proposed by the parser (on the basis of a training set) and either accepted or modified (at each particular instance) by the annotator. Relevant tools are freely available, e.g. TnT, a statistical part-of-speech tagger (cf. Brants 2000). The *Annotate* tool comprises a statistical parser based on *Cascaded Markov Models* which also allow the generation of structural annotations (cf. Brants 1999). Depending on the success of these methods, we envisage a further development of the relevant tools for Phase II, together with P6 (Cysouw/Quasthoff).

Development of a construction-based database As pointed out in Sect. 2.1.5, we aim at unifying the analysis of textual data and abstractions made in (classical) typological databases by equating the 'constructions' of a database with sets of annotations as specified in a query (hence the notion of 'emergent database'). This method is useful for mining the corpora in search of probabilistic generalizations. However, generalizations over abstract constructions will also be stored in a more robust (persistent) form, i.e. in a classical typological database. The database will serve as a control set for further corpus explorations in Phase II and at the same time as a platform of cooperation within the research unit and as a means of publishing our data on the internet. The system will take the already existing database described by Bickel (in press-b) as a basis but will be further developed in cooperation with the individual projects of the research unit. Technically, the database will be implemented using a MySQL database management system which is accessed by PHP pages. The database system will also comprise a catalogue of attributes and values (cf. above) which serves as a backbone for the standardization of linguistic annotation within the research unit.

3.2.3 Case studies on non-local dependencies

Syntax As noted above, non-local dependencies are frequent characteristics of clause linkage exemplars in many languages but they pose particular challenges for annotation systems. We therefore plan in-depth case studies answering these challenges. Among the syntactic non-local dependencies most frequently discussed are 'filler-gap dependencies' (e.g. Fodor 1978, 1989, Hawkins 1999, 2004) and 'long-distance anaphora' (e.g Koster & Reuland 1991, Giorgi 2007). The term 'filler-gap dependency' refers to syntactic configurations where some element occurs in a position other than the structural position where it is interpreted, as in (4) above or in the range of phenomena associated with 'raising' and related constructions. Long-distance anaphora are pronominal elements that are 'referentially dependent' (cannot refer on their own), but that do not have an antecedent within a 'local domain' (cf. Gast 2006c for a typology of referential independence). In (post)structuralist linguistics, long-distance anaphora were first studied in detail for African languages under the label 'logophoric pronouns' (e.g. Hagège 1974, Clements 1975), but they are also well known from the 'classical' grammar tradition, i.e. Latin and Greek grammatico-graphy, and they bear as yet little-understood similarities to switch-reference markers.

Non-local dependencies have played a prominent role in generative linguistics, where they have been regarded as an interesting set of problem, with topics like the constraints governing this type of dependency (starting with Ross 1967), their modeling in specific syntactic theories (see Alexiadou et al. forthcoming for a recent contribution) and the question of 'complexity' (cf. Hawkins 1999, 2004) figuring centrally. With the exception of processing-related research along the lines of Hawkins (1999, 2004), long-distance dependencies have mostly been studied with respect to categorical constraints such as 'boundary nodes' or the 'distance' between the filler and the gap in a tree structure. Much less attention has been paid to probabilistic matters such as lexical or contextual biases, i.e. non-grammatical factors determining the (un)likelihood for a non-local dependency to be established.

We will investigate such contextual conditions on the basis of richly annotated corpora for the following phenomena:

1. non-local *self*-forms in English (cf. Baker 1995, König & Siemund 2000, Gast 2004);
2. cross-clausal extraction in English (cf. Hawkins 1999, 2004);
3. long-distance anaphora in Latin (cf. Benedicto 1991);
4. raising in Chintang (Bickel & Nichols 2001, Bickel 2008b, Bickel et al. in press).

