Exploring probabilistic differences between genetically related languages*

Daniel Wiechmann

RWTH Aachen University (Germany)

The study proposes an analytical pipeline for quantitative corpus-based research in the domain of contrastive linguistic analysis. Specifically, it is proposed that a combination of conditional inference tree models for data exploration combined with subsequent statistical modeling via generalized linear models can help reveal and describe interesting contrasts between the languages under investigation. The main goal of this paper is to provide an approachable introduction into the rationale underlying these techniques and to illustrate their application on the basis of a contrastive analysis of English and German relative clause constructions.

Keywords: Contrastive Analysis, non-categorical contrasts, exploratory data analysis, Quantitative Corpus Linguistics, English/German

1.   Introduction

This study proposes an analytical pipeline for an exploratory quantitative contrastive analysis of English and German relative constructions. I take it that this type of analysis is comparatively uncommon in the field of contrastive linguistics, so let me start by clarifying what exactly this approach entails. The first thing to note is that the study focuses on quantitative differences, rather than qualitative ones. Traditional contrastive linguistic studies are characterized by the careful linguistic analysis of the expressive and combinatorial possibilities of the languages being compared. By way of assessing what is and what is not possible, it can make visible certain lines that mark important categorical contrasts and commonalities between the languages. In contrast to this approach, the present analysis focuses on the commonalities of the compared systems and aims at disclosing statistically significant quantitative differences within an area of grammar which exhibits shared properties. As a consequence thereof, the type of analysis envisaged here is
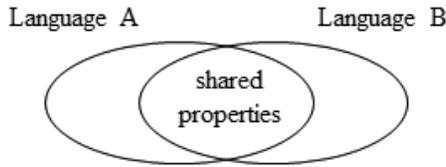
Figure 1. Venn diagram representing shared and unshared properties of two compared languages.

most useful when we wish to compare genetically related languages, as these are likely to show relatively larger areas of grammatical overlap. Figure 1 presents an illustration of the domain under investigation.

The ellipses represent sets of properties that can be ascribed to the two languages under comparison, Language A and Language B. A traditional contrastive analysis will tell us whether given linguistic property is shared by the languages or not. The intersection area in Figure 1 thus denotes the set of shared properties and it is this area, which will be looked at more closely here. To give an example: a comparison of German and English adpositions will reveal that both languages have prepositions as well as postpositions, and this is as far as one can get with qualitative statements. A quantitative contrastive analysis, on the other hand, can deepen our understanding of the two systems in question by complementing the traditional analysis with an assessment of the degree of (dis-)similarity between the two languages, e.g. by providing information on how much the languages differ with respect to their relative preference towards one type of adposition over the other.

By "exploratory" I mean that rather than testing a specific set of hypotheses, the research presented here exemplifies a scenario where there is no relevant theory in place about such preferences, which could guide the research. Typically, the goal of an exploratory study is to document the object of interest as completely as possible. In such contexts it is common to start by taking a holistic look at the objects in question, which means that one starts by gathering as much information about the objects as possible, and postpones the task of cutting away unnecessary data until we get a better picture about what is most important. For the present study, this procedure translates into (a) describing two comparable samples of relative constructions (one per language) with respect to a large number of dimensions of contrast (i.e. linguistic properties) and (b) subsequent weeding out of irrelevant dimensions of contrast. The so identified most discriminatory properties and their interrelations can then fuel future research, as they constitute empirical facts in need of an explanation. So, crucially, rather than trying to find an answer to a known problem, the goal is to identify new problems and generate new hypotheses.

Corpora of natural language can be considered the key resource for such approaches as they enable linguists to estimate statistical properties of grammars from naturalistic (i.e. non-artificial) contexts of language use. Of course, the usefulness of corpora to the domain of contrastive linguistics has long been acknowledged by many researchers and the number of corpus-based studies in the field has increased noticeably over the last ten years (for overviews and case studies cf., e.g., Filipovic, 1974; Baker, 1993; Johansson and Hofland, 1994; Granger, 1994; Hartmann, 1996; Johansson, 1998; Granger *et al.*, 2004). However, the approach taken here, namely to identify quantitative differences within a complex domain of grammar on the basis of a large set of descriptor variables, has not yet been exhausted in contrastive investigations (but cf., e.g., Stoll and Bickel, 2009 for a related approach in quantitative typology and Szmrecsanyi, 2011 for a related approach in dialectometry).

