Volkswagenstiftung
Förderinitiative Documentation of Endangered Languages –
Dokumentation bedrohter Sprachen (DoBeS)

The relative frequencies of nouns, pronouns, and verbs cross-linguistically

Applicant:
Frank Seifart

Co-applicants:
Hans-Jörg Bibiko
Balthasar Bickel
Swintha Danielsen
Roland Meyer
Sebastian Nordhoff
Brigitte Pakendorf
Alena Witzlack-Makarevich
Taras Zakharko
NN (computational linguistics Ph.D. student)

Table of contents

Summary

This project will investigate the relative frequencies of core parts of speech, such as nouns, verbs, and pronouns, in spoken language corpora of seven languages that represent a wide range of areal and typological diversity. We focus on two research questions:

(1) Why do languages vary so drastically in the relative frequencies of noun, pronoun, and verb tokens employed in discourse? Our pilot study for this project suggests that in some languages (such as Chintang) the overall number of nouns and pronouns taken together roughly equals the overall number of verbs, while in others (such as Sri Lanka Malay) this ratio is twice as high, i.e., the overall number of nouns and pronouns taken together is roughly double the overall number of verbs. What typological or other differences between languages can explain these differences in the use of parts of speech? One of the hypotheses we will test is the presence of argument indexing on verbs, which may make the overt realization of arguments as nouns or pronouns unnecessary, and may thus explain the low frequencies of nouns and pronouns in some languages.

(2) Why do the relative frequencies of nouns, pronouns, and verbs vary within texts? Our pilot study has shown that—consistently across languages—at the beginning of narrative texts, nouns are particularly frequently used, reflecting the introduction of new discourse participants, as expected. Furthermore, there are characteristic, sinusoidal alternations in the frequencies of noun use as narrative texts unfold, with regular peaks of heavy noun use roughly every 10-15 clauses. These peaks may reflect universal cognitive constraints on the activation of discourse participants, which necessitate their re-introduction by full lexical nouns after their activation has decayed, ultimately due to constraints of short-term memory.

We will also investigate the influence of further factors on the relative frequencies of nouns, pronouns, and verbs, such as the degree of speakers' and listeners' mutual acquaintance (known/familiar vs. unknown) and text genres. In this context we will empirically test the assumed universality of 'nouniness' of formal genres.

The newly available data compiled in the DoBeS framework allow us to develop and then appropriately address these new and exciting research questions for the first time, as they require data from diverse languages that are annotated for parts of speech by experts, time-aligned, and described with detailed metadata with respect to speakers' social status, mutual acquaintance, etc. These data allow us to capture subtle language usage patterns and explore their relation to typological differences between languages, narrative strategies, and other linguistic and non-linguistic factors. This project thus further develops documentary linguistics, connecting it with areas such as corpus linguistics, morphological typology, syntactic theory, discourse studies, and cognitive linguistics. In order to connect our findings with research on well-known languages such as English, we will additionally carry out analyses on published corpora of English.

The methods applied include advanced computational techniques for quantitative analysis of textual data of the type that has been produced by DoBeS projects, with as little additional manual annotation of data as possible. This permits us to analyze the huge amount of data necessary to detect and appropriately describe the subtle patterns under investigation. It will involve developing solutions for a number of technological and computational issues for cross-corpora studies, as additional valuable outcomes of this project.

Our team consists of experts in the seven languages who have compiled the corpora to be analyzed, and of linguists specializing in corpus linguistics and computational and quantitative linguists, who have unparalleled experience in the methods necessary for the current project.

1. Why study the relative frequencies of parts of speech?

1.1. An innovative line of research

The ratio of noun, verb, and pronoun tokens in discourse varies both within and across languages, but how exactly and why is so far largely unclear. The project proposed here aims to fill this gap by systematically describing this cross-linguistic variation for the first time, in order to uncover regular patterns within it. We will then investigate factors that explain various aspects of this variation by empirically testing correlations with morphosyntactic properties of languages, with social relations between speakers and listeners, and with cultural conventions of speech communities, such as genres and narrative strategies, as well as with general human cognitive constraints.

In addition to developing these new research questions, our cross-linguistic research will add a valuable new perspective on findings from a number of research strands, for which the ratio of nouns, verbs, and pronouns plays an important role, but which have so far focused almost exclusively on studies on individual languages. These research strands include stylistics, genre studies, and language acquisition.

The reason for the lack of cross-linguistic studies on the frequency of parts of speech is simple: the unavailability of appropriate cross-linguistic data. The situation is different now, after more than a decade of intensive language documentation activities, much of it in the DoBeS framework. These activities have resulted in large, time-aligned, electronic corpora of spoken language representing typologically and areally diverse languages, providing an appropriate database for the research we propose here.

We are thus now able to exploit the full potential of measurements of the relative frequencies of nouns, pronouns, and verbs to address fundamental research questions in a number of linguistic sub-disciplines, such as linguistic typology, text linguistics, sociolinguistics, and cognitive linguistics, with potential implications for linguistic relativity.

The research we propose is also innovative for undertaking quantitative cross-linguistic analyses on texts, rather than, for instance, grammatical features. Being one of the first studies of this kind, we focus on a feature of texts that is relatively simple to operationalize—the number of nouns, pronouns, and verbs. This feature is therefore appropriate for such a pioneering mass-comparison based on automatic counts in annotated data. In the future this approach may be extended to more complex phenomena, such as syntactic relations, etc., which will involve more extensive markup of the data and thus more intricate computational procedures.

For the coding of our data, i.e. the cross-linguistic identification of parts of speech, we rely on methods developed in recent descriptive-typological work. Our criteria are mainly derived from Croft's (1991; 2000) approach, combining the notion of prototypes with morphosyntactic, semantic, and discourse-pragmatic criteria. Many other, more strongly formalized approaches are not suitable for our purposes because they are not easily applicable cross-linguistically.

Based on a pilot study conducted by the members of the current proposal, reported in Seifart et al. (2010) and Seifart (2011), we propose here an in-depth study of two sets of research questions in connection with the relative frequencies of nouns, pronouns, and verbs (section 1.2-1.3), as well as an investigation of further factors (section 1.4).

1.2. Overall cross-linguistic variation and typological correlates

A first set of research questions focuses on the overall cross-linguistic variation in the relative frequencies of parts of speech and the factors that correlate with it, most prominently typological features such as argument marking systems.

In our pilot study we investigated the Amazonian languages Baure (Arawakan) and Bora (Boran), the Tibeto-Burman language Chintang, spoken in the Himalayas, the Tuu (Southern Khoisan) language N|uu, and Sri Lanka Malay (see section 3.1. for details on the data used in this pilot study). This study has yielded as preliminary results dramatic differences in the frequencies of nouns and pronouns relative to verbs, which are summarized in Figure 1, a so-called beanplot with mirrored density estimates (Kampstra 2008). For instance, in the Chintang corpus we used, approximately one noun or pronoun per verb is used, while in the Sri Lanka Malay data, approximately two nouns or pronouns per verb are used. Note that the standard calculation of such ratios (e.g., Stoll et al. 2010; Dhillon 2010) involves dividing the number of nouns and pronouns by that same number plus the number of verbs, to avoid division by zero for short stretches of text in which no verbs occur. Thereby we arrive at the figures of approximately 0.5 for Chintang (1/(1+1)=0.5) and approximately 0.667 for Sri Lanka Malay (2/(2+1)=0.667) given in Figure 1.