The basic hypothesis guiding these analyses will be that non-local dependencies are correlated with the type of clause linkage. In particular, in better-known languages adjunct clauses seem to be 'more distant' from their main clauses than complement clauses. In many cases where non-local dependencies may hold between an element from a complement clause and an element from the matrix clause, this is not possible with adjunct clauses. For example, English regularly allows extraction out of (finite) complement clauses while extraction out of finite adjunct clauses is not possible (cf. Hawkins 1986, König & Gast 2009:Ch. 12). Similarly, long-distance anaphora in Latin are mostly found in complement clauses rather than adjunct clauses. However, such associations are not universal (cf. Bickel in press-b for counter-examples) and we expect them to differ across lexical elements in the matrix clause and other contextual factors (e.g. Givon 1980, Cristofaro 2003), with usage aspects being mirrored in quantitative patterns.

Semantics The scope of certain operators is typically restricted by specific boundary nodes. For example, it is often assumed that focus particles may not take scope beyond the finite clause they are contained in (cf. for instance König 1993, Gast & van der Auwera forthcoming for discussion). However, there are instances where operators (either apparently or actually) do take scope beyond their (finite) host clause. For instance, the scalar additive operator *even* is sometimes analyzed as taking wide scope in examples like (5) (cf. the scoping in (5a)). Alternatively, *even* can be assumed to be ambiguous, and the type of *even* occurring in (5) is regarded as a negative polarity item (cf. Gast & van der Auwera forthcoming for an overview of the discussion).

(5) Every student [$_S$ who even [looks]$_F$ at me] will get into trouble.
   a. EVEN [Every student [$_S$ who [looks]$_F$ at me]] will get into trouble].
   b. Every student [$_S$ who$_i$ EVEN$_{\text{NEG}}$ [t$_i$ [looks]$_F$ at me]] will get into trouble.

While occurrences of *even* of the type illustrated in (5) are often said to be licensed in 'downward entailing' contexts (cf. Ladusaw 1979), little is known about the exact contextual conditions governing their distribution. Moreover, *even* (in some contexts) competes with *so much as*, which could substitute for *even* in (5). We suspect that the notion of 'downward entailing' is too coarse-grained for a precise distributional characterization of (different types of) *even* and its competitors, and that a thorough corpus study will bring to light contextual (lexical) biases which have not so far been noticed. These case studies will be carried out for English, using the (enhanced version of) the ICE-GB corpus.

3.2.4 Organization plan

The project work falls into four work packages, with the following responsibilities (bold-faced names: lead responsible) and deliverables at the end of the three-year funding period:

	Description	*Team*	*Deliverable*
WP1	Developing and evaluating the system of variables	**Bickel**, PhD student 1	Dissertation 1
WP2	Annotation of exemplar samples	**Bickel**, entire team	Model corpora
WP3	Case studies of non-local dependencies	**Gast**, PhD student 2	Dissertation 2
WP4	Development of the technical infrastructure	**Gast**, computer scientist	Database and programs

References

Aijmer, K., 2008. Parallel and comparable corpora. In Lüdeling & Kytö (2008), 275–292.

Alexiadou, A., T. Kiss, & G. Müller (eds.), forthcoming. *Local Modelling of Non-Local Dependencies*. Tübingen: Niemeyer.

Baker, C., 1995. Contrast, discourse prominence, and intensification. *Language* 71, 63–101.

Benedicto, E., 1991. Latin long-distance anaphora. In Koster & Reuland (1991), 171–184.

Bickel, B., 2003. Referential density in discourse and syntactic typology. *Language* 79, 708 – 736.

Bickel, B., 2007. Typology in the 21st century: major current developments. *Linguistic Typology* 11, 239 – 251.

Bickel, B., 2008a. A general method for the statistical evaluation of typological distributions. Ms. University of Leipzig, [http://www.uni-leipzig.de/~bickel/research/papers/testing_universals_bickel2008.pdf].

Bickel, B., 2008b. Grammatical relations in Chintang. Work In Progress Presentation, MPI for Evolutionary Anthropology, Leipzig, October 7, 2008 [http://http://www.uni-leipzig.de/~bickel/research/presentations/GR_chintang.pdf].