The goal of the study presented here is to make some steps in that direction and to propose a pathway of an exploratory quantitative contrastive analysis. To this end, it offers a first look at a set of shared linguistic properties of English and German relative constructions (RCs) and sets out to (i) identify those variables that discriminate most strongly between the two languages, and (ii) to assess the magnitude of the quantitative contrasts they introduce. In pursuit of this goal, we will employ two statistical tools, which will be used in a complementary fashion: first, the identification of interesting dimensions of contrast will be carried out by way of a tree-based recursive partitioning technique; second, the eventual measurement of the magnitude of contrasts and the assessment of interrelations among discriminating variables will be handled by way of binomial multivariable logistic regression modeling.

The remainder of this section will provide some linguistic background for the comparison of RCs and — where necessary — familiarize the reader with the variables used for their description. As relative constructions constitute a rather complex phenomenon, we will introduce a relatively large set of variables that serve as dimensions of (potential) contrast. Section 2 will present the corpus data and Section 3 is dedicated to describing the proposed methodology. Given the methodological focus of the paper, the method will be described in considerable detail (without becoming unnecessarily technical). Complex phenomena tend to require complex descriptions and complex linguistic structures — such as RCs — can vary along many grammatical and semantic dimensions. More often than not, it is hard to determine *a priori* which dimensions give rise to interesting contrasts and which do not.

The first step in the analysis is thus of an exploratory nature and involves a rich (ideally comprehensive) description of (ideally representative) samples of the compared structures. Having described the data, the goal is then to identify those

variables (i.e. dimensions of contrast) that are most distinctive for the languages in question (Sections 3.1 and 3.2). In the attempt to solve this task, we will propose an easy-to-use and statistically sound procedure, namely a type of recursive partitioning that employs large ensembles (i.e. forests) of so called *conditional inference trees*. Once the most distinctive variables have been identified, we will measure the magnitude of the probabilistic contrasts they introduce and investigate their interrelationships (Section 3.3). To this end we will apply a logistic regression model to the data. Such models can help us understand the systematic probabilistic differences underlying the languages investigated, as they can be set up to predict the class-membership of a RC-construction (English or German) on the basis of the distributions of the linguistic variables used in their description.

In summary, the analysis proceeds in three steps:

1. Sample from comparable sources and code for a broad range of variables to describe the data
2. Lose unnecessary complexity of the data and identify the variables that most strongly discriminate between the languages (Tool: Random forest of conditional inference trees)
3. Measure the magnitude of the contrasts and investigate interrelationships between variables (Tool: binomial logistic regression)

Finally, Section 4 will provide some discussion of the findings and Section 5 will conclude the study.

1.1   Bi-clausal relative clauses constructions: some preliminaries

Any contrastive analysis must begin with the explication of the thing to be compared. In cross-linguistic contexts this is particularly challenging as both notional and formal definitions tend to be problematic when applied to different linguistic systems and relative clauses are no exception. For the purposes of this paper, we will adopt a working definition (of what qualifies as an RC) that has good chances of being accepted by a large number of linguists.

> Definition: Relative Clause
> A relative clause is a type of subordinate clause that serves an attributive function and thus typically modifies a nominal element.

To maximize the comparability of the structures across the two languages, we will constrain ourselves to considering only those relative clauses that are tensed and whose dominating clause is the main clause of the respective sentence. The variables that will be used in the description of the resulting (bi-clausal) structures pertain to a set of properties that can be binned into three groups.

