Final draft of a chapter for *Clause Linkage in Cross-Linguistic Perspective*, eds. Volker Gast and Holger Diessel. Berlin: De Gruyter Mouton. 2012.

The Performance Basis of Grammatical Constraints on Complex Sentences: *A Preliminary Survey*

KARSTEN SCHMIDTKE-BODE
Friedrich Schiller University Jena and University of Cambridge[1]

Usage-based approaches to syntactic theory suggest that grammatical systems are best conceived of as abstractions from individual usage events, such that preferred structural choices in language performance are conventionalized into productive grammatical patterns, regularities and constraints. Consequently, morphosyntactic universals, as inferred from the comparison of synchronic grammatical descriptions, are predicted to be firmly rooted in language performance. The advent of rich and systematic corpus data for some languages opens up the possibility to substantiate this hypothesis on a sound empirical basis, and complex sentences provide a particularly wide range of constructions and phenomena amenable to such a research programme. The present paper surveys major previous contributions in this area and provides new data or perspectives on a variety of pertinent construction types, including relativization, purpose, avertive and complement clauses, as well as converbal constructions.

1. The Performance-Grammar Correspondence Hypothesis

The last two decades of linguistic research have witnessed a dramatic increase in studies that emphasize the usage-based nature of language, with regard to both individual linguistic representations and grammatical systems at large (cf. Bybee 2006 for an overview). On this view, grammatical categories, patterns and constraints are emergent phenomena that are abstracted from recurrent usage events. Typically, speakers can select, on any such event, from a variety of grammatical constructions

[1] This paper was written during my stay at the *Research Centre of English and Applied Linguistics* at the University of Cambridge/UK (2008-2009). I would like to express my gratitude to John Hawkins, Henriëtte Hendriks and John Williams for providing a stimulating research environment and several opportunities for discussing the theoretical (i.e. usage-based) foundations that this paper rests on. I am also grateful to Holger Diessel, Katja Hetterle, and the audiences at the *Leipzig Spring School 2008* and the *Cambridge Tuesday Colloquium* for invaluable feedback and suggestions on oral versions of this paper. Special thanks go to Bernd Kortmann and Daniel Wiechmann for sharing their original data and generously allowing me to reproduce it in altered formats. Needless to say, none of them is to blame for any errors in presenting and interpreting the results; I alone claim responsibility for any shortcomings. Author's correspondence address: <karsten.schmidtke@uni-jena.de>.

that provide a sufficient match to the specific conceptualization of the experience to be conveyed. In this way, variation is permanently generated and immanent in the way grammatical structures are used in actual discourse (cf. Croft 2010). At the same time, however, variation is kept in check by cognitive and communicative pressures, and it is these factors that lead to the more successful spread (and hence more frequent use) of one variant at the expense of a competing one. What emerges from language use, then, is a set of preferred discourse patterns, a conventionalized machinery of well-oiled constructions, and specific constraints on their distribution and applicability. This intrinsic link between patterns of language use and properties of grammatical systems is at the heart of the 'Performance-Grammar-Correspondence Hypothesis' (henceforth PGCH), fully articulated in Hawkins (2004). According to this view, one would expect to find systematic correspondences between structural constraints or conventionalized syntactic patterns in grammars, on the one hand, and preferred choices in the performance of languages that license several structural variants of a given phenomenon, on the other.

Over the last decades, quite a few such correspondences have been established. In fact, some of the classic typological generalizations, such as Greenberg's (1963) markedness hierarchies, Keenan and Comrie's (1977) work on relative clauses, or Du Bois' (1987) research on 'preferred argument structure', do not only incorporate intralinguistic variation, but have also been linked explicitly to frequency distributions within individual languages (in the sense that grammatical hierarchies on the above phenomena, for example, are argued to be conventionalizations of frequency rankings in language use). However, it is not until recently that this agenda has started to be pursued in a large-scale and more systematic fashion. Undoubtedly, apart from the increasing prominence of the usage-based approach as a theoretical framework, one reason for this development is the new availability of electronic corpora and more extensively gathered and annotated corpus material from field work on individual languages. This opens up the possibility to investigate significant quantitative distributions of linguistic phenomena in great detail, and study them against the backdrop of relevant findings from language typology. In this way, interesting new perspectives have been provided on phenomena as diverse as referential density in discourse (cf. Bickel 2003), person-voice interactions (cf. the ungrammaticality of 3>1 aligned active sentences such as *The man knows me* in Salishan languages, and their statistical dispreference in corpora of English, cf. Bresnan et al. 2001), ditransitive alignment splits (cf. Haspelmath 2007), or the analysis of linear order (Hawkins [1994, 2004], for example, observes that languages with considerable leeway in constituent order still show significant performance preferences for the orders that are most pervasively grammaticalized across the world's languages, notably SVO and SOV). A wide phenomenological array of performance-grammar correspondences can also be

found in Haspelmath (2008a), who proposes a unified account of grammatical coding asymmetries in terms of differential corpus frequencies. Importantly, the convergence of patterns from typology and corpus linguistics can be beneficial to both disciplines. On the one hand, universal trends in language structure can be more plausibly motivated in cognitive-functional terms (e.g. processing pressures, economical behaviour, conceptual preferences) if they are studied at the level where they actually operate, i.e. on individual usage events. On the other hand, Bresnan (2007) argues compellingly that typological findings can provide important (and previously unnoticed) variables for the study of language-specific phenomena.

In this spirit, the present paper investigates performance-grammar correspondences in the domain of complex sentences. Complex-sentence systems are known to vary immensely across the world's languages, yet they exhibit interesting constraints that are often shared by unrelated and typologically diverse languages. For this survey, we will consider two important aspects of interclausal relationships. Section 2 is concerned with the position of subordinate clauses relative to the matrix clause, focussing on purpose, converbal and complement constructions. Section 3, by contrast, will discuss various argument-structural constraints in relative, complement and converbal clauses. The discussion is rounded off by an outlook, in section 4, on further domains of application, specifically semantic ones. Across all sections, the primary aim of the paper is to provide an overview of recent work on complex sentences from the perspective of the PGCH, surveying previous studies in this spirit and presenting new data from currently progressing research. It needs to be emphasized from the outset that the choice of case studies is necessarily selective (and subjective) and should, therefore, be taken to be a *preliminary* survey (as reflected by the title of the paper). Nevertheless, it aims at a certain degree of comprehensiveness by including all major functional types of complex sentences (i.e. adverbial, relative and complement clauses) and converging evidence from a variety of different methodological approaches.

2. Constraints on linearization in complex sentences

One of the best-known constraints on grammatical structures is their position relative to other elements in a larger construction. Greenberg's (1963) seminal typological paper on word order patterns has spawned an impressive amount of research in this domain. Most notably, correlations of word order across different types of phrases, including their exceptions, have now been documented extensively (e.g. Dryer 2005a, b) and have been subjected to different kinds of functional explanations (e.g. Hawkins 2004 for a unified processing account, Givón 2001: ch.5 for a diachronic approach). Although complex sentences have often been included in the set of constructions for which head-dependent ordering correlations are postulated, their linearization has

only recently begun to be investigated in a more systematic fashion. In fact, a currently progressing research project[2] aims at developing the first large typological database on the positioning patterns of all types of subordinate-clause constructions. It can be expected that such a comprehensive database will ultimately lead to a clearer, but also more differentiated picture of the ordering constraints in complex sentences across the world's languages. Crucially, however, the *motivations* for those constraints can be studied more thoroughly if the typological distributions are related to patterns of internal variation in languages that license several ordering options of a given construction. In this spirit, several recent studies have been devoted to particular types of complex sentences (cf. below). For this overview article, I will briefly discuss an illustrative case study from my own work on purpose clauses (section 2.1), and point to ongoing research on other construction types (section 2.2).

2.1 Linearization of purpose clause constructions

In Schmidtke-Bode (2009), I present the first monographic study of the typology of purpose clauses, based on a genetically and geographically dispersed variety sample of 80 languages. By way of introduction, (1) provides an example of a preposed purpose clause from Ika (a Chibchan OV-language of Colombia), while (2) shows a postposed purpose clause from Tetun (an Austronesian VO-language of West Timor):

(1) Ika (Frank 1990: 107)
[*Kʌnɨ mus-ʌn-**guasi***], *mura ʌn-kʌ-taʔ-na.*
cane grind-IMPF-**PURP** mule REF-PERI-look.for-DIST
'He looked for the mule in order to grind sugar cane.'

(2) (Fehan) Tetun (van Klinken 1999: 317)
L'ao mai [***bat*** *ita atu hó malu hi'it ai.kanoik*].
walk come **PURP** 1PL.INCL IRR accompany each.other guess story
'Come here so that we'll tell riddles together.'

In his classic paper, Greenberg (1963) hypothesizes that examples like (1) and (2) above in fact reflect typical ordering patterns of purpose clauses in that such constructions would generally be found after their associated matrix clause except in OV languages, in which dependent elements are usually placed before the corresponding head ('Universal 15'). The larger and more principled sample underlying my study proved to corroborate Greenberg's claim. If we consider all distinct purposive *constructions* across the sample (N = 218), it turns out that they are overwhelmingly preferred in sentence-

[2] The project is entitled *Principles of Linear Ordering in Complex Sentence Constructions*. It involves a research group led by Holger Diessel at the University of Jena and is funded by the *German Research Foundation* (DFG).

final position, i.e. after the matrix clause (Fig. 1a). At the level of languages (N = 80), we can distinguish whether its purposive constructions are preferably preposed, postposed, flexible (readily allowing both options for each construction) or mixed (each construction has a different positioning pattern). At this level, too, postposing comes out as the preferred position for purpose clauses. As Fig. 1b shows, it is the unanimous pattern in VO languages and the dominant one in languages without a fixed basic constituent order (Fig. 1b). Crucially, we also find postposing as a majority pattern in OV languages, even though this ordering is non-harmonic under the predictions from (head-dependent) word-order correlations.