Bickel, B., in press-a. Absolute and statistical universals. In Hogan, P. C. (ed.) *The Cambridge Encyclopedia of the Language Sciences*. Cambridge: Cambridge University Press [preprint available at http://www.uni-leipzig.de/~bickel/research/papers/universals_cels_bb.pdf].

Bickel, B., in press-b. Capturing particulars and universals in clause linkage: a multivariate analysis. In Bril, I. (ed.) *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*. Amsterdam: Benjamins [preprint available at http://www.uni-leipzig.de/~bickel/research/papers/multivariate.clauselinkage.bickel2009.pdf].

Bickel, B. & J. Nichols, 2001. Syntactic ergativity in light verb complements. *Proceedings of the 27th Annual Meeting of the Berkeley Linguistics Society* .

Bickel, B., M. Rai, N. Paudyal, G. Banjade, T. N. Bhatta, M. Gaenszle, E. Lieven, I. P. Rai, N. K. Rai, & S. Stoll, in press. The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti). In Malchukov, A., M. Haspelmath, & B. Comrie (eds.) *Studies in ditransitive constructions: a comparative handbook*. Berlin: Mouton de Gruyter [preprint available at http://www.uni-leipzig.de/~bickel/research/papers/ctn_ditrans_bickeletal.pdf].

Bickel, B., S. Stoll, M. Gaenszle, N. K. Rai, E. Lieven, G. Banjade, T. N. Bhatta, N. Paudyal, J. Pettigrew, I. P. Rai, & M. Rai, 2010. *Audiovisual corpus of the Chintang language, including a longitudinal corpus of language acquisition by six children*: ca. 520,000 words transcribed and translated, of which ca. 400'000 glossed and translated, plus paradigm sets and grammar sketches, ethnographic descriptions, photographs. *DOBES Archive*, http://www.mpi.nl/DOBES.

Bird, S. & M. Liberman, 2001. A formal framework for linguistic annotation. *Speech Communication* 33, 23–60.

Brants, T., 1999. *Tagging and parsing with cascaded Markov Models – Automation of corpus annotation*. Ph.D. thesis, Universität des Saarlandes.

Brants, T., 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixths Applied Natural Language Processing Conference ANLP-2000, April 29–May 3, Seattle, WA*.

Brants, T. & O. Plaehn, 2000. Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000, May 31–June 2, 2000*. Athens.

Bryant, D. & V. Moulton, 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21, 255 – 265.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, & S. Pado, 2006. SALTO – a versatile multi-level annotation tool. In *Proceedings of LREC 2006*.

Clements, G. N., 1975. The logophoric pronoun in Ewe: Its role in discourse. *Journal of West African Languages* 10, 141–177.

Cristofaro, S., 2003. *Subordination*. Oxford: Oxford University Press.

Croft, W., 2000. *Explaining language change: an evolutionary approach*. Harlow: Longman.

Croft, W. & K. T. Poole, 2004. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. Ms., Center for Advanced Studies in the Behavioral Sciences, Stanford.

Cysouw, M., 2005. Quantitative methods in typology. In Altmann, G., R. Köhler, & R. Piotrowski (eds.) *Quantitative linguistics: an international handbook*, 554 – 578. Berlin: Mouton de Gruyter.

Cysouw, M., 2007. Building semantic maps: the case of person-marking. In Miestamo, M. & B. Wälchli (eds.) *New challenges in typology: broadening the horizons and redefining the foundations*, 225–248. Berlin: Mouton de Gruyter.

Diessel, H., 2004. *The acquisition of complex sentences*, *Cambridge studies in linguistics*, vol. 105. Cambridge: Cambridge University Press.

Dimitriadis, A., 2009. Designing linguistic databases: A primer for linguists. In Everaert et al. (2009), 13–75.

Dryer, M. S., 1998. Why statistical universals are better than absolute universals. *Papers from the 33rd Annual Meeting of the Chicago Linguistic Society* 123 – 145.

DuBois, J. W., 1987. The discourse basis of ergativity. *Language* 63, 805 – 855.