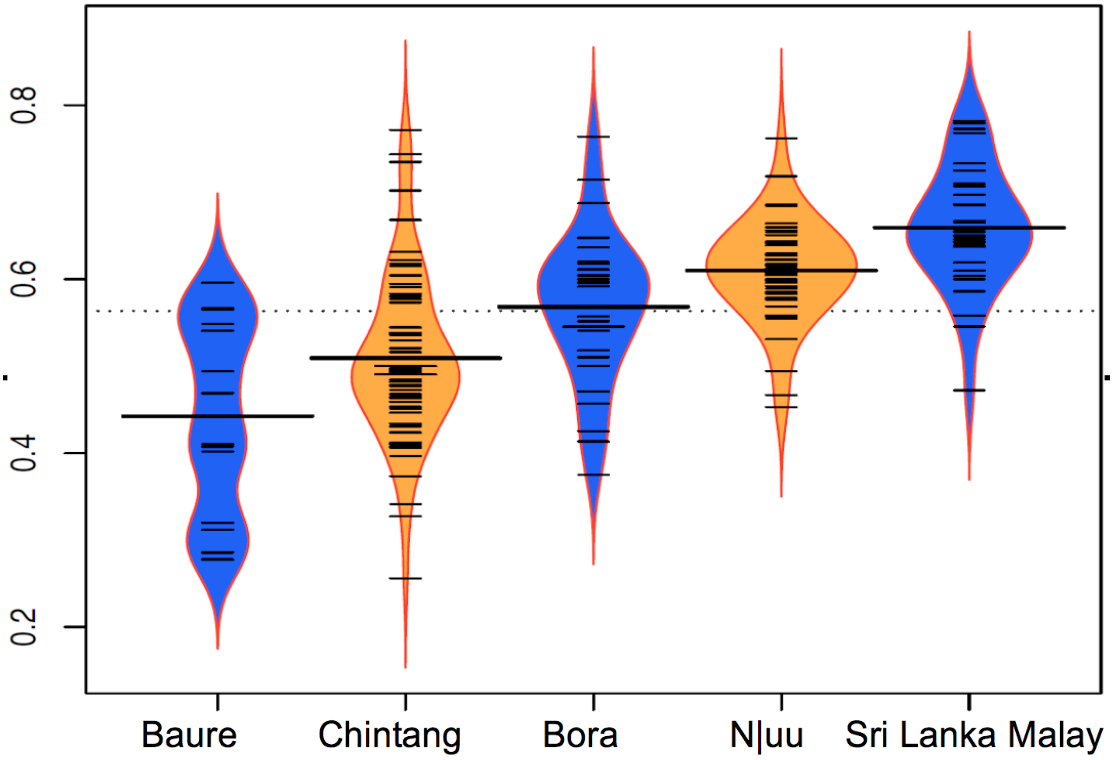


FIGURE 1: FREQUENCIES OF NOUNS AND PRONOUNS RELATIVE TO VERBS

In Figure 1, the mean frequencies in individual texts of a language are represented by short horizontal bars (—) and the mean frequency in all texts of a language is represented by long and thick horizontal bars (————). This representation clearly shows that the frequencies in different texts in individual languages vary considerably, e.g. some texts in Chintang have very low noun plus pronoun vs. verb ratios, while others have very high ratios. However, as indicated by the horizontal bulging of the beanplot in most languages, the intralinguistic variation in the relative frequencies of nouns plus pronouns vs. verbs tends to follow a normal distribution around the overall mean for each language (except for the Baure data considered in this pilot study, see below). This is indicative of the representativeness of the overall mean for a language as a valid measurement for cross-linguistic comparison.

However, the data represented in Figure 1 also reveal two important desiderata, to be followed up in the current project: First, the shape of the distribution of frequencies across individual texts in Baure—the smallest corpus considered in the pilot study—does not display the characteristic Gaussian concentration around the mean. This could point to different underlying distributions (as suggested by Bickel 2011 for argumental noun phrase ratios) or it could suggest that a certain minimum amount of data is necessary for this kind of study. As part of our project we will determine whether a general minimum corpus size can be fixed for the task at hand, in order to guarantee a distribution sufficiently close to normal. Otherwise, non-parametric methods of data analysis will have to be explored. Second, despite statistical non-significance in the data used in the pilot study, we are not convinced that genres have no detectable effect on the ratio of nouns, pronouns, and verbs generally, and therefore propose to further investigate the influence of genres (see section 1.4).

The precise description of the overall cross-linguistic variation of the relative frequencies of parts of speech—based on more extensive data for these five languages and from two further languages—within the proposed project will be an important contribution to our understanding of cross-linguistic variation of linguistic usage patterns. However, we would also like to find out what underlies such striking differences in the way speakers of different languages behave.

In an attempt to formulate appropriate hypotheses, we explored in our pilot study whether there is a correlation between the argument marking system of a language and the relative frequencies of nouns, pronouns, and verbs. The rationale behind this is that in languages with extensive argument marking on verbs, i.e. full paradigms of obligatory subject and object indices, less nouns and pronouns might be used, because the presence of these indices may make the realization of arguments as nouns or pronouns unnecessary. Our preliminary results seem to indicate that this typological characteristic of languages indeed correlates with frequencies of nouns, pronouns, and verbs: In languages such as Baure and Chintang, which have both subject and object indexing on verbs, relatively few nouns and pronouns are used (on the left in Figure 1). In a language like Bora, which has subject indices on verbs, but no object indexing, more nouns and pronouns are used. And finally, N|uu and Sri Lanka Malay (on the right in Figure 1) have no argument indexing on verbs at all, and accordingly use most nouns and pronouns.

Interestingly, the syntactic obligatoriness of arguments (i.e. *pro drop*) seems to be less closely correlated with the frequencies of nouns and pronouns: For instance, in Sri Lanka Malay, both subjects and objects are optional and can be dropped if the referents can be retrieved from context. Yet, in the Sri Lanka Malay corpus nouns and pronouns are very frequently used, suggesting that the option of dropping is not employed that often. Our use of real-world corpus data thus also allows us to empirically test impressionistic generalizations by field linguists (in this case co-applicant Sebastian Nordhoff), and correct them as necessary.

Our pilot study thus allows us to formulate the hypothesis that differences in the system of argument indexing on verbs of a languages—via the non-realization of arguments—may have an effect on the frequencies with which nouns and pronouns are used in that language. It suggests furthermore that this effect must be a fairly strong one, as it is clearly discernible in the overall frequencies of nouns and pronouns—i.e., including not only arguments, but also adjuncts—for which our automatic counting procedures on large amounts of data are most appropriate for.

These observations are highly suggestive and certainly warrant a careful and thorough follow-up in the context of the proposed project. One challenge will be to reconcile the observations from our pilot study with the findings in Bickel (2003) in which there is no significant effect of verbal morphology on the frequency of noun and pronoun use across a sample of three languages. It may be that the difference results from the kind of data (retellings of the Pear Story movie in Bickel [2003], cross-genre corpora here), from the different sample of languages (three Himalayan languages in Bickel [2003]), or the from the fact that Bickel's study focused on arguments alone, whereas here we include both arguments and adjuncts. In order to find out how the inclusion of adjuncts affects our results, we will also perform for the seven languages considered in this project limited case studies on arguments alone. Identification of arguments vs. non-arguments requires painstaking manual coding, which is clearly unattainable within our project for the total amount of data we consider. However, we will carry out such coding on limited, selected subsets drawn from the corpora of each language, to appropriately address the relevance of argumenthood for our results. If it then turns out that the cross-linguistic differences we observed in the pilot study are strongly influenced by the number of adjuncts—i.e. if the ratio of arguments vs. adjuncts also varies dramatically cross-linguistically—this would be another potentially highly relevant result, as it would suggest that we need to also reconsider the role of the functional balance between arguments and adjuncts.

1.3. Variation as narrative texts unfold

A second set of research questions addresses the variation of the relative frequencies of parts of speech within texts of individual languages. Our pilot study suggests that the frequencies of nouns and verbs are not uniformly or randomly distributed from the beginning until the end of narratives, but display regular patterns. Figure 2 summarizes measurements of the frequencies of nouns relative to verbs in progressive windows of five annotation units (window 1: unit 1-5, window 2: unit 2-6, etc.) and calculations of means for all narrative texts for each language.