I.   Properties of the relative clause, namely
   a.   the internal grammatical role of the head (i.e. type of RC)
   b.   the external grammatical role of the head (i.e. role of modified nominal)
   c.   the type of embedding of the RC
   d.   the voice of the RC
II. Properties encoded on the head noun
   a.   the animacy (of the referent) of the head
   b.   the definiteness of the head
   c.   the morphosyntactic realization of the head
III. Properties encoded on the subject of the relative clause (i.e. SRC)
   a.   the animacy (of the referent) of the RC subject
   b.   the definiteness of the RC subject
   c.   the morphosyntactic realization of the RC subject

Let us illustrate this on the basis of examples 1 and 2, which represent two pairs of semantically equivalent relative clause constructions.

(1)   a.   The man [who got shot] was Skip Martin.
       b.   *Der Mann* [, *der erschossen wurde,*] *war Skip Martin.*

(2)   a.   Peter hates anything [that his brother likes _].
       b.   *Peter hasst alles* [, *was sein Bruder mag _*].

The pairs of examples 1 and 2 illustrate some of the variability of relative clause constructions. They constitute two quite dissimilar relative clause patterns, where (dis-)similarity is measured over the values of the variables listed in I–III. With respect to group I, i.e. the variables associated with the RC proper, we may say that examples 1a and 1b both instantiate center-embedded passive subject RCs that modify the subject of the respective main clause. In contrast, examples 2a and 2b instantiate right-embedded active object RCs that modify the direct object of the respective main clause. Moving to the second group of properties, we observe that the heads of the RCs in examples 1a and 1b are lexical nouns that denote animate entities and that receive the value +DEF(INITE) from the articles that quantify the entire referring expressions. Again, the values for the relevant variables are orthogonally different in examples 2a and 2b, where the respective heads are realized as indefinite pronominal expressions referring to inanimate entities. Finally, in regard to group III, we may note that the subject NPs of the RCs in all examples refer to animate entities. However, in examples 2a and 2b the grammatical subjects of the RC are expressed though definite lexical NPs, whereas examples 1a and 1b constitute relativizations on the subject role and thus inherit their subject$_{RC}$ properties from the head noun. With these prerequisites in place we may now turn to the description of our data.

2. Corpus data

The data used in the comparative analysis of English and German relative clause constructions consist of a total of 500 corpus extractions, viz. 250 observations from each language. The English data were extracted from the written part of the British component of the International Corpus of English (ICE-GB R2, Nelson *et al.*, 1996). The German data were extracted from the IDS corpus compilation via COSMAS II$_{web}$. In order to control for potential differences introduced by genre-specific usage preferences, the ICE-GB data were matched to the IDS corpus extractions in terms of text types and comprise rather formal written texts only (newspaper texts and academic writing).

The corpus query was designed for maximal generality, so as to ensure that no true hits were missed. The rich grammatical annotation of the ICE corpus allowed a direct query for finite RCs. The German corpora are not parsed but tagged for part-of-speech information. Correspondingly, the query targeted structures that include relative pronouns. The search was stopped at n = 1000. From these data, random subsamples were extracted with n = 250 for each language. Each instance was manually checked. RCs that exhibit attachment ambiguities (e.g. […] we have the latest version (N1) of the car (N2) that […]) were excluded as they render the identification of the head problematic. More generally, if a data point was judged to be problematic (e.g. incomplete or grammatically anomalous in any way), it was removed and substituted with another random pick until both samples consisted of only true hits of the target patterns.

Table 1. Variables used to describe the corpus data (overview).

No	Variable	Scale of measurement				Labeling of levels
1	LANGUAGE	Factor	w/	2	levels	English, German
2	INTERN.ROLE	Factor	w/	3	levels	subj, do, other
3	EXTERN.ROLE	Factor	w/	3	levels	ext.subj, ext.do, ext.other
4	EMBEDDING	Factor	w/	2	levels	center, right
5	VOICE.RC	Factor	w/	2	levels	act, pass
6	ANI.HEAD	Factor	w/	2	levels	h.ani, h.ina
7	DEF.HEAD	Factor	w/	2	levels	h.def, h.indef
8	MS.HEAD	Factor	w/	2	levels	h.lex, h.prn
9	ANI.SRC	Factor	w/	2	levels	src.ani, src.ina
10	DEF.SRC	Factor	w/	2	levels	src.def, src.indef
11	MS.SRC	Factor	w/	2	levels	src.lex, src.prn