DuBois, J. W., L. E. Kumpf, & W. J. Ashby (eds.), 2003. *Preferred argument structure: grammar as architecture for function*. Amsterdam: Benjamins.

Erk, K. & S. Pado, 2004. A powerful and versatile XML Format for representing role-semantic annotation. In *Proceedings of LREC-2004*. Lisbon.

Everaert, M., S. Musgrave, & A. Dimitriadis (eds.), 2009. *The Use of Databases in Cross-Linguistics Studies*. Empirical Approaches to Language Typology, Berlin: Mouton de Gruyter.

Evert, S., J. Carletta, T. J. O'Donnel, J. Kilgour, A. Vögele, & H. Voormann, 2003. The NITE Object Model. URL `http://www.ltg.ed.ac.uk/NITE/documents/NiteObjectModel.v2.1.pdf`.

Fenk-Oczlon, G. & A. Fenk, 1985. The mean length of propositions is 7 plus/minus 2 syllables – but the position of languages within this range is not accidental. In D'Ydewalle, G. (ed.) *Proceedings of the 23rd International Congress of Psychology: selected papers, vol. 3*, 355 – 359. Amsterdam: North Holland.

Fodor, J., 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry* 9, 427–73.

Fodor, J., 1989. Empty categories in sentence processing. *Language and Cognitive Processes* 3, 155–209.

Gaenszle, M., B. Bickel, J. Pettigrew, A. Rai, S. K. Rai, & N. P. Sharma (Gautam), 2010. Binomials and the noun/verb ratio in Puma Rai ritual speech. Ms. University of Vienna.

Gast, V., 2004. The interpretation of logophoric *self*-forms, and some consequences for a theory of reference and denotation. In Branco, A., T. McEnery, & R. Mitkov (eds.) *Proceedings of the Fifth Discourse Anaphora and Anaphora Resolution Colloquium DAARC*, 75–80. Lisbon: Ediçôes Colibri.

Gast, V., 2006a. The distribution of *also* and *too* – a preliminary corpus study. In Gast (2006d), 163–176. Special issue of Zeitschrift für Anglistik und Amerikanistik, 54.2.

Gast, V., 2006b. Focus particles. In Brown, K. (ed.) *The Encyclopaedia of Language and Linguistics*, vol. 4, 518–519. Oxford: Elsevier.

Gast, V., 2006c. *The Grammar of Identity. Intensifiers and Reflexives in Germanic Languages*. Routledge.

Gast, V. (ed.), 2006d. *The Scope and Limits of Corpus Linguistics – Empiricism in the Description and Analysis of English*. Würzburg: Königshausen und Neumann. Special issue of Zeitschrift für Anglistik und Amerikanistik, 54.2.

Gast, V., 2009. A contribution to 'two-dimensional' language description: The typological database of intensifiers and reflexives. In Everaert et al. (2009), 209–234.

Gast, V., forthcoming. Contrastive analysis: Theories and methods. In Kortmann, B. & J. Kabatek (eds.) *Linguistic Theory and Methodology*. Dictionaries of Linguistics and Communication Science, Berlin: Mouton de Gruyter.

Gast, V. & J. van der Auwera, 2010. Vers une typologie des opérateurs additifs scalaires. In Haderman, P. & O. Imkova (eds.) *Approches de la Scalarité*, 226–247. Genève: Droz.

Gast, V. & J. van der Auwera, forthcoming. Scalar additive operators in the languages of Europe. *Language* .

Giorgi, A., 2007. On the nature of long-distance anaphors. *Linguistic Inquiry* 38, 321–342.

Givon, T., 1980. The Binding Hierarchy and the typology of complements. *Studies in Language* 4, 333–377.

Givón, T. (ed.), 1983. *Topic continuity in discourse: a quantitative cross-language study*. Amsterdam: Benjamins.

Greenberg, J. H., 1959. A quantitative approach to morphological typology. *International Journal of American Linguistics* 26, 178 – 194.