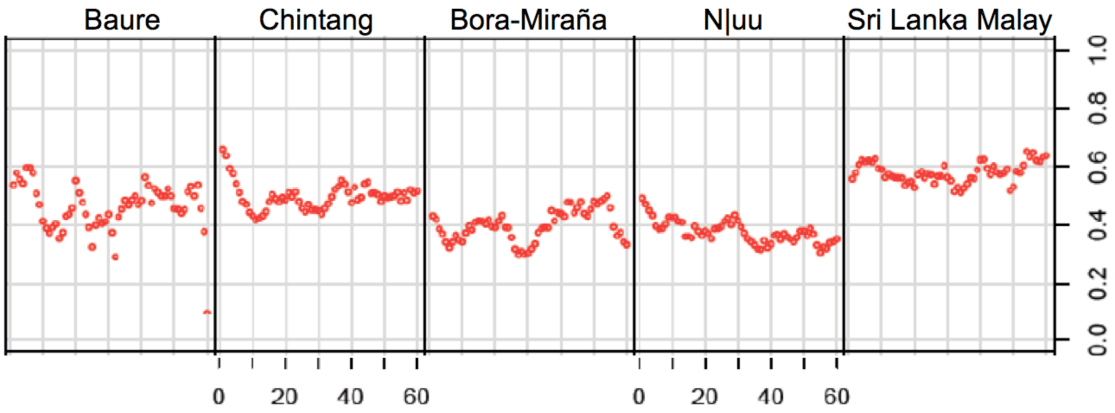


FIGURE 2: NOUN TO VERB RATIOS IN THE FIRST 60 ANNOTATION UNITS OF NARRATIVES

Figure 2 shows that there seems to be cross-linguistically a consistently higher noun to verb ratio at the beginning of narrative texts. This reflects the introduction of new discourse participants by means of nouns at the beginning of narrative texts, as one may expect—although this may not be a universal, as observed by Stoll and Bickel (2009). However, the most interesting observation in this pilot study—and the one that we will follow up in detail in the proposed project—is that there seem to be characteristic, sinusoidal alternations of the noun-to-verb ratio as narrative texts unfold, with peaks at about every 10-15 annotation units. We may hypothesize that these peaks in the noun-to-verb ratio reflect universal cognitive constraints on the activation of discourse participants, since discourse participants need to be re-introduced at regular intervals, often by full lexical nouns, when their activation status has decayed. The pattern we observe may thus ultimately be due to constraints of short-term memory

In Figure 2, measurements were carried out in progressive windows of annotation units. These annotation units are the basic units into which the narrative texts in our corpora are segmented for practical purposes in the ELAN or Toolbox software. They roughly correspond to intonation units, paragraphs, syntactic sentences, or a combination of these. These preliminary measurements already reveal the striking patterns given in Figure 2.

To avoid potential noise from inconsistencies in the segmentation into annotation units across corpora, and—more importantly—to address further research questions, we will carry out refined measurements of the periodic increase and decrease of the frequencies of nouns in two ways: Firstly, instead of in steps defined by annotation units, we will measure the frequencies of nouns in steps defined by verbs. This will involve, for instance, measurements of how many nouns are used between the first verb token and the 10th verb token of a text, between the second and the 11th verb token, between the third and the 12th verb token, etc. Secondly, we will measure the frequencies of nouns in steps defined by time, e.g., the number of nouns per verb in the first 20 seconds of a text, in the interval between second 5 to 25, in the interval between second 10 and 30, etc. First pilot studies with these two methods suggest that with both, the frequency of nouns relative to verbs is also far from evenly or randomly distributed and follows regular patterns of increase and decrease.

If the increase and decrease reflect properties of cognitive processes related to the overall management of nominal reference then the comparison of these two measurements can give us insights into the units on which these cognitive processes operate: are they more sensitive to actual time and relatively independent of speech rate? Or are they more sensitive to the actual flow of information and relatively independent of how densely this information is packed into time?

Note that by counting all noun occurrences, we are not simply looking at topic continuity in discourse (see, e.g., Givón 1983). Rather, our quantitative analyses aim at uncovering patterns in the overall distribution of nominal reference relative to verbal meanings, which are potentially more relevant for cognitive processes and language processing than linguistically-defined notions like topic.

An important desideratum for this project is the precise description of the curves that represent the increase and decrease of noun use in terms. This will involve descriptions of the phase length between peaks and the exact shapes of the curves, e.g., the steepness of the increase, and the cross-linguistic differences with respect to these properties. These differences may reflect culture-specific narrative traditions, and will help us to evaluate the potential universality of some features of the variation in the relative frequencies of nouns and verbs as narrative texts unfold.

In sum, our data and our methods, once these are fully developed, promise to reveal intriguing cross-linguistic patterns in narrative texts. Being so subtle, these patterns are only detectable by advanced quantitative methods, involving large amounts of textual data and measurements in progressive windows. They may offer important insights into the interplay between linguistic behavior, human cognitive processes, and narrative traditions.

1.4. Further factors

As mentioned above, the influence of genres is very likely to be an important factor in the variation of the relative frequencies of nouns, pronouns, and verbs. We will investigate this issue in detail by refined definitions of genres for each language and possibly the inclusion of a limited amount of additional data for individual languages in order to ensure a minimum of data for each genre.

As a second additional factor, we will investigate the influence of the sociological background of speakers and listeners on the frequency of noun and pronoun use, following up findings by Bickel (2006; 2011). The specific hypothesis we will further test is whether the degree of acquaintance of the speaker and the listener correlates with the use of fewer nouns, which may be dropped because of a high degree of shared knowledge. Information relevant to assess the closeness of acquaintance includes information on kinship relations, location, and the general size of the society. Since all corpora have been obtained in fieldwork by team members of this project, we can count on detailed information on all these parameters for each language.

Thirdly, we will investigate the potential influence of intralinguistic variation in the degree of synthesis (i.e. the number of morphemes per word) on the frequency of nouns and pronouns. That is, we will measure the morphological complexity of verbs in relatively small segments of texts, defined by either the number of verbs or time (see section 1.3), and then test whether this local index of synthesis correlates with the frequency of nouns and pronouns in that segment. Since our data are segmented into morphemes, these measurements can be easily done automatically. The idea behind this is that a higher degree of synthesis of verbs correlates with more information about arguments given in argument indices in verbs, which we will first examine based on manual coding of subsets of data. If that is the case, we can use automatic counting procedures on large amounts of data to pick up signals of an increase of information about arguments given in verbs, even though these procedures are blind to the distinction between, for instance, tense-aspect-mood markers and argument indices. Comparison of these measurements may thus give insights into the relationship between morphological properties of verbs and the frequency of noun and pronoun use not only across languages, but also within languages.

Finally, we will follow up research on the influence of case-based vs. non-case-based agreement whose influence on the frequency of argumental noun phrases has been demonstrated in Bickel (2003), a study that redefined the state of the art in research on noun phrase usage in discourse (see also section 2.2).

2. Connections to other research

Our proposed research connects to a number of research areas for which the use of nouns, pronouns, and verbs is a central issue. It builds upon findings from these and at the same time has important implications for them. These include research on *pro drop* (section 2.1), referential density (section 2.2.), and a few others (section 2.3). By addressing relevant issues in these areas, our project will also help to further raise the visibility of DoBeS in the general linguistic community.

2.1 Pro drop and syntactic theory

Underlying most thinking on *pro drop* (see also section 1.2) is the old idea that subject marking on verbs stands in for the subject, that expressing it twice would be redundant, and that therefore subjects are not obligatory in languages with subject marking on verbs (Apollonius Dyscolus: book 1, §17, in Householder 1981:25; Duponceau 1819:xxix; Jespersen 1924:213; Taraldsen 1978; Van Valin and LaPolla 1997:330–335). While the traditionally assumed close correlation between verbal morphology and the realization of pronominal and nominal arguments has been known to break down in many languages at least since Gilligan (1987) (see also, e.g., Huang 1984:6–13; Roberts and Holmberg 2009:5–13), it continues to play a highly relevant role in syntactic theories, as shown, for instance, by the recent controversy between Newmeyer vs. Roberts and Holmberg (Newmeyer 2004; Newmeyer 2006; Roberts and Holmberg 2006; Roberts and Holmberg 2009). Typological research with a strong empirical base, as it is to be conducted in the present project, will make a significant contribution here. We intend to widen the perspective on this specific issue by doing justice to actual usage data, by taking into account not only subjects, but all arguments (in addition to adjuncts), and by relating the discussion to general noun/verb marking.