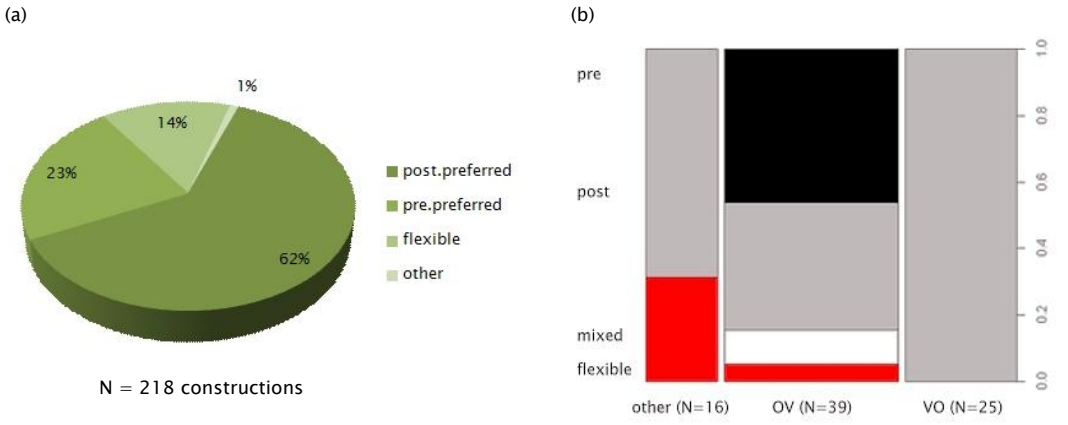


Figure 1. (a) Positioning patterns of purpose clause constructions (in percentages from N = 218) (b) Positioning patterns of purpose clauses across the sample of 80 languages, broken down by constituent order types (OV, VO, 'other'). The bar chart reflects the proportion of languages of each type in which purpose clauses are preferably preposed (black shading), postposed (grey), mixed (white) or flexible (red). The distribution is highly significant (randomised $\chi^2 = 44.29$, $p < .001$, B = 100,000, Cramer's $V = 0.526$), chiefly due to the large amount of postposing across the sample.[3]

It can also be shown that this distribution is found across all of Dryer's (1989) macro-areas (cf. Schmidtke-Bode 2009: §3.5.1), and that it even overrides correlations from the position of the subordinator (cf. Diessel 2001): Contrary to expectation, we find quite a few OV languages which postpose their purpose clauses even though the subordinator or 'linking' morpheme does not occur in between the two connected clauses. Given that such constellations have been claimed to be suboptimal from a processing perspective (cf. section 3.2 below for further details), there appear to be other factors that systematically override those from consistent branching directions and easy constituent recognition. In Greenberg's explanation for his 'Universal 15', these factors

[3] All statistical analyses presented in this paper were performed with the open-source software *R*, version 2.7.1 (*R Development Core Team* 2008).

boil down to iconicity: The postposed purpose clause diagrammatically mirrors the logical sequence of action > (intended) result. Though certainly parsimonious, this monofactorial account needs to be evaluated more carefully, especially because the notion of iconicity has very recently come under attack in the functional literature (e.g. Haspelmath 2008b).[4] To this end, it is worthwhile considering data from language performance; it is here that motivations for structural choices manifest themselves, and detailed studies of usage may bring to light which (relative) importance we attach to the concept of iconicity of sequence.

A recent corpus-linguistic study actually provides convincing evidence that for some types of complex sentences, iconic ordering plays an important role. Diessel (2008) performed a logistic regression analysis on English temporal clauses (*when, before, after, once, until* clauses). The merit of such multivariate statistical methods lies precisely in their ability to estimate the relative impact of certain independent variables on a structural choice, i.e. clause order. Step-wise regression procedures allowed Diessel to conclude that iconicity is a superior predictor variable since it accounts for a significant proportion of the variation in linear order. What is more, Diessel also shows that initial temporal clauses in English are significantly more often iconic than their final counterparts. Building on psycholinguistic work, Diessel assumes that iconic clause order in temporal relations reduces the processing cost of the complex sentence. Considering that preposed subordinate clauses are – on structural grounds – already associated with higher processing costs in VO languages (cf. Hawkins 1994), it seems that an additional processing load that would result from non-iconic clause order is strongly avoided. In order to keep processing cost in tolerable limits, then, preposed temporal clauses are preferably iconic. Postposed temporal clauses, by contrast, are readily processed in structural terms, so that they tolerate non-iconic ordering to a significantly higher degree.

This makes us confident that iconicity plays an important role, but for purpose clauses, unfortunately, iconicity is more problematic. This is because it can be (and has been) claimed that even preposed purpose clauses are iconic since the conceptualization of the purpose necessarily precedes the action (intention > action > result, cf. Hwang 1995). This move, however, makes iconicity vacuous for the present phenomenon; it simply shifts the burden of evidence to explaining why one of the two conceptualizations is grammaticalized much more often in the world's languages. Luckily, however, there are performance studies of a different kind that can be brought to bear on our typological findings. In particular, studies in (quantitative) discourse

[4] Although Haspelmath's (2008b) critique is not directed against iconicity of sequence, he makes an important general point about iconicity: He cautions us against blindly accepting that what may look iconic to the linguist is also iconically *motivated*. It is in this awareness that we look at performance studies that may corroborate the impact of iconic motivation or else suggest alternative factors for a given prevailing order.

analysis have suggested that the position of adverbial clauses is often a reflex of their discourse-pragmatic function and/or information-structural value (e.g. Chafe 1984 for a classic paper). Building on this line of research, I argue in Schmidtke-Bode (2009) that prototypical purpose clauses can be understood as the mirror image of 'scene-setting' adverbial clauses, which are preposed even in VO languages (and are tolerated in this position even with clause-initial conjunctions) because they lay out a thematic ground for the interpretation of the associated main clause. This commonly applies to conditional and many temporal adverbial constructions, which tend to provide topical (often presupposed) information. Purpose clauses are fundamentally different in this respect, and a discourse-analytical study by Thompson (1985) provides evidence *ex negativo* for this hypothesis.

Specifically, Thompson investigates pre- and postposed infinitival purpose clauses in various genres from written English discourse. She shows that the final purpose clause, i.e. the cross-linguistically unmarked type of purposive construction, is rigidly constrained to the (narrow) function of modifying the proposition in the preceding matrix clause. It provides the very motive for the main clause action and as such typically introduces new, focal information into the ongoing discourse. Its tight bond with the matrix clause is reflected in Thompson's study in the lack of commas at the left clause boundary and the fact that the postposed purpose clause only ever has its matrix, i.e. the immediately preceding clause, in its scope. Whenever purpose clauses are preposed in English (which is a perfectly grammatical variant), these rhematic characteristics do not apply anymore. Thompson finds that initial purpose clauses help "to *guide the attention of the reader,* by signalling, within the portion of text in which it occurs, how the reader is expected to associate the material following the purpose clause with the material preceding it" (Thompson 1985: 61, emphasis in original). Because of their discourse-organizing function, such clauses often have a much wider scope (average number of clauses in scope = 3.8) than final purpose clauses (average scope = 1.0 clauses). These findings suggest that it is, in fact, the discourse-pragmatic status of purpose clauses that strongly favours their sentence-final position, even in languages that generally prepose adverbial clauses.

Indeed, if we look at the typological data from this perspective, we find a number of phenomena that tie in with Thompson's proposal. I refrain from reproducing the full list here, but it is particularly instructive to see, for instance, that some languages open their focal constructions (such as cleaving) only to purpose, but not other adverbial relations, or to find that many OV languages postpose both causal and purpose clauses (while keeping other adverbial clauses sentence-initial). This is a relevant finding since for causal (i.e. cause-effect or reason-result) clauses, iconicity cannot be invoked as a postposing factor. It rather seems that the common information-structural value underlying both causal and purpose clauses (foregrounding, focussing) leads to their

peculiar positioning patterns. A recent contribution by Diessel and Hetterle (to appear) systematically relates discourse-pragmatic studies on causal clauses from a variety of languages to a large typological sample of the structural properties of such clauses. They, too, find a close match between performance and grammatical data, which confirms the necessity (i) to bring detailed language-specific material to bear on typological distributions, and (ii) to examine the relative weight of alleged motivations. Iconicity of sequence, in this light, does not appear to be the only, and perhaps not even the major factor underlying the linearization of purpose clauses. The scenario is probably best captured in terms of Hawkins (2004: 223), who claims that structural asymmetries in performance and grammar (such as the right-left asymmetry in the position of purpose clauses) result from multiple motivations working into the same direction, reaffirming each other.

2.2 Linearization in converbal and complementation constructions

Before we leave this section, we shall take a brief look at other types of complex sentences for which PGCH endeavours have either been made or are currently underway. A peculiar type of adverbial clause that ties in neatly with the preceding discussion is the converbal construction (cf. Haspelmath 1995, König and van der Auwera 1990 for typological surveys, and Creissels' article in this volume). As we shall see in more detail later on, two important structural parameters for classifying converbs are the presence of the subordinate-clause subject, and the 'augmentation' of the converbal clause by means of an explicit subordinator that provides a link between the two clauses. In (3), we thus find an implicit-subject converb from Huallaga Quechua (Quechuan: Peru), while (4) illustrates an explicit-subject converb from Hungarian (Uralic: Hungary). Notice that its English translation also includes a converb, often called the Past Participle in English grammar, and that this converb (in contrast to the Hungarian original) is augmented by an overt subordinator (*with*):

(3) *implicit-subject converb or 'free adjunct'*
Huallaga Quechua (Weber 1989: 304) and English
[*Aywa-ra-yka-**r***] *parla-shun.*
go-STAT-IMPF-**CONV** converse-12IMP
'Let's converse as we go along.' (lit. '**Going along**, let's converse.')