Hagège, C., 1974. Les pronoms logophoriques. *Bulletin de la Societé de Linguistique de Paris* 69, 287–310.

Haspelmath, M., 2008. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery* 6, 40–63.

Hawkins, J., 1986. *A Comparative Typology of English and German – Unifying the Contrasts*. London: Croom Helm.

Hawkins, J., 1999. Processing complexity and filler-gap dependencies across languages. *Language* 75, 244–285.

Hawkins, J., 2004. *Efficiency and Complexity in Grammars*. Cambridge: Cambridge University Press.

Hawkins, J. A., 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Huson, D. H. & D. Bryant, 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 32, 254 – 267.

Ide, N. & L. Romary, 2006. Representing linguistic corpora and their annotations. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy.*

Ide, N. & K. Suderman, 2007. GrAF: A Graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007, Prague, June 28-29*, 1–8.

Janssen, D., B. Bickel, & F. Zúñiga, 2006. Randomization tests in language typology. *Linguistic Typology* 10, 419 – 440.

Johansson, S., 1998. On the role of corpora in cross-linguistic research. In Johansson, S. & S. Oksefjell (eds.) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*, 3–24. Amsterdam/Atlanta: Rodopi.

Justeson, J. & L. Stephens, 1990. Explanation for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists*. Berlin: Mouton de Gruyter.

Jäger, G., 2007. Maximum entropy models and stochastic entropy theory. In Zaenen, A., J. Simpson, T. H. King, J. B. Grimshaw, J. Maling, & C. D. Manning (eds.) *Architectures, rules, and preferences: variations on themes by Joan W. Bresnan*, 467 – 479. CSLI.

Koenig, E. & V. Gast, 2002. Reflexive pronouns and other uses of *self*-forms in English. *Zeitschrift für Anglistik und Amerikanistik* 50, 225–238.

Kortmann, B. & B. Szmrecsanyi, 2009. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119, 1643–1663.

Koster, J. & E. Reuland (eds.), 1991. *Long-Distance Anaphora*. Cambridge: Cambridge University Press.

Kountz, M., U. Heid, & K. Eckart, 2008. A LAF/GrAF based encoding scheme for underspecified representations of syntactic annotations. In *Proceedings of the Language Resources & Evaluation Conference 2008, LREC2008, Marrakech, Morocco, Mai 2008*.

König, E., 1993. Focus particles. In Jacobs, J. & A. von Stechow (eds.) *Syntax. Ein internationales Handbuch zeitgenössischer Forschung*, *Handbücher der Sprach- und Kommunikationswissenschaften*, vol. 9, 978–987. Berlin: Mouton de Gruyter.

König, E. & V. Gast, 2009. *Understanding English-German Contrasts*. Berlin: Erich Schmidt, 2nd edn.

König, E. & P. Siemund, 2000. Locally free *self*,-forms, logophoricity and intensification. *English Language and Linguistics* 4, 183–204.

Ladusaw, W., 1979. *Polarity sensitivity as inherent scope relations*. Ph.D. thesis, University of Texas, Austin.

Lehmann, C., 1988. Towards a typology of clause linkage. In Haiman, J. & S. A. Thompson (eds.) *Clause combining in grammar and discourse*, 181 – 226. Amsterdam: Benjamins.

Lezius, W., 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis, University of Stuttgart.

Lüdeling, A. & M. Kytö (eds.), 2008. *Corpus Linguistics*, *Handbooks of Linguistics and Communication Science*, vol. 1. Berlin: Mouton de Gruyter.

McEnery, T., R. Xiao, & Y. Tono, 2006. *Corpus-based Language Studies: An Advanced Source Book*. Routledge Applied Linguistics, London: Routledge.

Müller, C. & M. Strube, 2006. Multi-level annotation of linguistic data with MMAX2. In Braun, S., K. John, & J. Mukherjee (eds.) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, 197–214. English Corpus Linguistics, Frankfurt: Peter Lang.

Nichols, J., 1992. *Linguistic diversity in space and time*. Chicago: The University of Chicago Press.