Thus, while research on *pro drop* focuses on the relation between two structural properties of language, namely verbal morphology and syntactic obligatoriness of arguments, the project proposed here will instead investigate the relationship between verbal morphology and the frequency of use of nouns and pronouns, i.e. a property of discourse. This will help to reframe this basic relationship in a perspective of fine-grained patterns of actual language use, beyond the binary *pro drop* vs. *non pro drop* distinction.

2.2. Referential density and linguistic relativity

Our proposed research potentially has important implications for linguistic relativity, similarly to related research on referential density (Bickel 2003:733; Stoll and Bickel 2009). There are good reasons to hypothesize that frequent use of nouns and pronouns in discourse entails heightened attention to participants, and the frequent use of verbs in discourse heightened attention to internal structure of events. Preliminary evidence for this comes from Stoll and Bickel (2009), who show that differences in referential density extend to differences in the extent to which speakers use lexical noun phrases and the extent to which they explicitly describe the characters in a story: speakers of low referential-density languages tend to minimize descriptions of characters while high referential-density languages seem to invite speakers to describe each character in close detail, and often going far beyond what actual stimuli contain.

Whether referential density is (co-)determined by the kind of agreement system that the language has (case-based vs. non-case-based; as proposed by Bickel 2003) or if referential density is (co-)determined by the elaboration of agreement morphology (as suggested by the pilot study), under each scenario we end up with a finding that some structural property of grammar has an impact on narrative style and thereby possibly also on how we balance our attention outside language. This would constitute a very strong case of linguistic relativity, relating core morphosyntax to cognition (unlike, e.g., better established relativity effects that relate lexical features, e.g. in the coding of spatial relations, to cognition). Experimental psychological research on such effects is beyond the scope of the present project but detailed research on discourse is an essential preliminary for such further explorations of linguistic relativity—here and in general (see, e.g., Pederson et al. 1998).

2.3. Further areas

Our research also connects to various aspects of existing corpus-linguistic research and some psycholinguistic research. None of these have investigated cross-linguistic variation to a significant extent, with the possible exception of language acquisition studies. Our research will thus help to put findings from these research areas into a wider perspective.

In corpus linguistics, the relative frequencies of parts of speech is used as a marker for text genres (Biber, Conrad, and Reppen 1998:68–69; Gaenzle et al. 2010), as well as for text styles, including the identification of author styles (Biber, Conrad, and Reppen 1998:66; Tuldava 2005:377). From this research, we will follow up in particular the repeated reports that in a number of individual languages, nouns are more frequently used in more formal genres than in other genres, maybe because nouns have less face-threatening potential than verbs (Brown and Levinson 1987:208). The carefully and uniformly annotated data from seven diverse languages represented in our project provides an excellent empirical basis to test to what extent this holds cross-linguistically.

The ratio of nouns to verbs has also been used in two strands of psycholinguistic research. Firstly, language acquisition research has investigated the development of this ratio in early childhood with contradicting results. Some scholars argue for a universal noun bias in early childhood, and others against it, making this a hotly debated issue (Gentner 1982; Stoll et al. 2010). Secondly, the ratio of nouns to verbs is also important in research on pathological language attrition such as in aphasia (Miceli et al. 1984; Schnitzer 1989:154; Bird et al. 2000; Thompson, Fix, and Gitelman 2002), which suggests that in these situations, more nouns are used. Both of these psycholinguistic research areas should significantly benefit from a better understanding of the cross-linguistic variation in the noun-to-verb ratio that our project will provide, since this will clarify an important parameter used in that research.

3. Data

3.1. The sample of languages

The corpora used in the project proposed here include five corpora compiled in the DoBeS framework, in addition to one compiled with funding from the Hans Rausing Endangered Languages Project (ELDP) (Güldemann et al. 2010), and one compiled by Pakendorf (Pakendorf 2007) with funding from the Wenner-Gren Foundation for Anthropological Research, Inc. and the Max Planck Society (Table 1). These seven languages represent, on the one hand, vast areal diversity. The inclusion of two languages from two broad areas each (Amazonia and Siberia) will allow for some initial hypotheses about the effects of areal vicinity in the variation of frequencies of parts of speech, e.g., through the influence of local narrative traditions. On the other hand, the languages chosen represent vast typological diversity, for instance in terms of their argument indexing systems (see section 1.2). The overall index of synthesis (calculated as the mean number of morphemes per word) given for each language in Table 1 is also indicative of their typological heterogeneity, ranging from nearly perfectly isolating languages such as N|uu to fairly polysynthetic languages such as Ėven and Bora.

Language (Family) (data source)	**Region**	**Number of speakers**	**Index of syn-thesis**	**Total number of words**	**Words with PoS annota-tion (8/2011)**
Baure (Arawakan) (Danielsen et al. 2009)	Amazonia	84	1.78	~35,000	~25,000
Chintang (Sino-Tibetan) (Bickel et al. 2011)	Himalaya	~ 1,500	1.96	293,506	193,744
Bora (Boran) (Seifart 2009)	Amazonia	~ 1,500	2.34	~90,000	13,069
N|uu (Southern Khoisan) (Güldemann et al. 2010)	South Africa	6	1.11	100,000	31,691
Sri Lanka Malay (Austronesian) (Ansaldo, Lim, and Nordhoff 2009)	Sri Lanka	~ 45,000	1.52	22,904	14,050
Ėven (Tungusic) (Pakendorf et al. 2010)	Siberia	~ 2,500	2.16	41,700	41,700
Sakha (Yakut) (Turkic) (Pakendorf 2007)	Siberia	~ 360,000	1.81	30,600	30,600

TABLE 1: CORPORA USED IN THIS PROJECT

All corpora include texts from various narrative genres, as well as conversation, procedural texts, and some songs, or at least—in the case of Sakha—various kinds of other spoken genres such as oral life histories. They include a range of speakers, both male and female, with different social statuses and from different locations where the languages are spoken.

We aim at a minimum of 30,000 words for each language that are fully annotated for parts of speech by the end of the first year of the project (i.e., by May 2013) and that can thus be fully included in the quantitative analyses during the second and third years of the project. Although some corpora seem to be far from this mark, this goal is feasible for different reasons. For Bora and Baure, we can count on a number of trained student assistants that continuously annotate our data with increasing speed. For instance, the student-assistant annotator for Bora is currently already able to annotate at least 1,000 words per month, with increasing speed of annotation. By the beginning of the project, in June 2012, at the very least 20,000 Bora words will thus be annotated, and the annotation of the rest by the end of the first project year should pose no problem at all, given an increased labor force by additional assistants hired within the project. For Baure, we can count on a whole group of trained annotators who continuously annotate Baure data and who are available to continue doing so in the context of the proposed project. For Sri Lanka Malay, various hours of recordings of additional data have been transcribed in the context of a project funded by the National Science Foundation. These data will become available later this year.

All corpora, except Ėven and Sakha, are entirely or almost entirely (at least 90%) time-aligned with video or audio recordings, so they can be directly used for the analyses discussed in section 1.3, above. About 60% of Ėven data are time-aligned, i.e. a sufficient amount to include Ėven in the analyses that require time-alignment. Sakha data are not time-aligned yet. Student assistants will time-align a selected subcorpus of Sakha, and possibly some additional Ėven data, during the first year, a task that can be relatively easily done even by assistants who do not know the language, as our experience has shown. Therefore we will be able to include Sakha data also in at least some of the analyses that require time-aligned data.

3.2. Why use DoBeS corpora?

The data from the DoBeS archive that we will use have some unique properties that make the kind of research proposed here possible for the first time. Firstly, the data are in a sufficiently standardized, well-structured format, the interchangeable ELAN and Toolbox file formats (*.eaf and *.tbt) and include time-alignment information. This makes it possible to relatively easily convert them into a format that allows for the comparative quantitative analyses proposed here (see section 4 on details of this method).