(4) *explicit-subject converb or 'absolute'*
Hungarian (Kenesei et al. 1998: 55)
[*Az eső el-áll-**ván***], *elindultunk a hegytetőre.*
the rain PRFV-stop-**CONV** left.1PL the hilltop.SUBL
'**With** the rain stopped, we left for the hilltop.'

A detailed corpus-linguistic study on English free adjuncts and absolutes (cf. Kortmann 1991) reveals interesting differences between those structural types of converbs with regard to their preferred positioning patterns. In accordance with our discourse-pragmatic findings from above, Kortmann claims that absolute converbs typically serve the local modifying function of adverbial clauses, often as (focal) afterthoughts, while free adjuncts more readily assume the topical role of grounding or 'guide-posting' the interpretation of the matrix clause. This is reflected in the preferred order for each type in Kortmann's data (Fig. 2): Even though all types of converbs are predominantly found after the matrix clause, the odds for preposing significantly increase from unaugmented to augmented absolutes, and again from augmented absolutes to free adjuncts.

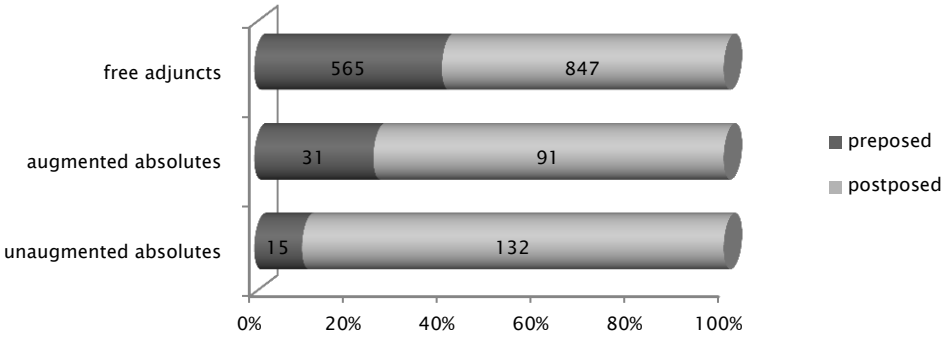


Figure 2. Scaled bar chart of Kortmann's (1991) corpus data on the position of English converbs. The distribution is highly significant, with a moderate effect size ($\chi^2 = 57.94$, $p < .001$, Cramer's $V = 0.186$).

In a more typologically oriented paper, Kortmann (1995) discusses these patterns again and makes clear predictions for cross-linguistic distributions: For languages which do not categorically restrict the position of their subordinate clauses, explicit- and implicit-subject converbs should display marked differences in ordering preferences. Unfortunately, according to Igor Nedjalkov (pers. comm.), the present quantitative database on converbs is not (yet) sufficient to investigate this issue in detail. Therefore, it is to be hoped that future efforts will be directed at compiling and extending the relevant typological data in this area (e.g. in the *Typological Database System*).

The situation is very similar in the realm of complementation. While both Grosu and Thompson (1977) and Dryer (1980) provide interesting hypotheses on (and possible explanations for) the positioning patterns of complement clauses, a principled and comprehensive cross-linguistic study of this phenomenon is currently lacking in the typological literature. A current research endeavour[5] seeks to fill this gap. Here, too, a

[5] This is part of the author's PhD project on complement clause constructions and complementation systems in the world's languages.

usage-based perspective is likely to be illuminating since (i) recent corpus-based analyses make us aware of the functions associated with structural choices in complementation (e.g. Kaltenböck 2004 on extraposition), (ii) advances in processing theories have relativized (or at least differentiated) our conceptions of the processing costs attributed to extraposition, centre-embedding and other weight effects (cf. Hawkins 2007, Yamashita and Chang 2001, MacWhinney and Pléh 1988), and historical explanations, especially recent insights from the grammaticalization of subordinate clauses, need to be incorporated into the discussion (e.g. Deutscher 2009, among many others). Viewing those language-specific studies against a fresh and sufficiently large typological database can contribute to our understanding of the principles that underlie the linearization of complex sentences.

3. Constraints on co-reference patterns in interclausal constructions

Another major structural dimension concerns the NPs found in clause combining constructions. When two propositions are integrated into a complex sentence (whether co- or subordinated), they are conceptualized as a relatively coherent chunk of the experience to be conveyed. Not surprisingly, then, the two clauses often, though by no means necessarily, share one or more participants involved in this experience. Grammatical systems are extremely diverse in their explicit means to keep track of participants across clause boundaries, and they often allow for economical coding when participants are shared between adjacent clauses. However, there are interesting constraints on how this tension between explicitness and economy needs to be resolved for individual constructions. Those constraints, if studied across languages, give rise to typological scales, such as the Accessibility Hierarchy in relativization or various co-reference hierarchies in control constructions. In this section, we will examine to what extent typological hierarchies of argument-structural conventions correspond to significant frequency rankings in the performance of individual languages. Section 3.1 will evaluate corpus data on relativization, while section 3.2 is concerned with argument realization in complement and converbal clauses.

3.1 Argument structure in relativization

In the domain of relative clause constructions (henceforth RCCs), perhaps the central concern of typological research has been with the sets of grammatical relations that are defined by relativization within and across languages. Two (by now 'classic') cross-linguistic observations are particularly relevant here.

On the one hand, the seminal work by Keenan and Comrie (1977) uncovered that argument roles within a proposition differ with regard to their 'accessibility' to

relativization. More specifically, cross-linguistic data give rise to an 'Accessibility Hierarchy' (henceforth AH) such that any position on this scale implies that all higher positions will also be accessible relativization sites in the vast majority of languages:

(5) Accessibility Hierarchy (AH) according to Keenan and Comrie (1977)
Subject > Direct object > Indirect object > Oblique > Genitive

On this view, by far the most frequent clustering of grammatical roles in relativization is that of {S,A}, which outranks {P}. Typological data appear to provide ample evidence for this subject > object contrast, which comes in the form of three phenomena. (i) We can find conventionalized cut-off points between subjects and objects such that the only available RCC in a given language is categorically restricted to {S,A} relativization. Such a situation is found, for instance, in Awa Pit (Barbacoan: Equador), in which anything other than subject relativization "is simply not possible" (Curnow 1997: 282). (ii) This situation contrasts with languages in which other arguments can be transformed into relativizable {S,A} positions by (anti)passivization or similar (diathesis-changing) constructions, as in many Austronesian languages (e.g. Begak Ida'an, Goudswaard 2005). (iii) The cut-off between {S,A} and {P} may also be marked by the use of a different relativization strategy. In Krongo (Kadugli: Sudan), for example, the gapping RC-construction is restricted to subjects, while all other syntactic roles need a resumptive-pronoun construction (cf. Reh 1985: 253). In sum, the {S,A} prevalence in relativization seems to have a firm cross-linguistic basis, and plenty of psycholinguistic studies have related this subject-object asymmetry in grammars to apparent processing differences between {S,A} and {P} (cf. Wanner and Maratsos 1978 for the most frequently-cited study in this context).

On the other hand, it is well-known that quite a few languages exhibit syntactic ergativity in that they open their RCCs to {S,P} sites only, or grammaticalize distinct constructions for {S,P} relativization on the one hand and {A} on the other. Crucially, this does not only apply to the 'usual suspects' with ergative alignment in coding constructions, such as Yidiɲ (cf. Dixon 1977), but also to others like Urarina (isolate: Peru), which does not have ergative case marking (cf. Olawsky 2006).[6] Therefore, {S,P}

[6] Note that the organization of grammatical relations in RCCs is a syntactic (or behavioural) issue and hence in principle independent of the grammatical relations defined by other (coding) constructions, such as case marking or agreement in main clauses. For the latter, ergativity is known to be a recessive feature in the languages of the world and this, in turn, might be due to preferred on-line processing strategies in the functional interpretation of arguments, for which there appears to be a clear {S,A} preference in experimental research on typologically diverse languages (cf. Bornkessel-Schlesewsky et al. 2008). This finding, however, does not *necessarily* carry over to the treatment of arguments by an entirely different construction, i.e. relativization, as we can see in Urarina. In fact, the interplay of coding and behavioural constructions, for which Kazenin (1994) first proposed implicational universals, is at the forefront of

clustering by RCCs appears to be a notable phenomenon and its recognition has actually led to revisions of the original AH. Thus Lehmann (1984) incorporated the common {S,P} preferences into the AH:

(6) (Upper part of the) revised AH according to Lehmann (1984)
{SA} or {SP} > A or P > indirect object or secondary object > adjuncts

We are thus left with two seemingly competing patterns for the upper part of the AH, and there have been multiple attempts to relate both patterns to language-specific preferences in processing and language use. In an early corpus study, Keenan (1975) finds corroborative evidence for the original AH (which neatly ties in with many classic psycholinguistic studies); Fox (1987), by contrast, argues that S and P are the preferred relativization sites in (especially spoken) discourse. Under the PGCH, this would provide a motivation (though not yet an explanation) for languages to conventionalize easy access to S and P sites. Ever since those early studies, more extensive and more fine-grained corpus-linguistic evidence has been collected, and here we shall briefly review to what extent those new studies are insightful pieces of 'converging evidence' in the spirit of the PGCH.