Once extracted and checked, the corpus data were manually annotated with information regarding the variables introduced in the preceding section. Table 1 presents the variables investigated (in small caps), followed by a short description of their respective types. The rightmost column indicates the labeling of the respective factor levels for a given variable.[1]

The resulting data set is considered to be not atypical for corpus-based studies. More often than not such studies analyze 'unbalanced' data sets, in which the logically possible combinations of factor levels are not equally represented and which typically exhibit many empty cells, i.e. unattested factor level combinations.

3.   Methods and results

Our data set comprises information of ten variables that could potentially distinguish English RCs from German ones. The goal now is the uncovering of structural properties of the data set that constitute substantial probabilistic differences between English and German relative constructions.[2] To attain this goal, we will eventually apply a multivariable logistic regression model to the data that considers a set of variables and all their potential interactions simultaneously.[3] At this point, however, this is not recommendable as the sheer number of variables prevents any serious attempt to apply a complex logistic regression model on the basis of the available corpus data. Instead, we will employ an iterative statistical procedure to identify those variables that are interesting in the sense that they somehow figure in important probabilistic differences between the two languages. This is to say that, rather than applying the regression model right away, we will employ a recursive partitioning procedure to identify the most important variables (thereby reducing the complexity of the data without losing too much valuable information). The so identified variables will subsequently be used in the multivariable regression model.

Logistic regression models are fairly common in many areas of linguistic inquiry and postponing their application here may require some elaboration. The strength of a multivariable regression model is that it allows us to assess the effects of many variables and their interactions in concert. However, applying a regression model that includes all ten main effects and at least all two-way interactions on the basis of the available limited and unbalanced data is problematic. Doing so would result in a "small n [for cases] to large p[redictors]"-scenario, which "may well lead to cell counts too sparse for parameter convergence" (cf., e.g., Peduzzi *et al.*, 1996; Strobl *et al.*, 2009). If, on the other hand, we excluded the interaction terms and included only main effects, we would lose much of the original motivation for employing multivariable models in the first place, namely the identification of

potential interaction effects. In addition to these complexity problems, we should also decide against applying a regression model whenever some of the variables correlate with each other, as this would lead to multicollinearity problems.[4]

In consequence, we would be forced to build and compare many regression models to identify all interesting variables and variable interrelationships, making this approach rather tedious and unpractical. One way of avoiding such trouble, which is proposed here, is to turn to tree models, a class of statistical classifiers that proceed in an iterative fashion, i.e. they evaluate one variable at a time. Tree models produce accurate and interpretable models with relatively little user intervention (cf. Breiman, 2001). The variables identified by that procedure will then be passed on to logistic regression modeling. The regression model then serves a double-function: (1) it allows us to ascertain the results obtained through the recursive partitioning via tree models and (2) it provides us with complementary perspectives on our data, thereby helping us better understand the contrasts.

In the light of these considerations, this section has the following structure. We will first (briefly) investigate our data using a series of bivariate analyses (Section 3.1). This will already reveal certain variables that mark interesting quantitative contrasts between English and German relative constructions. In Section 3.2 we will employ the above-mentioned recursive portioning technique to identify the most important variables. Specifically, we will introduce large ensembles (forests) of conditional inference trees.[5] Having identified the most important variables, we will then pass them on to a multivariable logistic regression model that will use that information to estimate the type and magnitude of probabilistic differences between English and German RCs (Section 3.3).

3.1   Bivariate perspectives

We can help ourselves to a first idea by simply plotting the frequency distributions of each variable for both languages, as shown in Figure 2.