Secondly, the data are accompanied with detailed metadata in the IMDI format containing information on text genres and the participants involved in the speech situation. This detailed information in the metadata, together with knowledge of the team members on further aspects of the linguistic, cultural, social, and personal context of each recording, allows us to closely control for a wide range of possible factors in our analyses, such as genres, social status of speakers, etc.

3.3. Incorporating English corpora

DoBeS corpora have a number of features that are necessary for the research proposed here, and they offer the advantage that corpora of a wide variety of languages follow the same standard. Therefore our project's primary focuses is on these corpora. However, we will include in at least some of our analyses also corpora of English, for two main reasons. Firstly, we want to optimally connect our research with the vast literature on relevant aspects of English, including corpus-linguistic research on the frequencies of parts-of-speech (e.g., Biber, Conrad, and Reppen 1998:66–69) and syntactic research and on argument obligatoriness and realization (see section 2.1. for references). Second, we want to make use of the huge and publically available corpora of English (e.g., Davies 2004; Davies 2008; Nelson 2011) in order to enhance our sample of languages for cross-linguistic comparison.

Extending our investigation to corpora of English poses a number of challenges. One is that part-of-speech tagging in these was mostly done automatically, and its quality is sometimes difficult to assess. Also, detailed metadata is often lacking that would allow carrying out detailed testing for the effects of genres, speaker identity and mutual acquaintance. In addition, the spoken part of these corpora tends to be small and is often made up to a large extent of non-spontaneous genres such as news broadcasts. For these reasons the extension of our research to English will focus on the fundamental question of the correlation between overall frequencies of parts of speech and typological features of languages (see section 1.2), and not so much on the analyses that require detailed information on, e.g., genres, speakers' mutual acquaintance, and time-aligned data (see sections 1.3-1.4).

The analyses of English also will serve as a stepping stone for the incorporation of corpora from other languages into further cross-corpora studies, e.g. Italian (e.g., Schneider 2003), Spanish (e.g., Davies 2002), and Portuguese (e.g., Davies and Ferreira 2006).

4. Methods

4.1. Data preparation and coding

As laid out in section 3.1, the great majority of our data already include part-of-speech tags, as the most important kind of annotation for the current project. For some languages (Chintang, Sri Lanka Malay, Sakha, and Ėven), part-of-speech information currently resides mainly in lexica and morpheme lists, and is not necessarily written into the files that contain the texts. These text files do, however, include segmentation into morphemes and morpheme glosses. Our task for these corpora is to fill in part-of-speech information by automatically matching morphemes in texts with morphemes in lexica and morpheme lists and automatically writing the part-of-speech tags from the lexica and morpheme lists into the files containing the texts. This process may involve some complications that require additional programming, such as the recoding of words containing noun roots if they include a verbalizing morpheme or the disambiguation of the part-of-speech tag for polyfunctional roots by combinations of such roots with word-class specific morphology. However, for this task we can build on experience and pieces of code that have already been developed for Chintang and Sri Lanka Malay.

We will also, in a very early stage of the project, thoroughly reassess the criteria for the cross-linguistic identification of parts of speech applied in the coding of our data, based on the joint expertise in descriptive-typological linguistics represented by the team members. This is especially important for potentially controversial cases such as locative nouns used as adpositions or cliticized pronouns. We will then amend individual coding choices where necessary. This can be done by automatic procedures on data that are already annotated. As mentioned above, some additional Bora and Baure data will be annotated during the first year, using existing setups within the Toolbox software. We will also reassess the characterization of argument indexing systems, again paying special attention to cliticized pronouns.

All data will be imported into the R statistical software package, further developing scripts developed by Bickel and Zakharko. Relevant information from the IMDI metadata, such as information on genres and speakers, will be extracted automatically from the IMDI metadata files and linked to the corresponding texts within R, again based on scripts developed by Bickel and Zakharko.

Our philosophy for the processing of data and the development of computational techniques is to use modular structures instead of complex aggregations. For instance, we apply a number of relatively small, self-contained scripts for specific purposes, instead of a small number of large and complex scripts for complex procedures. This facilitates the manipulation of individual aspects of our procedures as the necessity arises, and the re-use of these scripts for other purposes.

4.2. Analyses

During the first year of the project, we will carry out further hypothesis-generating analyses in order to formulate more precise hypotheses about the factors that influence the variation in relative frequencies of nouns, pronouns, and verbs. We will then test these hypotheses using standard statistical modeling strategies, paying particular attention to careful assessment of what distributions can be assumed (normal, uniform) and to state-of-the art techniques for controlling individual variation (e.g. mixed models)

For the characterization of the curves that describe the increase and decrease of the relative frequencies of parts of speech as narrative texts unfold (see section 1.3), we will use various models, among them models from thermodynamics which are highly developed to cope with these kinds of complex distributions.

4.3. An international workshop

We will hold an international workshop in Leipzig to gain a broad overview of the state of the art of research on the relative frequencies of parts of speech in different linguistic subdisciplines. It will involve two invited speakers as well as a call for papers aimed at specialists in corpus linguistics, language acquisition, genre studies, stylistics, and other linguistic subdisciplines to accommodate a total of about 20 presentations, including presentations by team members. This workshop will help to clarify some fundamental issues for our project work and at the same time serve to disseminate first results from our project. It will include a special session to discuss among the participants future directions of cross-disciplinary corpus-based research on relative frequencies of parts of speech and more generally.

5. Projected outcomes

5.1. Scientific publications

Team members will present results of our research at international conferences during the project, for instance, at the International Congress of Linguistics in Geneva in July 2013 and at the Biannual Conference of the Association for Linguistic Typology in Leipzig in August 2013, and other conference that are still to be announced, including conferences specializing in quantitative and corpus linguistics.

Based on these presentations, we will produce a set of co-authored papers reporting on results of our research to be published in journals specializing in computational, quantitative, and corpus linguistics, on typology, on discourse studies, text linguistics and stylistics, and on theoretical linguistics. In addition, our Ph.D. student will complete a Ph.D. thesis, which will include a thorough discussion of the methodological issues involved in this project, and a detailed treatment of at least one research question. In both our presentations and publications we will ensure proper acknowledgement of all colleagues and native speakers that have contributed data used by us, including those that are not team members.

We also aim at publishing an edited volume with about 10 selected papers from the international workshop organized by us during this project, including an introductory chapter by team members summarizing the state of the art of research on the relative frequencies of parts of speech.

5.2. Enhanced data

Another outcome will be a large set of cross-linguistic data with consistent, detailed annotations with respect to morpheme glosses and part-of-speech information, of at least 30,000 words each for seven languages. This set of data can be used for a variety of other cross-corpora studies, which require cross-linguistically consistent annotation, for instance corpus-based research on clause combining or on the distribution of tense-aspect-mood-evidentiality markers, among many others. These data will be stored under a separate node in the central archive of the DoBeS program at the Max Planck Institute for Psycholinguistics in Nijmegen, with links to the original collections in that same archive and elsewhere.

5.3. Methodological contribution

In the course of this project we will develop solutions for a number of important methodological issues in quantitative cross-corpora studies. These include improved techniques to import data into R and techniques for making consistent changes to part-of-speech tags. We will thus contribute to building a research infrastructure for the novel field of cross-corpora studies, providing services such as enhancing the interoperability of software such as ELAN and R. Any piece of code we develop in this project will be published online and will thus be publically available, as already practiced by Balthasar Bickel (see http://www.spw.uzh.ch/software.html).

6. Team members, collaborations, and the institutional setting

6.1. Team members and their expertise

The applicant and co-applicants of this project jointly represent a large amount of expertise in both typological description of languages and in quantitative analyses of corpora (for details, see the CVs in the appendix).