Gordon and Hendrick's (2004) study sets out to systematically compare how well the original AH and Fox's revised AH fare across different corpora from English, specifically the BROWN, SWITCHBOARD and CHILDES corpora. They basically claim that the original AH proves superior because it consistently matches the frequency distributions in all three corpora. This is shown in Table 1, which also includes the corpus findings from a very recent comprehensive study on English RCCs in the ICE-GB corpus (cf. Wiechmann 2009):

Table 1. Corpus distributions of traditional grammatical relations in English relative clauses across different corpora

		Relativization site in terms of Keenan and Comrie (1977)			Row totals	P_{binom} Sub>Obj	Medium	Genre
		Subject	Object	Oblique				
Gordon and Hendrick 2004	BROWN	**1606**	**848**	174	2628	2.2e-16	written	mixed
	SWITCH	**506**	**394**	98	998	2.2e-4	spoken	conversation
	CHILDES	**473**	**247**	52	772	2.2e-16	spoken	conversation
Wiechmann 2009	ICE-GB	**605**	**255**	129	989	2.2e-16	mixed	mixed

currently progressing research on a comprehensive database of grammatical relations (cf. Witzlack-Makarevich 2007).

As can be seen, subject extraction significantly outranks object extraction across all four corpora (cf. the binomial test results), and the complete distributional pattern also conforms to the original AH. If the traditional grammatical relations are now split up to S, A and P, the following distribution emerges in Gordon and Hendrick's data (Table 2):

Table 2. Corpus distributions of sematico-grammatical roles in English relative clauses across different corpora

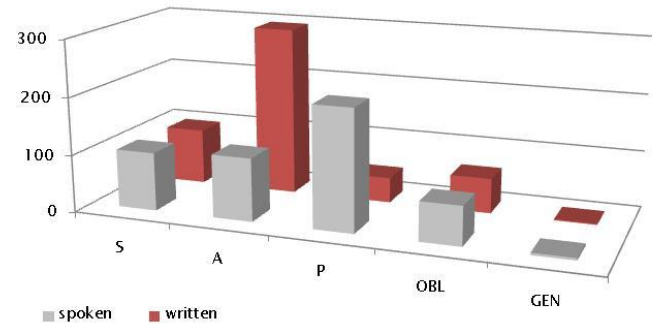
		Relativization site in terms of Fox (1987)			Row totals	Ranking of S, A and P	Medium	Genre
		S	A	P				
Gordon and Hendrick 2004	BROWN	590	1016	841	2447	**A > P > S**	written	mixed
	SWITCH	286	220	393	899	**P > S > A**	spoken	conversation
	CHILDES	271	198	246	715	**S > P > A**	spoken	conversation
Wiechmann 2009	ICE-GB	200	405	255	860	**A > P > S**	mixed	mixed

These results show that Fox's hypothesis can only account for two corpora, SWITCHBOARD and CHILDES. In addition, while the original AH proved to correctly predict the frequency differences for *all* grammatical roles it includes on the typological scale, Fox's hypothesis is weaker in that it does not predict the distribution of S and P relative to each other. In fact, as we can see in Table 2, in the two corpora for which Fox's overall prediction is correct, the relative ranking of S and P is inconsistent. For these reasons, Gordon and Hendrick argue that the original AH turns out to be the more robust one, resisting differences between individual corpora, and given that it also receives vast experimental support, should be preferred to Fox's proposal.

It turns out, however, that Gordon and Hendrick's criticisms are in parts ill-founded and that a more refined interpretation of the corpus results is necessary. To start with, their data collection from the CHILDES corpus is heavily biased since it includes only relative clauses with an overt relativizer (*that*). This creates various problems since it is well-known that for object-, but not subject-extracting relatives, the relativizer is optional. Other things being equal, this by itself will bias the query towards subject RCs. (For illustration, the reader may take a brief look at the examples of children's early object RCs in Diessel and Tomasello's (2000) acquisition study: None of them contains an overt relativizer!) What is more, it has been shown that subject RCs are easier to master (and hence more frequent in the speech of young children) than object RCs even though the latter are more frequent in the ambient language (cf. Diessel 2004). This finding, too, makes the portion of the CHILDES corpus extracted here unsuited for the point to be made. Finally, the sampling procedure also most likely results in the different internal rankings of S and P in Table 2: Gordon and Hendrick's criticism that the figures for P relativization are significantly different in SWITCHBOARD and CHILDES

may simply have been inapplicable if P relativization had been covered *exhaustively* in the CHILDES database.

Apart from this sampling error, a closer look at the medium differences between the corpora is revealing for evaluating Gordon and Hendrick's conclusion. Notice, first, that the binomial test for the entirely spoken SWITCHBOARD corpus (cf. Table 1) yields a markedly different result from the other corpora, which contain either only written or mixed data (disregarding the suboptimal CHILDES collection). Indeed, if we partition Wiechmann's (2009) data into the spoken and written component of his ICE-GB sample, the picture becomes much more differentiated (Fig. 3):



Frequency of traditional grammatical roles

	{SA}	{P}	TOTAL
spoken	212	211	423
written	393	44	437
TOTAL	605	255	860

Figure 3. Frequency distribution of relativization sites in Wiechmann's (2009) sample from the ICE-GB (OBL = oblique (indirect or prepositional objects and adjuncts), GEN = genitive/possessor). Wiechmann extensively coded 1000 randomly selected instances of RCCs, 500 each from spoken and written language. While the graph on the left visualizes the frequency differences between all individual roles in the two genres, the grid on the right lists the exact figures for the crucial contrast between subject- and object relatives.

The distribution in Fig. 3 reveals a straightforward contrast between spoken and written data with regard to preferred relativization sites. Wiechmann's data from written language testify to the rankings predicted by the original AH, with all individual differences being significant under an exact binomial test ($p < .001$). The subject-object extraction asymmetry in written language is best reflected by the table in the right panel of Fig. 3 ($\chi^2 = 163.33$, $p < .001$, with an effect size of $\phi = 0.436$). This essentially confirms Keenan's (1975) corpus findings for written language. In spoken interaction, by contrast, {P} comes out as the most preferred relativization site, so that, on aggregate, subject- and object RCs are actually equally frequent in this medium. This finding mirrors the results of Fox (1987) in a larger and more principled data set, and thus provides supporting evidence for a preferred {SP} > {A} ranking of the relativization sites in spoken discourse.

In accordance with Fox (1987), Wiechmann (2009) also finds that non-subject RCs in spoken language have a distinct structural profile. Specifically, they are often headed by a pronominal, generic and inanimate NP (e.g. *that*, *something*) and contain an internal pronominal (rather than lexical) subject. More fine-grained analytical techniques, such as *Configural Frequency Analysis*, allow Wiechmann to break down this general profile even further into significantly recurring construction types of non-subject RCCs. This provides a firm quantitative underpinning of Fox and Thompson (2007), who argue that our 'competence' with relative clauses is best captured in terms of lexically-specific templates, or concrete RC constructions at low levels of syntactic generalization which dominate actual language use (e.g. *All* $_{RC}$[NP_{pron} *wants to do*] *is* V. as in *All she wants to do is sleep.*) Consequently, such frequent RCC patterns are by no means difficult to activate in language processing. This is confirmed by a number of recent psycholinguistic studies that subscribe to experience- (or usage-) based processing theories. Reali and Christiansen (2007), for instance, conducted a series of experiments which show precisely that for RCCs in which the first internal NP is pronominal (e.g. *the consultant that* [*you called*]$_{RC\text{-}Obj}$ or *the consultant* [*that called you*]$_{RC\text{-}Subj}$), object RCs significantly outrank subject RCs in terms of processing ease.

The findings summarized in this section thus illustrate how carefully investigated distributional information from language use refines our understanding of the constraints on relativization sites (and of preferred RC constructions more generally). First, they strengthen Fox's (1987) idea that syntactic ergativity is a preferred discourse pattern of relativization in spoken language.[7] In a usage-based theory like the PGCH, such distributional properties in discourse can give rise to conventionalized {SP} > {A} role constraints in grammars, i.e. typologically relevant patterns. Second, if the typologically dominant subject > object asymmetry in relativization is to be explained in terms of language processing, we need to acknowledge that many structure-based parsing theories (such as dependency domains and ensuing working-memory costs, or solely syntactically-based accounts) can actually make wrong predictions for the majority of RCCs occurring in spoken language (which is the sole medium of conversation in many indigenous languages, after all). In other words, those processing accounts are only successful if preferred discourse patterns and construction types are glossed over or factored out, casting doubt on the alleged existence of a *universal* 'cognitive' or 'processing' subject prominence in relativization. It rather appears that in RCCs, typological distributions are shaped by both cognitive and communicative pressures.

[7] Fox (1987) and especially Fox and Thompson (1990) also provide an elaborate argumentation as to *why* such syntactic ergativity makes sense from the perspective of anchoring referents in discourse and integrating them with the current flow of information, and *why* spoken language is considerably more prone to these considerations than written language. For the present paper, it suffices to note that there is convincing evidence for a discourse-based perspective on relativization.

We will round off this section by pointing to one particular RCC whose accessibility properties promise to be an interesting avenue of future research. Apart from having a canonical RC, many languages have functionally expanded a purposive (notably infinitival) or modal construction into the domain of relativization. (7) provides an example from Mayogo (Adamawa-Ubangi: Congo):

(7) Mayogo (Sawka 2001: 98)
Ma ne gʉmʉ [***na****-dji engu*].
1SG with need **INF**-drink water
'I have (a) need to drink water.'