Nonparallel lines in a given plot suggest that there is a quantitative difference between English and German RCs for that variable. For example, Figure 2 strongly suggests that the factor VOICE does not discriminate between English and German relative constructions: the lines describing the frequency distributions of active and passive constructions overlap almost perfectly. In contrast, the factor EXTERN. ROLE makes for a good candidate for a discriminatory variable. For many variables though, e.g. for the variable DEF.HEAD, it seems difficult to decide on the issue by just eye-balling these data. At this point, we can, of course, statistically analyze all bivariate relationships and test for statistical associations between each of the variables and LANGUAGE. In other words, we can analyze in a one-by-one fashion all 10 cross-tabulations like the one given in Figure 3, which visualizes the contingency
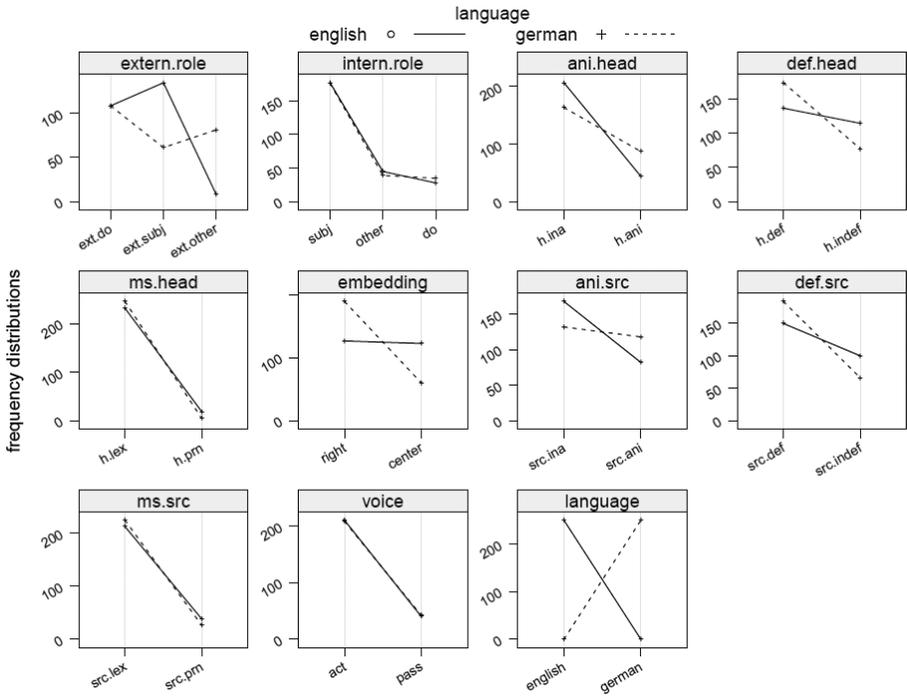
Figure 2. Distributional differences between English and German RCs.

table for EXTERN.ROLE versus LANGUAGE as a balloonplot. Such plots represent the frequencies in the table in terms of circles of proportional size, thereby allowing for a quick and convenient first interpretation of the distribution.

Using such contingency tables as input, we can compute the association strengths between each variable in our description and LANGUAGE, and identify those variables that exhibit the most pronounced values. Figure 4 presents the results of such a procedure, where association strength was measured using Cramer's
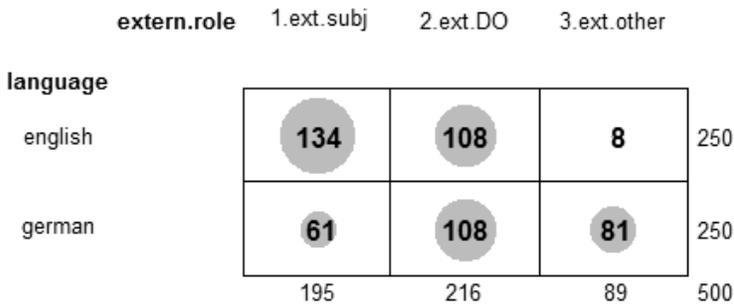