Balthasar Bickel, Swintha Danielsen, Sebastian Nordhoff, Brigitte Pakendorf, Frank Seifart, and Alena Witzlack-Makarevich were all directly and centrally involved in collecting the data of the languages of their expertise during extended field work on site in language communities, if they were not the sole collectors (in the cases of Nordhoff, Pakendorf with respect to Sakha, and Seifart). They have all published numerous descriptive studies on these languages (see the lists of publications in the CVs for the most important ones, in addition to Ernszt et al. in prep.; Ernszt et al. 2011; Ernszt, Güldemann, and Witzlack-Makarevich to appear; Ernszt, Siegmund, and Witzlack-Makarevich 2008). Danielsen and Nordhoff have produced full-length reference grammars of the languages of their expertise (Danielsen 2007; Nordhoff 2009). Team members have also previously published on issues relating to parts of speech and on relevant aspects of argument indexing (for instance, Nordhoff 2004; Witzlack-Makarevich 2006; Witzlack-Makarevich 2011).

Hans-Jörg Bibiko, Balthasar Bickel, Roland Meyer, and Taras Zakharko are experts in computational, quantitative, and corpus linguistics. Balthasar Bickel is additionally a leading typologist and Roland Meyer has additionally a strong interest in syntactic theory, especially with applications in Slavic languages. They have published numerous contributions to computational, quantitative, and corpus linguistics (e.g., Bickel and Stoll 2008; Bickel in press; Kallmeyr, Meyer, and Wagner 2001; Meyer 2003; Meyer 2011).

As an additional team member, we will hire a Ph.D. student with a background in both computer sciences and linguistics. S/he will be expected to work half time for the project and will be given the opportunity to work on a Ph.D. thesis directly related to this project in the remaining time. We will widely publicize a vacancy notice in addition to doing active searching ourselves. We will contact our colleagues at the Abteilung Automatische Sprachverarbeitung (department of Natural language processing) of the Universität Leipzig, for possible candidates, and colleagues at German universities that have departments specializing in corpus linguistics such as Regensburg, Humboldt University Berlin. Informal inquiries among these colleagues in August 2011 have shown that there is a very good chance of finding a number of highly qualified candidates among which we will be able chose one.

S/he will receive a three-year contract from the Max Planck Institute for Evolutionary Anthropology, where s/he will have a workspace and access to superb research infrastructure. S/he will receive intensive training and continued supervision throughout the project in the computational methods required for this project primarily from Hans-Jörg Bibiko, Balthasar Bickel, Roland Meyer, and Taras

Zakharko. S/he will receive advice in the linguistic issues involved from the linguists among the team members, primarily from Frank Seifart at the Max Planck Institute for Evolutionary Anthropology, where s/he will work. Since the Max Planck Institute for Evolutionary Anthropology cannot award Ph.D. degrees, s/he will enroll as a Ph.D. student in a German university with a program that has a focus on computational linguistics, where s/he will have an additional Ph.D. advisor.

We will hire a number of student assistants for the labor-intensive tasks of annotation, time-alignment, and to some extent identification of arguments. We will offer contracts to student assistants who are already trained in annotating Bora and Baure data and who will thus be able to fulfill these tasks very efficiently, and we will actively search for further student assistants, primarily at the University of Leipzig, where we have already identified a number of suitable candidates in the context of our teaching activities. As with the already trained student assistants, we will hire additional students who are also interested in using the data they will process for research projects within their studies, especially for B.A. and M.A. theses. In this way we further promote young researchers in cross-corpora linguistic analyses and encourage such research at German universities beyond our project.

6.2. Previous collaboration between team members and beyond

Team members have previously successfully collaborated on a number of projects related to the current proposal. In 2010, all team members were located in Leipzig, where they founded the informal collaborative project "Comparative Corpus Research in Leipzig" (see http://lingweb.eva.mpg.de/compacorp) which has held regular meetings and carried out the pilot study reported in Seifart et al. (2010) and Seifart (2011). The current project proposal stems directly from that collaboration. In early 2011, Bickel, Witzlack-Makarevich, and Zakharko have moved to the University of Zurich, so that our team has now two main institutional bases, in addition to Pakendorf joining the CNRS Laboratoire Dynamique du Langage in Lyon in early 2012.

In addition team members have previously collaborated on a number of research projects with direct relevance for the current proposal. For instance, Meyer and Seifart are working on automatic tagging procedures for Bora (Meyer and Seifart 2011), Seifart and Pakendorf are collaborating on a corpus study on borrowed affixes (Seifart, Pakendorf, and Steinkrüger in prep.), and Bickel and Zakharko collaborate with Volker Gast and colleagues at the University of Jena on a DFG-funded project entitled "Towards a corpus-based typology of clause linkage: An analytical framework and case studies on non-local dependencies" (see http://www.personal.uni-jena.de/~mu65qev/LinkType/index.html), one of the first projects that explicitly applies modern quantitative corpus-linguistic methods to a DoBeS corpus. Our project will closely collaborate with that project in order to make full use of synergy effects in developing quantitative methods specifically aimed at the analyses of DoBeS corpora.

Seifart is organizing a workshop entitled "Potentials of Language Documentation: Methods, Analyses, and Utilization" (see http://www.eva.mpg.de/lingua/conference/2011_DOBES/), funded by the Volkswagen Foundation and to be held in Leipzig in November 2011. This workshop will include panels on how computational methods can efficiently enhance the annotations and improve the analyses of linguistic data, on how language documentation data can change important aspects of analyses in linguistics and related disciplines, and on how language documentation data can be utilized for research, language maintenance, and to raise awareness of linguistic diversity. It will thus directly contribute to clarifying some fundamental issues in the research we propose here. Bibiko and Bickel are also centrally involved in this workshop, as well as members of The Language Archive unit at the MPI for Psycholinguistics which houses the central DoBeS archive.

Team members have also collaborated with members of The Language Archive unit in various publications (Berck et al. 2006; Trilsbeek et al. 2007). We aim to continue close collaboration with The Language Archive with respect to tools that are currently being developed there, in particular ELAN and its extension for corpus-linguistic queries.

In sum, the team of applicant and co-applicants form a closed-knit research network, in which they have already jointly addressed a number of issues that are fundamental for the research proposed here. In addition, we are actively engaged in collaborations with various relevant institutions and related projects that will greatly facilitate our project work.

6.3. Institutional setting

The Max Planck Institute for Evolutionary Anthropology offers an ideal institutional setting for this project for its exceptional research infrastructure and the expertise of the members of the department of linguistics in some of the issues that are important for our research. This includes the valency classes project, led by Martin Haspelmath and Andrej Malchukov, which has relevance for the realization of arguments, as well as unparalleled expertise in typological research and linguistic databases, such as the World Atlas of Language Structures WALS (Haspelmath et al. 2005; Dryer and Haspelmath 2011).

The Max Planck Institute for Evolutionary Anthropology is also highly suited for the international workshop we propose to organize within this project, as it can count on experienced personnel for the organization of such events.

7. Working plan and time schedule

7.1 Working plan

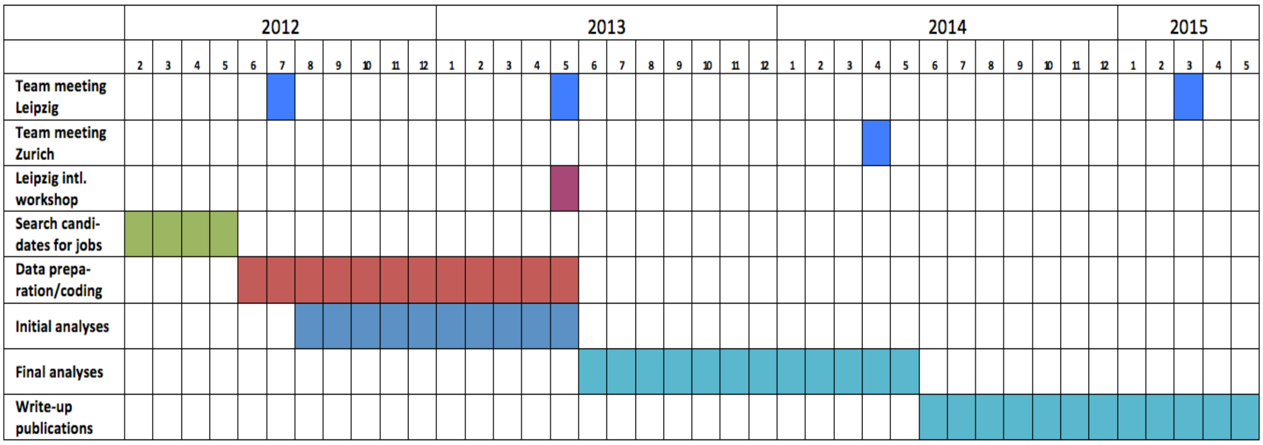
During the first year of the project we will concentrate on reassessing our coding choices, amending existing annotations accordingly, and annotating new data, in order to have the fully annotated dataset ready by the end of the first year. The annotation of new data will be carried out by up to five student assistants during the first year. We plan to continue to employ two student assistants in the second and third years for further data curation.