Such modal relatives are at present rather little studied along the major dimensions of relativization, but we know that there are several languages which place interesting constraints on their relativization sites. Modern Hebrew, for example, which makes all syntactic roles accessible to canonical finite relativization (albeit with different strategies), reverts the AH in only allowing lower syntactic roles to be relativized by the infinitive. Specifically, as the comparison of (8a) and (8b) goes to show, subject relativization is categorically excluded from this construction:

(8a) Modern Hebrew (Glinert 1989: 373)
Hine me'il [*lilbosh*].
here coat wear.INF
'Here's a coat to wear.'

(8b) **Hine káma tmunot* [*leanyen otaH*].
here some picture.PL interest.INF 2SG
'Here are some pictures to interest you.'

English, again, proves to be a language with unrestricted site access in infinitives, but corpus studies reveal significant performance preferences in keeping with the hard grammatical constraints of other languages. In Geisler's (1998) study of the London-Lund Corpus of Spoken English, for instance, object extraction in infinitival relatives outnumbers subject extraction by 2:1 (206 : 105 instances in the corpus). Subject extraction is even slightly outnumbered by or at least on a par with adverbial relativization (106 : 105). If this trend continues to show up in further typological research, language-internal skewings might again help to substantiate and motivate such AH reversals. Interestingly, Geisler (1998) finds significant associations between relativization site and external, i.e. matrix, function of the relativized NP, and proposes

a discourse-based account that also builds on Fox and Thompson's (1990) notions of referent grounding and anchoring.

3.2 Co-reference patterns of complement and converbal subjects

The patterns of argument realization in complement clauses have generated a great deal of research in language typology. A recent broader study is found in Cristofaro (2003), where argument-structural properties of various types of complement clauses are systematically compared to those in adverbial and relative clauses and interpreted in the context of the functional parameters of subordination more generally. For the present survey, we will exemplarily focus on desiderative complementation (*want* V).

A key issue in desiderative complements is the overt realization of the subordinate subject. What we find recurrently in corpora is a strong preference for the subject of the main clause to be co-referential with the subject of the subordinate clause (*Tom$_i$ wants* [*Ø$_i$ to go to the movies tonight*].) even though, of course, it is perfectly possible to have a desire for someone else to do something (*Tom$_i$ wants* [*her$_j$ to take care of the children*].). Haspelmath (2008a) offers text counts for Italian and Greek *want*-clauses, and I performed an exhaustive query of all *want*-complements (and functionally analogous clauses governed by *(would) like*) in the ICE-GB.[8] Even though Italian, Greek and English use formally different clause types for the expression of desiderative complementation, the corpus findings reveal the same trend (Table 3):

Table 3. Usage frequencies for subject reference in *want*-constructions

Language	Verb	Same subject (SS)	Different subject (DS)	Row totals
Greek	*thélo*	444	65	509
Italian	*volere*	38	5	43
English	*want*	540	76	616
English	*like*	217	26	243

Needless to say, all of the individual distributions are significantly skewed under an exact binomial test ($p < .001$) and, interestingly, the four volitional verbs in Table 3 are also almost *equally* skewed towards the same-subject preference ($\chi^2 = 0.69$, $df = 3$, $p = .88$). Under the PGCH, this language-internal trend gives rise to several predictions for typological distributions. First, the overwhelmingly more frequent pattern in performance should be more productively grammaticalized across languages. Indeed,

[8] For both *want* and *like* complementation, a fuzzy-tree fragment search in the ICE-GB was performed to extract all instances of the relevant pattern. Different-subject *want*-clauses, for instance, were characterised as a processing unit with two daughter nodes: a lexical node containing the main verb *want*, and a clausal node introduced by an NP of any kind followed by the particle *to*.

Haspelmath (1999, 2008a) reports that some languages, while having a productive *want*-SS complementation pattern, simply cannot render the corresponding DS-meaning in the same way. Speakers of Acehnese or Tümpisa Shoshone, for instance, need to resort to a syntactic paraphrase to fill the 'constructional gap' of *want*-DS complementation.

The second prediction relates to the fact that the higher odds for the occurrence of the SS-pattern in language use make it a well-predictable pattern and hence prone to economical coding (cf. Zipf 1949). Across languages, then, we would expect SS-*want* constructions to preferably leave the embedded subject implicit, giving rise to well-known 'control' constructions. In a recent typological survey, Haspelmath (2005) finds substantial evidence for this hypothesis; the distribution of overt subject realization in SS-*want* constructions is displayed in Fig. 4.

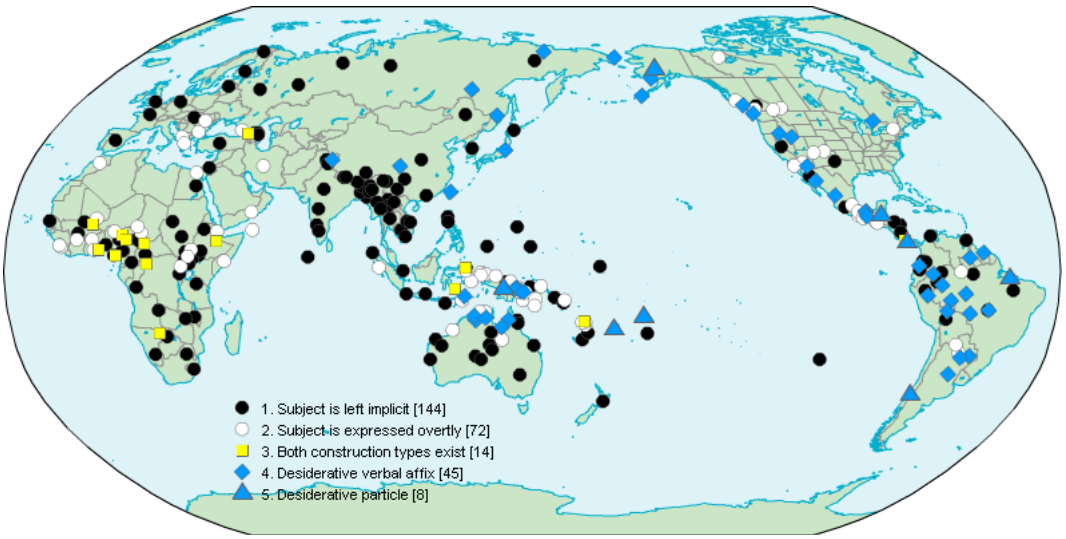


Figure 4. Subject realization in SS-want complements (Haspelmath 2005)

As we can see, the implicit-subject pattern dominates the sample by far (144/230 = 62.6%, $p_{binom} < .001$, discounting desiderative morphemes) and has robust world-wide coverage (at least in the sense of Dryer's (1989) macro areas). But the map also suggests certain areal effects, notably the circum-Pacific distribution of desiderative morphemes, the restriction of constructional choice (yellow rectangles) to central Africa and Oceania, as well as more local trends such as the overt subject expression in Papua New Guinea or south-eastern Europe (cf. the lack of infinitives or predilection for finite subordination in the Balkan region). Clearly, however, the implicit-subject pattern prevails, and Haspelmath (1999) goes on to argue that the higher usage frequency of SS also gives rise to other economical effects in grammars. For instance, the complementizer may be reduced or even absent in the SS-version of the construction (e.g. Maltese, Hopi) or it coalesces phonologically with the main verb (cf. English *I wanna go.* versus *I want him to go.*). A more radical effect is the separation of SS-

and DS-*want* complements into two different constructions altogether (e.g. infinitival versus finite subordination in German or Lango), or the lexicalization of two different verb lexemes for *want* (the SS-version of which, crucially, is shorter than the DS one, e.g. in Japanese or Samoan). These differential effects all demonstrate how speakers tend to 'minimize form' (cf. Hawkins 2004: 38) in proportion to the usage frequency of particular patterns in actual discourse, such that *want*-DS-clauses end up being at least as complex in form as SS-clauses across the world's languages. Crucially, then, it is the availability of corpus-linguistic data that enables us to anchor such grammatical generalizations in usage events, and to make a case for performance-based syntactic theory.[9]

Essentially the same observations also apply to converbal constructions (cf. also §2.2 above). Typologically, we find the whole range of argument patterns in converbs, from those constructions that cannot be used with overt subjects (e.g. the Hungarian 'simple converb' in Kenesei et al. (1998)), those that must be used with an overt subject (e.g. adverbial 'gerund' clauses in Glinert's (1989) description of Modern Hebrew), and those that leave a choice (e.g. Lezgian [cf. Haspelmath 1993], English). Amongst the latter, we usually find performance preferences for subject omission, motivated by participant sharing across the two clauses. Thus in Kortmann's (1991) corpus on English, implicit-subject converbs ('free adjuncts', N = 1,412) outnumber explicit-subject converbs ('absolutes', N = 269) by far. If English is representative of a more general trend, the quantitative expectation across grammars would thus clearly be in favour of implicit-subject converbs. What is more, Kortmann also proposes a 'Controller Accessibility Hierarchy' for those converbs, which he believes "may well reflect universal tendencies" (Kortmann 1995: 227). This hierarchy comprises a whole set of implicational relations, such as matrix > non-matrix control, NP > non-NP control, and a ranking of subject > object > complement for matrix NP controllers. As was noted earlier, the cross-linguistic validity of those generalizations must await the compilation of a substantial typological database of converbs that systematically records the pertinent variables, but given that there are very plausible functional motivations for

[9] It must be added that Haspelmath (1999, 2008a) also develops an account of the diachronic mechanisms that lead to the conventionalization of different grammatical patterns for SS- and DS-situations, which is essential for a comprehensive usage-based theory. The basic idea is that of differential selection in syntactic change (cf. also Croft 2000), and it remains to be investigated how well this proposal accords with diachronic data from pertinent languages. (From this perspective, it would also be particularly interesting to observe the future development of the 14 languages in Fig. 4 which currently leave a constructional choice for same-subjects to be expressed or left out (e.g. Koromfe, Obolo, Ternate)). As far as the ultimate motivation of the SS > DS asymmetry in usage is concerned, Haspelmath (1999: 3) claims that it is due to human egocentrism: "Our own actions are much more important to us, so we talk more about wishes concerning these than about wishes concerning the action of others."