Figure 3. Balloon plot of EXTERN.ROLE VS. LANGUAGE (diameter is proportional to frequency).
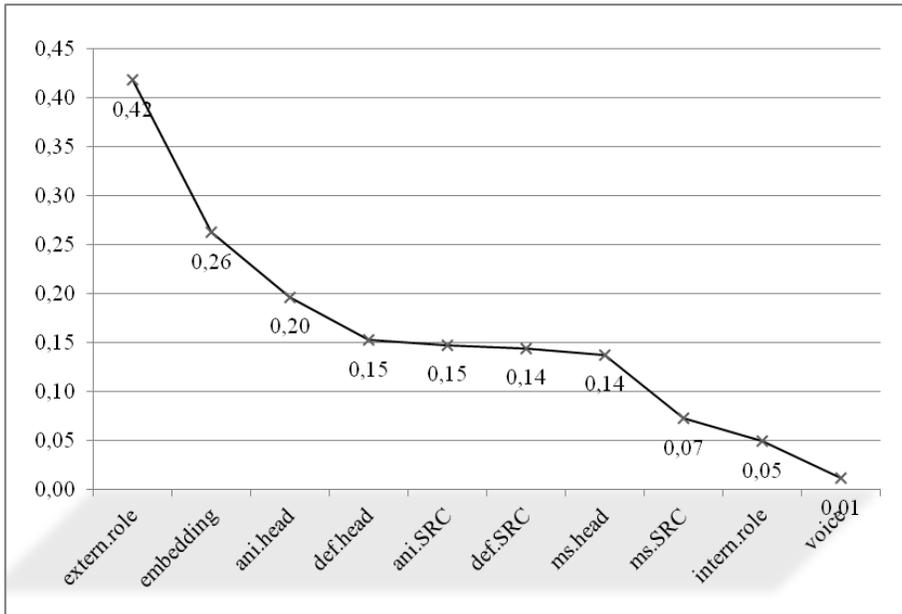
Figure 4. Association strength (Cramer's V): LANGUAGE versus all descriptors.

V.[6] We observe that the variable EXTERN.ROLE exhibits the highest score (Cramer's V = 0.42), meaning that this variable discriminates most strongly between English and German relative constructions. Conversely, VOICE has close to no discriminatory effect (Cramer's V = 0.01).

The advantage of this bivariate analysis clearly is that we do not run into the data sparsity problems we have alluded to before: there are sufficient data points for each bivariate test. However, it has the disadvantage of not being able to show us any interrelationships between the predictors.

3.2 Exploring data sets with Tree(-Based) Models

Having investigated all cross-tabulations with LANGUAGE, this section introduces a recursive partitioning technique, which belongs to a class of procedures that exploit the strength of a serial bivariate approach while remedying some of its disadvantages, most notably the inability to detect interactions between variables. Quite generally, whenever the goal is data exploration, tree models constitute a valuable resource in the analyst's tool kit as they do not suffer from the problems of more widely used linear regression models, i.e. problems associated with empty cells and extreme differences in cell counts, which we have alluded to earlier. The method chosen here employs so called Conditional Inference Trees (ctree algorithm, Hothorn *et al.*, 2006). The procedure involves a recursive partitioning

algorithm that seeks to find the strongest discriminators. The ctree-algorithm works as follows:

1. Test the global null hypothesis of independence between any of the input variables and the response (here: LANGUAGE). Stop if this hypothesis cannot be rejected. Otherwise select the input variable with the strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response.
2. Implement a binary split in the selected input variable.
3. Recursively repeat steps 1 and 2.

Before we look at the results of the recursive partitioning technique, a last precaution is in order: a single tree can sometimes produce problematic results. Most importantly, it can produce results that may adequately describe structural properties of the data at hand, but that are not robust enough to generalize to other data sets (a few cases may change the decision as to whether or not a variable is included into the tree). One way to remedy this is to build not just one tree but many trees and focus on the patterns they "agree on". This is what was done in the present study.[7] Specifically, a total set of 1000 trees was grown using a bootstrapping technique, in which 1000 different random subsamples were taken from the original data. The resulting estimated importance of the 10 variables is given in Figure 5 (cf. Strobl, 2009 for details).