During the first year we will also continue to carry out hypothesis-generating analyses, continuously refine our hypotheses accordingly, and begin hypothesis-testing analyses, which we will continue to do on the complete set of data primarily during the second year, in order to interpret the results and write up publications primarily during the third year.

At the end of the first year, the international workshop on the relative frequencies of parts of speech that our project will organize will take place in Leipzig. Among other things, this workshop will help us to appropriately address issues currently debated in various linguistic subdisciplines and neighboring disciplines.

The team members will interact closely throughout the duration of the project, facilitated by the use of a password-protected project WIKI. All team members will meet four times during the project, for in-depth discussions and planning our research activities. The first of these meetings, during which we will develop a detailed plan for the project, will take place within a month of the beginning of the project in Leipzig. The second meeting will be adjacent to the international workshop and mainly serve to discuss our hypotheses-testing procedures. We will hold our third general meeting towards the end of the second year in Zurich, to discuss our joint publications and presentations, among other issues. A final meeting, again in Leipzig, will serve to wrap up the project work, including the finalization of publications.

7.2. Schedule details



	2012											2013												2014												2015				
	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
Team meeting Leipzig						🟦										🟦																						🟦		
Team meeting Zurich																											🟦													
Leipzig intl. workshop																🟪																								
Search candi-dates for jobs	🟩	🟩	🟩	🟩																																				
Data prepa-ration/coding					🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥																								
Initial analyses							🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦																								
Final analyses																	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦												
Write-up publications																													🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

References

Ansaldo, Umberto, Lisa Lim, and Sebastian Nordhoff (eds.) 2009. *Sri Lanka Malay Documentation*. Nijmegen: DOBES-MPI. http://www.mpi.nl/DOBES/projects/slm.

Berck, Peter, Hans-Jörg Bibiko, Marc Kemps-Snijders, Albert Russel, and Peter Wittenburg. 2006. Ontology-based language archive utilization. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2295-2298. Genoa.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bickel, Balthasar, and Sabine Stoll. 2008. Quantitative Analysis of DOBES Corpora Using R. Paper presented at the DOBES Workshop on Language Documentation Methods, Nijmegen, June 13, 2008. Nijmegen. http://www.uni-leipzig.de/~bickel/research/presentations/dobesmtg2008.pdf.

Bickel, Balthasar, Sabine Stoll, Martin Gaenszle, Novel Kishore Rai, Elena Lieven, Goma Banjade, Toya Nath Bhatta, et al. (eds.) 2011. *Audiovisual corpus of the Chintang language, including a longitudinal corpus of language acquisition by six children: ca. 650,000 words transcribed and translated, of which ca. 450,000 glossed, plus paradigm sets and grammar sketches, ethnographic descriptions, photographs*. Nijmegen, Leipzig: DoBeS, Universität Leipzig. http://www.mpi.nl/DOBES.

Bickel, Balthasar. in press. Statistical modeling of language universals. *Linguistic Typology*. http://www.uzh.ch/spw/bickel/papers/bickel2011universals.pdf.

--- 2003. Referential density in discourse and syntactic typology. *Language* 79.708-736.

--- 2006. Referential density in typological perspective. Plenary talk, Leipzig Spring School on Linguistic Diversity, March 22, 2006. Leipzig. http://www.uzh.ch/spw/bickel/presentations/rd2006.ppt.pdf.

--- 2011. The role of genealogical units in explaining linguistic distributions: a case study on referential density. Talk given at Workshop “Cross-linguistic and language-internal variation in text and speech: focus on the joint analysis of multiple characteristics”, Freiburg, Germany, 9-11 February 2011. Freiburg. http://www.uzh.ch/spw/bickel/presentations/pears2011.pdf.

Bird, Helen, Matthew A. Lambon Ralph, Karalyn Patterson, and John R. Hodges. 2000. The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. *Brain and Language* 73.17-49.

Brown, Penelope, and Stephen C Levinson. 1987. *Politeness : some universals in language usage*. Studies in Interactional Sociolinguistics ; 4. Cambridge: Cambridge University Press.

Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press.

--- 2000. Parts of speech as typological universals and as language particular categories. *Approaches to the typology of word classes*, ed. by Petra M. Vogel and Bernard Comrie, 65-102. Berlin, New York: Mouton de Gruyter.

Danielsen, Swintha, Franziska Riedel, Femmy Admiraal, and Lena Terhart (eds.) 2009. *Baure Documentation*. Nijmegen: DOBES-MPI. http://www.mpi.nl/dobes/projects/baure/.

Danielsen, Swintha. 2007. *Baure: an Arawak language of Bolivia*. Indigenous Language of Latin America 6. Leiden: CNWS Publications.

Davies, Mark, and Michael Ferreira. 2006. *Corpus do Português (45 million words, 1300s-1900s)*. Provo: Brigham Young University. http://www.corpusdoportugues.org.

Davies, Mark. 2002. *Corpus del Español (100 million words, 1200s-1900s)*. Provo: Brigham Young University. http://www.corpusdelespanol.org.

--- 2004. *BYU-BNC: The British National Corpus*. Provo: Brigham Young University. http://corpus.byu.edu/bnc.

--- 2008. *The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present*. Provo: Brigham Young University. http://www.americancorpus.org.

Dhillon, Rajdip. 2010. Examining the "Noun Bias": A Structural Approach. *University of Pennsylvania Working Papers in Linguistics* 16.

Dryer, Matthew S., and Martin Haspelmath (eds.) 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library. http://wals.info/.

Duponceau, Peter Stephen. 1819. Report of the corresponding secretary to the committee, of his progress in the investigation committed to him of the general character and forms of the languages of the American Indians: Read, 12th Jan. 1819. *Transactions of the Historical & Literary Committee of the American Philosophical Society, held at Philadelphia, for promoting useful knowledge* 1.xvii-xlvi.

Ernszt, Martina, Tom Güldemann, and Alena Witzlack-Makarevich. to appear. Valency classes in N|uu. *Valency Classes: A Comparative Handbook*, ed. by Bernard Comrie and Andrej L. Malchukov. Berlin, New York: de Gruyter Mouton.

Ernszt, Martina, Tom Güldemann, Sven Siegmund, and Alena Witzlack-Makarevich. in prep. The pronoun system of N|uu. Leipzig: Max Planck Institute for Evolutionary Anthropology, ms.

--- 2011. The other "he". Reference tracking in N‖ng. Poster presented at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Ernszt, Martina, Sven Siegmund, and Alena Witzlack-Makarevich. 2008. Serial verb constructions in N|uu. *Proceedings of the 3rd International Symposium, July 6-10, 2008, Riezlern/Kleinwalsertal.* Köln: Rüdiger Köppe.

Gaenszle, Martin, Balthasar Bickel, Judith Pettigrew, Arjun Rai, Shree Kumar Rai, and Narayan P. Gautam Sharma. 2010. Binomials and the noun/verb ratio in Puma Rai ritual speech. Manuscript. Leipzig, ms.