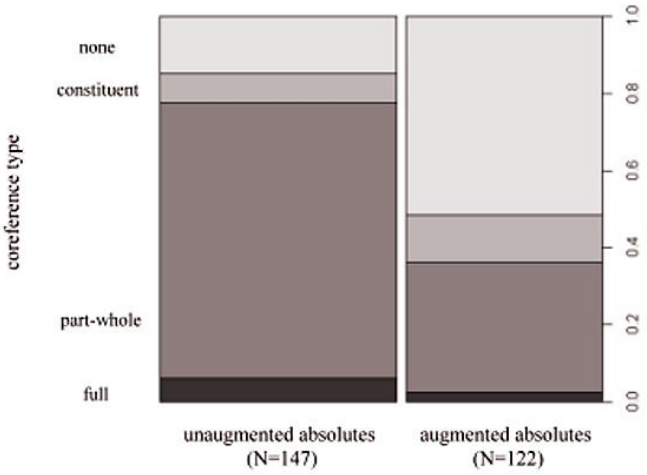
controller hierarchies (cf. Kortmann 1995), we would not be surprised to see correspondences between intra- and cross-linguistic variation in this domain.

Perhaps a more subtle phenomenon concerns the class of explicit-subject converbs (i.e. Kortmann's 'absolute participles'). In Kortmann's study, this class is divided again by the presence of an 'augmentative' subordinator such as English *with*:

(9) unaugmented absolute converb
Tommy screamed, [*his cracked voice riding clearly through the still air*].
(ICE-GB W2F-002 184)

(10) augmented absolute converb
Another possibility is that Samuel and his new wife moved there when they married, [***with*** *Eugenie returning to her mother's house*]...
(ICE-GB W2B-002 049)

Interestingly, augmented and unaugmented absolutes also differ markedly with regard to the co-reference relations that their explicit subject contracts with other discourse participants. Specifically, while unaugmented absolutes are overwhelmingly tied to their matrix clause by a part-whole co-reference[10] of the subject (as exemplified in [9] above), the subjects of augmented absolutes tend not to show any co-reference relations to a matrix participant (cf. (10)). The more precise co-reference patterns in Kortmann's study are reproduced here as a spine plot (Fig. 5), which allows easy visual access to the significantly different proportions of each co-reference relation (χ^2 = 49.55, p < 0.001, Cramer's V = 0.429):



[10] The term is adopted from Kortmann and comprises meronymic relationships in the traditional sense as well as possessive relationships like the one in (9).

Figure 5. Spine plot for the co-reference relations in augmented and unaugmented English converbs (data from Kortmann 1995: 213)

The graph reflects that the odds for augmentation increase when the absolute clause contains a subject participant not found in the matrix. According to Kortmann, argument co-reference and the overt presence of the subordinator parcel out their work in a way that maximizes the "acceptability and ease of processing" of the sentence (Kortmann 1995: 213): The more the two clauses are connected by virtue of argument sharing, the less explicitly the connectedness needs to be marked by an overt clause linker (i.e. adverbial conjunction or adposition). This echoes very recent developments in the processing assumptions underlying the PGCH. Hawkins (2009) proposes that two fundamental operations in on-line parsing consist in (i) the 'construction' (i.e. identification) of the syntactic unit in question (e.g. NP, PP, subordinate clause, complex sentence), and (ii) in the 'attachment' of all material that belongs to this unit (e.g. determinatives of all sorts in NPs, or the two or more clauses in a complex sentence). Importantly, *efficiency* plays a key role in parsing in Hawkins' theory, such that, for example, the amount of material that is needed for reliable construction and attachment of a unit is preferably small ('Minimize Domains'), and in that readily available contextual information can be exploited at the expense of overt linguistic material ('Minimize Forms'). On this account, the subordinator in absolute converbs is a useful cue for constructing the complex sentence structure. By virtue of occurring at the left clause boundary, the complex-sentence ('mother') node can be immediately constructed when the subordinator is encountered. If the absolute clause started with a referentially unrelated subject NP, then this NP would delay the construction of the complex sentence node and the attachment of the converbal clause. This, in turn, means that the NP itself is left 'unattached' since there is no mother node that guides the processing of further material. Consider thus the unattached nature of the underlined NPs in the following examples from the ICE-GB:[11]

(11) [...] *but nowadays most people actually want the words spread out <u>the page</u> going to the edge of each boundary.* (ICE-GB S2A-058 016)

(12) *Eventually it was waved through, <u>the final red-and-white barrier</u> rising to let it past,* [...] (ICE-GB W2F-012 019)

If these examples contained a connecting *with*, the listener's expectation would be cued towards the correct subordinate-clause interpretation, so that the following NP *the page*

[11] The examples were extracted from the ICE-GB with a fuzzy-tree fragment search for adverbial clauses containing a subject NP and a present participial form.

or *the final red-and-white barrier* could readily be interpreted as subjects of this attached clause.[12] If this processing interpretation of Kortmann's corpus findings is correct, then it strengthens Hawkins' idea – established chiefly on the basis of cross-linguistic data – that construction and attachment are critical in parsing, while the specific semantic relation holding between the two clauses (for whose expression *with* is "fundamentally unsuited", Kortmann 1995: 227) is largely left to ex post pragmatic inference.

The tasks of future research in this domain are evident: Much more typological data on converbs need to be examined from this processing perspective, paying due attention also to less frequently encountered constructions (here: explicit-subject converbs), for which corpus studies and their interpretation in terms of the PGCH make clear distributional predictions. Furthermore, recall from above that Kortmann does not only appeal to on-line processing, but also to the 'acceptability' of absolute converbs, and it is here that the second type of converging evidence in the PGCH paradigm suggests itself: Performance preferences from corpora translate into rankings in acceptability judgements, which can be tested in experimental psycholinguistic research. On yet another level of processing – our neurological architecture – the relatively young but burgeoning field of 'neurotypology' (cf. Bornkessel-Schlesewsky et al. 2008) has already made important advances and promises further contributions to the interdisciplinary effort of validating language universals as required by the PGCH.

4. Further areas of application

Linearization and co-reference relations do not, of course, exhaust the structural variables of complex sentence constructions that can be approached from a usage-based perspective. Especially with semantic properties, such as the association of lexical verbs and particular grammatical constructions, performance data can be helpful whenever the corresponding typological sample faces certain limitations.[13] For instance, my typological sample of purpose clauses (N = 80) yielded 20 languages with a distinctly grammaticalized avertive construction, i.e. negative purpose clause, as exemplified in (13):

[12] It is also possible, of course, that *with* after a main clause is initially misinterpreted as introducing a prepositional object or adjunct of the matrix, but intonation in spoken language and the comma in written discourse make this misassignment rather unlikely.

[13] The reader may object at this point that significant associations of lexical and grammatical material are always language-specific and hence not amenable to 'converging-evidence research'. I principally agree, subscribing to a 'Radical Construction Grammar' in Croft's (2001) sense myself. However, the present proposal applies to verb-construction interactions for which there is significant typological evidence in the first place, i.e. which are attested in geographically and genetically independent languages.

(13) Slave (Athapaskan: Canada; Rice 1989: 1262)
[*Daniel yegúhʔ ále* ***ch'á***] *goghádehk'a.*
Daniel 3OPT.find 4.NEG **LEST** 1SG.threw
'I threw it so Daniel wouldn't find it.'

Upon closer inspection, it turned out that this construction type is well worth investigating in its own right since it exhibited the opposite grammatical characteristics of positive purpose clauses, such as a tendency for balanced (instead of deranked) verb forms, different (and notably semantic-patient) subjects (instead of same-agent subject sharing), and a lack of allative and recipient-benefactive markers. Furthermore, whereas some positive purposive constructions obligatorily require a matrix verb of motion (as in Hixkaryana or Krongo (cf. (14)), there seems to be no language in which this grammatical requirement applies to negative purpose clauses.

(14) *motion-cum-purpose construction*
Krongo (Kadugli: Sudan; Reh 1985: 351)
M-áa cáaw òmúno-ŋ éekwàarà.
CONN.F-COP go.INF (DAT.)INF.call-TR chief
'And she goes to call the thief.'

In the parlance of the PGCH, such grammaticalized (and hence often categorical) associations of verbs and syntactic constructions have arisen diachronically from performance preferences. If the typological data on avertive constructions reflect the cross-linguistic trend correctly, then we would expect that motion verbs are not among the typical matrix verbs of such constructions. Modern quantitative corpus linguistics has provided specific techniques for testing such semantic constraints. Again, the reader is referred to Schmidtke-Bode (2009) for a full exposition; we will only sketch the potential of this methodological approach here.