Ostler, N., 2008. Corpora of less studied languages. In Lüdeling & Kytö (2008), 457–483.

Plaehn, O. & T. Brants, 2000. Annotate – an efficient interactive annotation tool. In *Proceedings of ANLP-2000*. Seattle, WA. URL `http://www.coli.uni-saarland.de/publikationen/softcopies/Plaehn:2000:AEI.pdf`.

R Development Core Team, 2010. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, `http://www.r-project.org`.

Ross, J. R., 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT, Cambridge, MA.

Schackow, D., B. Bickel, S. K. Rai, N. P. Sharma (Gautam), A. Rai, & M. Gaenszle, in press. Morphosyntactic properties and scope behavior of 'subordinate' clauses in Puma (Kiranti). In Gast, V. & H. Diessel (eds.) *Clause-combining in cross-linguistic perspective*. Berlin: Mouton de Gruyter [preprint available at `http://www.uni-leipzig.de/~autotyp/download/schackowetal2009puma.pdf`].

Schmidt, T. & K. Wörner, 2009. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In *Corpus-based pragmatics*, vol. 19.

Stoll, S., B. Bickel, E. Lieven, G. Banjade, T. N. Bhatta, M. Gaenszle, N. P. Paudyal, J. Pettigrew, I. P. Rai, M. Rai, & N. K. Rai, 2009. Nouns and verbs in Chintang: children's usage and surrounding adult speech. Ms. Max-Planck-Institute for Evolutionary Anthropology, [`http://www.eva.mpg.de/lingua/staff/stoll/pdf/Nouns&Verbs_Chintang.pdf`].

Tomasello, M., 2003. *Constructing a language: a usage-based theory of language acquisition*. Cambridge, Mass.: Harvard University Press.

Wälchli, B., 2007. Advantages and disadvantages of using parallel texts in typological investigations. *Sprachtypologie und Universalienforschung* 60, 118–134.

Xiao, R., 2008. Well-known and influential corpora. In Lüdeling & Kytö (2008), 383–457.

Zipser, F. & L. Romary, 2010. A model oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010, Malta*.

4 Beantragte Mittel

4.1 Personalkosten

Für die Durchführung des Projektes benötigen wir zwei Doktoranden mit linguistischer Ausbildung (TV-L E13/2), eine/n Informatiker/in mit linguistischem Hintergrundwissen (TV-L E13/2) und fünf studentische Hilfskräfte à 40h/Monat. Die linguistischen Doktoranden werden hauptsächlich mit der Entwicklung eines Variablensystems (WP1) und den Fallstudien zu nicht-lokalen Abhängigkeiten (WP3) befasst sein. Der/die Informatiker/in wird für die Entwicklung der Annotationssoftware und der Datenbank verantwortlich sein (WP4). Hierbei handelt es sich weitgehend um konzeptionelle Aufgaben, die Vertrautheit sowohl mit aktuellen Problemen der Annotation (z.B. Mehrebenenannotationen) als auch mit dem Design und Management von Datenbanksystemen und linearen Datenstrukturen voraussetzen. Für die Erstellung spezifischer Skripte werden außerdem Mittel für Werkverträge benötigt, die flexibel eingesetzt werden können (siehe Abschn. 4.4 'Sonstige Kosten').

Weiterhin benötigen wir insgesamt fünf Hilfskräfte, davon eine für WP1 (Entwicklung des Variablensystems), drei für die Annotation von Korpora (WP2) und eine für die Entwicklung einer technischen Infrastruktur (WP4). Die Annotation von Korpora (WP2) ist sehr zeitaufwendig und eignet sich hervorragend für eine qualifizierende Beschäftigung von Studierenden. Annotationsaufgaben lassen sich gut mit Abschlussarbeiten (BA, MA) verbinden, so dass fortgeschrittene Studierende die Projektarbeit mit ihrem Studium sinnvoll verbinden können. (Wir haben mit Annotationen durch Hilfskräfte in anderen Projekten zum Chintang exzellente Erfahrungen gemacht.)