Gentner, Dedre. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language development. Vol 2*, ed. by Stan A. Kuczaj, 38-62. Hillsdale: Erlbaum.

Gilligan, Gary M. 1987. A cross-linguistic approach to the pro-drop parameter. Los Angeles: University of Southern California, ph.d. thesis.

Givón, Talmy (ed.) 1983. *Topic continuity in discourse. A cross-language study.* Amsterdam: Benjamins.

Güldemann, Tom, Alena Witzlack-Makarevich, Martina Ernszt, and Sven Siegmund (eds.) 2010. *A text documentation of N|uu*. Leipzig, London: MPI-EVAN, ELDP. http://www.eva.mpg.de/lingua/research/nuu.php.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds.) 2005. *The world atlas of language structures : WALS*. Oxford: Oxford University Press.

Householder, F. 1981. *Syntax of Apollonius Dyscolus*. Amsterdam, Philadelphia: John Benjamins.

Huang, C.-T. James. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15.531-574.

Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen and Unwin.

Kallmeyr, Laura, Roland Meyer, and Andreas Wagner. 2001. Guidelines for the TUSNELDA Corpus Annotation Standard. Universität Tübingen. http://www.sfb441.uni-tuebingen.de/c1/tusnelda-guidelines.html.

Kampstra, Peter. 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets* 28.1-9.

Meyer, Roland, and Frank Seifart. 2011. Automatic Tagging for Lesser Studied Languages: The Case of Bora-Miraña. Universität Regensburg, MPI für Evolutionäre Anthropologie, Leipzig, ms.

Meyer, Roland. 2003. Halbautomatische morphosyntaktische Annotation russischer Texte. *Linguistische Beiträge zur Slavistik aus Deutschland und Österreich. X. JungslavistInnen-Treffen, Berlin 2001*, ed. by Robert Hammel and Ljudmila Geist, 92-105. München: Sagner.

--- 2011. Old wine in new wineskins? Tagging Old Russian via annotation projection from modern translations. *Russian Linguistics* 35.267-281.

Miceli, G., M. C. Silveri, G. Villa, and A. Caramazza. 1984. On the basis for the agrammatic's difficulty in producing main verbs. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 20.207-220.

Nelson, Gerald (ed.) 2011. *The International Corpus of English (ICE)*. Hong Kong: The Chinese University of Hong Kong. http://ice-corpora.net/ice/.

Newmeyer, Frederick J. 2004. Against a parameter-setting approach to typological variation. *Linguistic Variation Yearbook* 4.181-234.

--- 2006. A rejoinder to "On the role of parameters in Universal Grammar: a reply to Newmeyer" by Ian Roberts and Anders Holmberg'. Revised version, Febraruy 2006. Manuscript. Seattle, ms.

Nordhoff, Sebastian. 2004. *Nomen/Verb-Distinktion im Guarani*. Arbeitspapiere des Instituts für Sprachwissenschaft der Universität zu Köln 48 (Neue Folge). Köln: Institut für Sprachwissenschaft der Universität zu Köln.

--- 2009. *A grammar of upcountry Sri Lanka Malay*. LOT Dissertation Series 226. Utrecht: LOT.

Pakendorf, Brigitte (ed.) 2007. *Documentation of Sakha (Yakut)*. Leipzig: MPI-EVA.

Pakendorf, Brigitte, Dejan Matić, Natalia Aralova, and Alexandra Lavrillier (eds.) 2010. *Documentation of the dialectal and cultural diversity among Ėvens in Siberia*. Nijmegen, Leipzig: DOBES, MPIP, MPI-EVA.

Pederson, Eric, Eve Danziger, David P. Wilkins, Stephen C. Levinson, Sotaro Kita, and Gunter Senft. 1998. Semantic Typology and Spatial Conceptualization. *Language* 74.557-589.

Roberts, Ian, and Anders Holmberg. 2006. On the role of parameters in Universal Grammar: a reply to Newmeyer. *Organizing grammar: Linguistic Studies in honor of Henk van Riemsdijk*, ed. by Hans Broekhuis, Norbert Corver, Riny Huybregts, Ursula Kleinhenz, and Jan Koster, 538-553. Berlin, New York: Mouton de Gruyter.

--- 2009. Introduction: parameters in minimalist theory. *Parametric Variation: Null Subjects in Minimalist Theory*, ed. by Theresa Biberauer, Anders Holmberg, Ian Roberts, and Michelle Sheehan, 1-57. Cambridge: Cambridge University Press.

Schneider, Stefan. 2003. *BAnca Dati dell'Italiano Parlato*. Graz: Karl-Franzens-Universität Graz.

Schnitzer, Marc L. 1989. *The pragmatic basis of aphasia: a neurolinguistic study of morphosyntax among bilinguals*. Hillsdale: Lawrence Erlbaum Associates.

Seifart, Frank, Roland Meyer, Taras Zakharko, Balthasar Bickel, Swintha Danielsen, Sebastian Nordhoff, and Alena Witzlack-Makarevich. 2010. Cross-linguistic variation in the noun-to-verb ratio: Exploring automatic tagging and quantitative corpus analysis. Paper presented at the DobeS Workshop "Advances in Documentary Linguistics" Nijmegen, 14-15 October 2010.

Seifart, Frank, Brigitte Pakendorf, and Patrick Steinkrüger. in prep. The productivity of borrowed affixes: A comparative study on Resígaro (Arawakan, Peru), Sakha (Turkic, Siberia) and Chabacano (Creole, Philippines). Max Planck Institute for Evolutionary Anthropology, Leipzig, Manuscript, ms.

Seifart, Frank. 2009. Bora documentation. *A multimedia documentation of the languages of the People of the Center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions*, ed. by Frank Seifart, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri. Nijmegen: DOBES-MPI. http://corpus1.mpi.nl/qfs1/media-archive/dobes_data/Center/Info/WelcomeToCenterPeople.html.

--- 2011. Cross-linguistic variation in the noun-to-verb ratio:the role of verb morphology and narrative strategies. Poster presented at the Association for Linguistic Typology 9th Biennial Conference, The University of Hong Kong, July 21-24, 2011.

Stoll, Sabine, and Balthasar Bickel. 2009. How deep are differences in referential density? *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, ed. by Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Şeyda Özçalişkan, 543 – 555. New York, Hove: Psychology Press.

Stoll, Sabine, Balthasar Bickel, Elena Lieven, Goma Banjade, Toya N. Bhatta, Martin Gaenszle, Netra P. Paudyal, et al. 2010. Nouns and verbs in Chintang: children's usage and surrounding adult speech. Manuscript. Max Planck Institute for Evolutionary Anthropology, Leipzig, ms.

Taraldsen, Tarald. 1978. *On the NIC, Vacuous Application, and the That-trace Filter*. Bloomington: Indiana University Linguistics Club.

Thompson, Cynthia K., Stephen Fix, and Darren Gitelman. 2002. Selective impairment of morphosyntactic production in a neurological patient. *Journal of Neurolinguistics* 15.189-207.

Trilsbeek, Paul, Sebastian Drude, Bruna Franchetto, Jürg Gasché, Lucía Golluscio, Carlos Leinho, Victor Miyakawa, Frank Seifart, Daan Boeder, and Peter Wittenburg. 2007. Repositories and Archives as Nodes in a Grid. Paper given at the DELAMAN meeting, Mexico, November 2007.

Tuldava, Juhan. 2005. Stylistics, author identification. *Quantitative Linguistik: ein internationales Handbuch*, ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund Genrikhovich Piotrowski, 368-387. Berlin, New York: Mouton de Gruyter.

Van Valin, Robert D., and Randy J. LaPolla. 1997. *Syntax: structure, meaning and function*. Cambridge: Cambridge University Press.

Witzlack-Makarevich, Alena. 2006. Aspects of information structure in Richtersveld Nama. Leipzig: University of Leipzig m.a. thesis.

--- 2011. Typological Variation in Grammatical Relations. Leipzig: University of Leipzig doctoral dissertation.