The procedure of 'collostructional analysis' (cf. Stefanowitsch and Gries 2003) essentially relies on the exhaustive retrieval of a particular grammatical construction from a representative corpus and computes, for every verb that occurs in the construction, whether its occurrence in this environment is significantly more or significantly less frequent than expected given the verb's overall frequency in the corpus. It thus abstracts away from absolute frequencies since it filters out, as it were, those verbs that are frequent in the construction because they are high-frequency items to begin with (e.g. *do, make, be, have, get*). The method rather assesses the difference between observed and expected frequencies: If a verb occurs significantly more often than expected in a given constructional environment, it is said to be particularly 'attracted' to the construction; its significant under-representation in the

construction is interpreted as semantic 'repellence' (e.g. due to incompatibility with the semantics of the construction). In order to examine the verbal associations of avertive clauses, I retrieved all purposive instances of LEST-clauses and all purposive *so-that-not* clauses from the *British National Corpus* (BNC).[14] For the collostructional analysis, the matrix verbs were lemmatized and each of them was scrutinized for its distribution in the corpus. Specifically, we need to know four types of distributional information for each relevant verb: the frequency of the verb (lemma) in the matrix clauses of LEST-constructions, its frequency in all other constructions, the total number of LEST-clauses in the corpus, and the total number of lexical verbs in the entire corpus. Taking the verb *cut* as an example, Table 4 illustrates how this information (highlighted in grey shading) can be gathered schematically in a contingency table:

Table 4. Contingency table for a collostructional analysis of *cut* as a matrix verb of LEST-clauses

	cut	other verbs	row totals
matrix of LEST-clause	3	256	259
all other clauses	18,508	7,195,981	7,214,489
column totals	18,511	7,196,237	7,214,748

The remaining cells can now be filled in by subtraction, and the bold-marked part of the table can then be submitted to an association test. Stefanowitsch and Gries (2003: 218) argue that the *p*-value of the Fisher-Yates exact test (FET) can be interpreted as a measure of the 'association strength' between a given lexeme and a grammatical construction. To stay with our example, *cut* exhibits a significant attraction of $p = 0.02977$ under a FET and thus qualifies as a significant 'collexeme' of the LEST-environment (since $p < .05$). Crucially, the most interesting results are only revealed when collostructional analysis is repeated for all verbs occurring as matrix verbs of LEST. We then obtain an ordered list of verbs, ranked according to their *p*-value and hence their relative degree of attraction to or repulsion by the LEST-matrix. Table 5a first lists the most attracted collexemes of English LEST-clauses; Table 5b then shows the most strongly repelled lexemes of this construction, and Table 5c, finally, also provides the repelled items of the *so-that-not* construction for comparison:

[14] Retrieval and coding of the data are non-trivial issues in a corpus-linguistic study and need careful documentation and justification, for which the reader is referred to chapter 3.6.1 in Schmidtke-Bode (2009). The same applies to the subsequent identification of observed and expected corpus frequencies, which is not straightforward either and often involves minute semi-manual procedures.

Table 5. Significantly attracted and repelled matrix collexemes of avertive clauses in English

[CollStr(ength) directly reflects the FET *p*-value for each verb, but since *p*-values tend to be very small for the most attracted lexemes, they were logarithmically transformed for ease of interpretation: The higher the CollStrength value reported here, the higher the degree of attraction, with any score ≥ 3 corresponding to $p < .001$. For significance criterion at $p < .05$, the score must be ≥ 1.3.]

(a) LEST-clauses attraction		(b) LEST-clauses repulsion		(c) *so-that-not* clauses repulsion	
Collexeme	CollStr	Collexeme	CollStr	Collexeme	CollStr
beware	6.91	*be*	30.45	*be*	77.67
guard	4.46	*have*	9.11	*have*	15.70
chain	4.39	*say*	2.70	*do*	7.65
add	4.15	*get*	2.54	*say*	7.64
temper	4.15	*come*	1.57	*go*	3.82
restress	4.14	*look*	1.04	*get*	3.16
hurry	4.00	*go*	1.02	*come*	2.72

Although the most attracted items in Table 5a provide interesting insights into the semantics and usage of English LEST-constructions (cf. Schmidtke-Bode 2009), it is the repelled items of negative purpose clauses that shall concern us here. Interestingly, even though *lest* and *so that not* are associated with entirely different genres and contexts of language use, their dispreferred lexical choices are almost identical. In the present context, it strikes us that the most general motion verbs, *come* and *go*, which are categorically required for some positive purpose clause constructions across the world, are significantly repelled by negative purpose clauses (although it has to be conceded that *go* does not reach the $p < .05$ significance criterion in LEST-clauses). Looking back on our typological sample, this robust performance pattern seems to carry over to other languages, and we may now more safely conclude that there is a cross-linguistic tendency for positive and negative purpose clauses to differ markedly in their constraints on lexical co-occurrences. In conclusion, the quantitative study of lexical and grammatical association patterns presents another avenue of research for usage-based typological theory. In future work, this approach may be suited to motivate more specific lexico-grammatical constraints on complex sentence constructions. If it turned out, for example, that (desiderative) control constructions show significant performance preferences for intransitive subordinate predicates, then this would shed new light on why such restrictions show up as conventionalized rules in some languages of the world (cf. Bickel [2010] for such control clauses from the Mayan stock). Of course, recording such performance-grammar correspondences does

not yet provide an explanation for *why* particular restrictions occur, but they strongly suggest searching for motivations in the realm of language use.

5. Concluding remarks

This paper has reviewed a body of research in support of the 'Performance-Grammar-Correspondence Hypothesis' (PGCH) as applied to complex sentence constructions. For a variety of construction types, we have seen how typological generalizations (or similar categorical constraints in independent languages) can be systematically related to frequency asymmetries in languages with structural choices in a given domain. Such correspondences between statistical and categorical grammatical constraints across unrelated languages are a non-trivial issue in usage-based theories of language. On those accounts, grammatical properties are but 'frozen' or conventionalized performance preferences, and frequency of use is a central determinant of the representation of linguistic knowledge and the organization of grammatical systems. I have tried to demonstrate in this paper that quantitative corpus linguistic research cannot only reveal significant performance distributions, but also uncover and weigh the alleged motivations for those distributions. Although the present discussion has centred on phenomenological and methodological issues, I hope to have shown that when it comes to *explaining* frequency asymmetries in language use, many structural variables relating to clause combining present multifactorial problems which cannot easily be reduced to a single explanatory concept. For linear ordering and relativization, for example, our discussion suggested that the most promising explanations combine insights from language processing (notably parsing theories) with those relating to the structure of discourse. Sometimes such information-structural considerations override optimal processing conditions (as is the case in purpose clauses), which in turn can outrank the need for semantic explicitness (as we saw in absolute converbs).[15]

In order to exploit the full potential of combining corpus linguistic and typological analyses, future work will have to (i) create yet more substantial typological databases

[15] It should be emphasized again that even a combination of processing- and discourse-related factors does not provide a full account of many typological distributions. Other important factors, often seen as 'confounds' in the functional analysis, include patterns of language contact and areal diffusion (cf. Bickel 2007), and especially also forces of grammaticalization that may directly lead to certain grammatical arrangements without leaving much room for processing or discourse pressures to assert themselves (cf. e.g. Bybee 2010: 111 on historical explanations of certain word-order correlations). In the present paper, however, my major concern was with those functional motivations that we can argue for particularly well on the basis of synchronic corpus data of individual languages, hence the emphasis on processing and communicative pressures.

that, crucially, code the same fine-grained variables as those used to make insightful corpus-linguistic analyses (cf. also Bickel 2007: 247 on this point); (ii) integrate methodological advances from both disciplines; (iii) continue to make the interdisciplinary effort of drawing on the explanatory variables suggested by advances in all potentially relevant disciplines, such as psycholinguistics (cf. Hawkins 2007 for a position paper), neurolinguistics, quantitative discourse analysis, sociolinguistics, cognitive linguistics, etc.; (iv) develop and employ appropriate methods for probing the diachronic mechanisms that lead to the conventionalization of preference patterns. With regard to this latter point, both Haspelmath (2008a) and Croft (2000) provide theoretical proposals along the lines of differential selection and diachronic adaptation. With the advent of annotated diachronic corpora, this crucial aspect of usage-based theories of language can now also be studied in considerable depth, paying due attention to shifts in usage frequencies of grammatical variants (e.g. Hilpert 2007 for a quantitative perspective on grammaticalization). Pursuing these goals will, ultimately, enable us to address the performance basis of grammatical constraints in a yet more principled and satisfactory way.

Abbreviations

Interlinearization in this paper conforms to the *Leipzig Glossing Rules*. Additional or deviant glossing is listed below.

CONN	connector	REF	referential	SUBL	sublative (case)
CONV	converb	PERI	peripheral		

References

Bickel, Balthasar (2003). Referential density in discourse and syntactic typology. *Language* 79: 708–736.

Bickel, Balthasar (2007). Typology in the 21st century: Major current developments. *Linguistic Typology* 11: 239–251.

Bickel, Balthasar (2010). Grammatical relations typology. In: *The Oxford Handbook of Language Typology*, ed. Jae Jung Song. Oxford: Oxford University Press. 399-444.

Bornkessel-Schlesewsky, Ina, Kamal Kumar Choudhary, Alena Witzlack-Makarevich and Balthasar Bickel (2008). Bridging the gap between processing preferences and typological distributions. In: *Scales*, eds. Andrej Malchukov and Marc Richards. *Linguistische ArbeitsBerichte* 86. University of Leipzig. 397–436.

Bresnan, Joan (2007). A few lessons from typology. *Linguistic Typology* 11: 297–306.

Bresnan, Joan, Shipra Dingare and Christopher D. Manning (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In *Proceedings of the LFG01 Conference*. Available online at <http://csli-publications.stanford.edu/LFG/6/lfg01-toc.html>.