Anzahl	Arbeitszeit	Vergütung	Qualifikation	Arbeitspaket	Arbeitsplatz
1	50%	TV-L E13	Linguist/in	WP1	Leipzig
1	50%	TV-L E13	Linguist/in	WP3	Jena
1	50%	TV-L E13	Informatiker/in	WP4	Jena
1	10h/Woche	student. Hilfskraft		WP1	Leipzig
3	10h/Woche	student. Hilfskraft		WP2	Leipzig
1	10h/Woche	student. Hilfskraft		WP4	Jena

4.2 Wissenschaftliche Geräte

Wir benötigen einen Laptop für jeden Projektmitarbeiter, einen Desktop-Computer für Softwareentwicklung und Datenbankmanagement (mehrfache Betriebssysteme) sowie einen Server. Der Server wird im Rechenzentrum der Friedrich-Schiller-Universität Jena aufgestellt und auch von diesem gewartet werden.

3	Laptops à €2000	€ 6 000
1	Desktop à €2500	€ 2 500
1	Server (Hardware) à €4500 (z.B. HP DL380G6)	€ 4 500
3	Server Software/Wartung für 3 Jahre à €1000	€ 3 000
Gesamt:		€ 16 000

4.3 Reisen

Gemäß den Vorgaben bei der Beantragung einer Forschergruppe beantragen wir jährliche Pauschalen für die Teilnahme an Konferenzen:

1	Projektpauschale à €1750	für 3 Jahre	€ 5 250
3	Mitarbeiterpauschalen à €500	für 3 Jahre	€ 4 500
Gesamt:			€ 9 750

4.4 Sonstige Kosten

Für die Entwicklung von Software für die Annotation von Korpora benötigen wir Sachmittel, die flexibel für Werkverträge eingesetzt werden können. Diese Arbeiten können teilweise von Studierenden der Informatik durchgeführt werden, verlangen z.T. aber auch spezifische Qualifikationen. Wir gehen von einem durchschnittlichen Stundenlohn von €30 aus und veranschlagen für die Entwicklung von Skripten 1000 Stunden für den ersten Antragszeitraum:

Mittel für die Erstellung von Werkverträgen für Skripte	€ 30 000
Sachmittel gesamt	**€ 55 750**

5 Voraussetzungen für die Durchführung des Vorhabens

5.1 Zusammensetzung der Arbeitsgruppe (Composition of the research team)

Apart from the two PIs, the two linguistic PhD students and the computer scientist the team includes three Ph.D. students at the University of Leipzig, who work on Chintang or related languages: Netra P. Paudyal (working on a descriptive grammar of Chintang, DAAD scholarship), Robert Schikowski (working on differential argument coding in Chintang and Nepali, EuroBABEL/DFG grant), and Diana Schackow (working on a descriptive grammar of Yakkha and on clause linkage in Puma, Sächisches Landesstipendium).

5.2 Zusammenarbeit mit anderen Wissenschaftlerinnen und Wissenschaftlern (Cooperation)

The project will proceed in close collaboration with all projects in the development of the system of variables. While the test suite that we use in the project ensures that the system will be compatible with a wide variety of languages, we will seek the groups' typological expertise to strengthen this compatibility. In the work on Latin, we will collaborate closely with P4 (Lühr/Zeilfelder), in the work on English with P5 (Diessel) and P3 (Gast/Schäfer). Especially towards the end of Phase I, and even more so in Phase II, we will work most closely with P6 (Cysouw/Quasthoff).

Outside the research unit we will collaborate closely with our colleagues at the Sfb 632 (Berlin/Potsdam), especially Anke Lüdeling, Amir Zeldes (HU Berlin) and Manfred Stede (Potsdam). We are also in contact with Florian Zipser (HU Berlin), who has developed a data model and a conversion tool for corpus annotations (cf. Sections 2.1.4 and 3.2.2).

5.3 Interessenkonflikte bei wirtschaftlichen Arbeiten

Keine

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.