Bybee, Joan. (2010). *Language, Usage and Cognition.* Cambridge: Cambridge University Press.

Bybee, Joan (2006). *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Chafe, Wallace (1984). How people use adverbial clauses. *Berkeley Linguistics Society* 10: 437–49.

Cristofaro, Sonia (2003). *Subordination.* Oxford: Oxford University Press.

Croft, William (2000). *Explaining Language Change: An Evolutionary Approach.* London: Longman.

Croft, William (2001). *Radical Construction Grammar.* Oxford: Oxford University Press.

Croft, William (2010). Language structure in its human context: New directions for the language sciences in the 21st century. In: *The Cambridge Encyclopaedia of the Language Sciences,* ed. Patrick Colm Hogan. Cambridge: Cambridge University Press. 1-11.

Curnow, Timothy J. (1997). *A Grammar of Awa Pit (Cuaiquer).* PhD dissertation: Australian National University.

Deutscher, Guy (2009). Nominalization and the origin of subordination. In: *Syntactic Complexity*, eds. Talmy Givón and Masayoshi Shibatani. Amsterdam, Philadelphia: John Benjamins. 199-214.

Diessel, Holger (2001). The ordering distribution of main and adverbial clauses: A typological study. *Language* 77: 345–365.

Diessel, Holger (2004). *The Acquisition of Complex Sentences.* Cambridge: Cambridge University Press.

Diessel, Holger (2008). Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19.3: 457–482.

Diessel, Holger and Katja Hetterle (to appear). Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. In: *Linguistic Universals and Language Variation,* ed. Peter Siemund. Berlin: Mouton de Gruyter.

Diessel, Holger and Michael Tomasello (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics* 11.1–2: 131–151.

Dixon, Robert M.W. (1977). *A Grammar of Yidiɲ*. Cambridge: Cambridge University Press.

Dryer, Matthew S. (1980). The positional tendencies of sentential noun phrases in universal grammar. *The Canadian Journal of Linguistics* 25: 123–196.

Dryer, Matthew S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13.2: 257–292.

Dryer, Matthew S. (2005a). Relationship between the order of object and verb and the order of relative clause and noun. In: *The World Atlas of Language Structures,* eds. Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie. Oxford: Oxford University Press. 390–393.

Dryer, Matthew S. (2005b). Relationship between the order of object and verb and the order of adjective and noun. In: *The World Atlas of Language Structures,* eds. Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie. Oxford: Oxford University Press. 394–397.

Du Bois, John (1987). The discourse basis of ergativity. *Language* 63: 805–855.

Fox, Barbara A. (1987). The noun phrase accessibility hierarchy revisited. *Language* 63: 856–870.

Fox, Barbara A. and Sandra A. Thompson (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language* 66: 297–316.

Fox, Barbara A. and Sandra A. Thompson (2007). Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language* 31.2: 293–326.

Frank, Paul (1990). *Ika Syntax.* Arlington: Summer Institute of Linguistics and The University of Texas at Arlington.

Geisler, Christer (1998). Infinitival relative clauses in spoken English. *Language Variation and Change* 10: 23–41.

Givón, Talmy (2001). *Syntax: An Introduction.* Amsterdam, Philadelphia: John Benjamins.

Glinert, Lewis (1989). *The Grammar of Modern Hebrew*. Cambridge: Cambridge University Press.

Gordon, Peter C. and Randall Hendrick (2004). Relativzation, ergativity, and corpus frequency. *Linguistic Inquiry* 36: 456–463.

Goudswaard, Nelleke E. (2005). *The Begak (Ida'an) Language of Sabah*. Dissertation. Utrecht: LOT Publications.

Greenberg, Joseph H. (1963) [1966]. Some universals of grammar, with particular reference to the order of meaningful elements. In: *Universals of Language,* 2nd ed., ed. Joseph H. Greenberg. Cambridge, Mass.: MIT Press. 73–113. [1st ed. 1963]

Gries, Stefan Th. (2006). Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54.2:191–202.

Grosu, Alexander and Sandra A. Thompson (1977). Constraints on the distribution of NP clauses. *Language* 53: 104–151.

Haspelmath, Martin (1993). *A Grammar of Lezgian.* Berlin, New York: Mouton de Gruyter.

Haspelmath, Martin (1999). On the cross-linguistic distribution of same-subject and different-subject complement clauses: Economic vs. iconic motivation. Paper presented at the *6th International Cognitive Linguistics Conference* (ICLC), Stockholm, July 1999.

Haspelmath, Martin (2005). 'Want' complement clauses. In: *The World Atlas of Language Structures,* eds. Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie. Oxford: Oxford University Press. 502–509.

Haspelmath, Martin (2007). Ditransitive alignment splits and inverse alignment. *Functions of Language* 14.1: 79–102.

Haspelmath, Martin (2008a). Creating economical morphosyntactic patterns in language change. In *Language Universals and Language Change*, ed. Jeff Good. Oxford: Oxford University Press. 185–214.

Haspelmath, Martin (2008b). Frequency versus iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19.1: 1–33.

Haspelmath, Martin and Ekkehard König (eds.) (1995). *Converbs in Cross-Linguistic Perspective*. Berlin, New York: Mouton de Gruyter.

Hawkins, John A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, John A. (2004). *Efficiency and Complexity in Grammars.* Oxford: Oxford University Press.

Hawkins, John A. (2007). Processing typology and why psychologists need to know about it. *New Ideas in Psychology* 25: 87–107.

Hawkins, John A. (2009). The typology of noun phrase structure from a processing perspective. Ms., Cambridge University. Available online at <http://linguistics.ucdavis.edu/People/jhawkins/recent-papers/NPStructureLingTypologyDouble.pdf>.

Hwang, Shin Ja J. (1995). Purpose clauses in Korean and iconicity. In: *Interfaces in Korean linguistics: 1993 Ohio State LSA Workshop*, ed. Chungmin Lee. Seoul: Thaehaksa. 125–142.

Kaltenböck, Gunther (2004). It-*extraposition and Non-extraposition in English.* Wien: Braumüller.

Kazenin, Konstantin I. (1994). Split syntactic ergativity: Toward an implicational hierarchy. *Sprachtypologie und Universalienforschung* 47: 78–98.

Keenan, Edward (1975). Variation in Universal Grammar. In: *Analyzing Variation in Language*, eds. Ralph Fasold and Roger Shuy, Washington, D.C.: Georgetown University Press. 138–148.

Keenan, Edward and Bernard Comrie (1977). Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8: 63–99.

Kenesei, István, Robert M. Vago and Anna Fenyvesi (1998). *Hungarian*. London, New York: Routledge.

König, Ekkehard and Johan van der Auwera (1990). Adverbial participles, gerunds and absolute constructions in the languages of Europe. In: *Towards a Typology of European Languages*, eds. Johannes Bechert, Giuliano Bernini and Claude Buridant. Berlin, New York: Mouton de Gruyter. 337–355.

Kortmann, Bernd (1991). *Free Adjuncts and Absolutes in English: Problems of Control and Interpretation.* London,New York: Routledge.

Kortmann, Bernd (1995). Adverbial participial clauses in English. In: *Converbs in Cross-Linguistic Perspective*, eds. Martin Haspelmath and Ekkehard König. Berlin, New York: Mouton de Gruyter. 189–237.

Lehmann, Christian (1984). *Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik.* Tübingen: Narr.

MacWhinney Brian and Csaba Pléh (1988). The processing of restrictive relative clauses in Hungarian. *Cognition* 29.2: 95–144.

Olawksy, Knut J. (2006). *A Grammar of Urarina*. Berlin: Mouton de Gruyter.

Reali, Florencia and Morten H. Christiansen (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language* 57: 1–23.

Reh, Mechthild (1985). *Die Krongo-Sprache (Nìinò Mó-Dì)*. Berlin: Dietrich Reimer Verlag.

Rice, Keren (1989). *A Grammar of Slave*. Berlin: Mouton de Gruyter.

Sawka, Kenneth S. (2001). *Aspects of Mayogo Grammar*. MA thesis, University of Texas at Arlington.

Schmidtke-Bode, Karsten (2009). *A Typology of Purpose Clauses*. Amsterdam, Philadelphia: John Benjamins.

Stefanowitsch, Anatol and Stefan Th. Gries (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2: 209–43.

Thompson, Sandra A. (1985). Grammar and written discourse: Initial vs. final purpose clauses in English. *Text* 5.1-2: 55–84.

Van Klinken, Catharina L. (1999). *A Grammar of the Fehan Dialect of Tetun, an Austronesian Language of West Timor.* Canberra: Australian National University.

Wanner, Eric and Michael Maratsos (1978). An ATN approach to comprehension. In *Linguistic Theory and Psychological Reality.*, eds. Moris Halle, Joan Bresnan and George A. Miller. Cambridge, MA: MIT Press.

Weber, David John (1989). *A Grammar of Huallaga (Huánuco) Quechua.* Berkeley: University of California Press.

Wiechmann, Daniel (2009). *Understanding Complex Constructions: A Quantitative Corpus Linguistic Approach to the Processing of English Relative Clauses.* PhD dissertation: University of Jena.

Witzlack-Makarevich, Alena (2007). Construction-centered database on grammatical relations. Paper presented at the 7th Biennial Meeting of ALT, Paris, September 25–28, 2007. Available online at < http://www.uni-leipzig.de/~witzlack/downloads.html>.

Yamashita, Hiroko and Franklin Chang (2001). "Long before short" preference in the production of a head-final language. *Cognition* 81.2: B45–B55.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.