Figure 5 suggests that the variables EXTERN.ROLE, EMBEDDING, VOICE and ANI. HEAD are important in the description of probabilistic differences between English and German RCs (all remaining variables have values very close to 0). The variable VOICE was the least discriminatory of all variables in Figure 4, so it is certainly surprising to see it pop up here. This is a hint towards it figuring in an interaction. To better understand these results, we can visualize the tree-based modeling on the basis of the results from the "best" tree, i.e. the tree that was fed with all 500 observations of English and German relative clauses. The resulting classification tree in Figure 6 is best interpreted from top to bottom.[8]
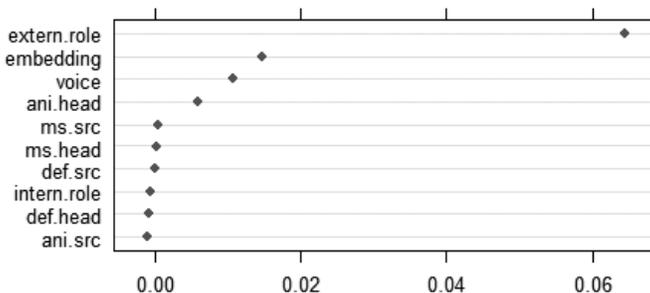

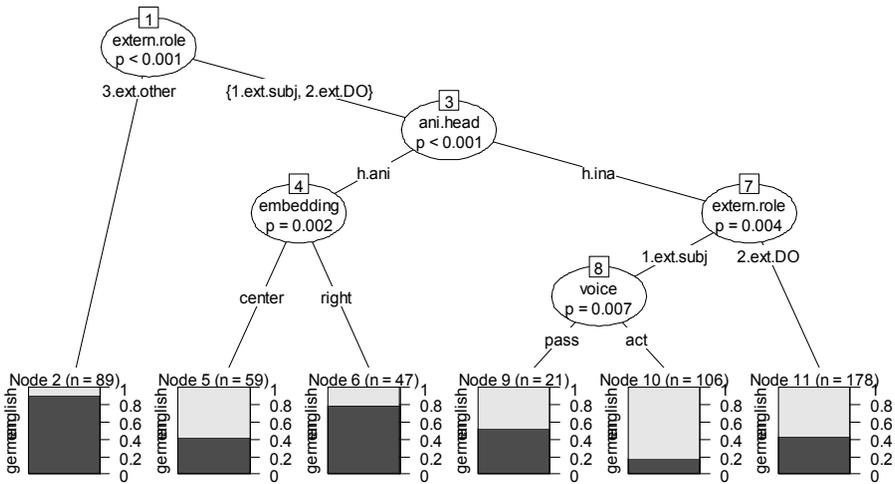**Figure 5.** Variable importance for random forest with all predictors.

Figure 6. Results of recursive partitioning: Conditional inference tree.

We observe that the variable EXTERN.ROLE is judged to be the most distinctive one, which corresponds to the bivariate results we presented in Figure 4.[9] This variable is used to introduce a first split. The two resulting groups differ significantly with respect to the response, i.e. they differ in terms of the proportions of English and German instances. The proportion of German instances is a lot greater if the external role is low on the accessibility hierarchy (i.e. "ext.other") than it is for the remaining factor levels, i.e. "ext.subj" and "ext.DO". The (spine)plot in Node 2 in the left-hand corner tells us (a) that there are 89 instances of this type and also (b) that the lion's share (81/89) of these constructions is from the German part of the data. The proportion of German instances is significantly lower for the rest of the groups in the tree.

Moving to the right-hand side of the tree, i.e. Node 3, we observe that the remaining (500 − 89 =) 411 cases, whose value for EXTERN.ROLE is either "ext.DO" or "ext.subj", are best split on the basis of the variable ANI.HEAD, resulting in Nodes 4 and 7. The tree structure then unfolds according to the logic of the ctree-algorithm and terminates when no subset of the data can be further split. The tree model suggests that there are interactions between variables: in classification trees, interactions between two variables k and m are suggested, if in one branch created by some predictor m it is not necessary to a split in k, while in the other branch created by m it is necessary. Applied to the tree in Figure 6, this means, for example, that there may be an interaction between EXTERN.ROLE and VOICE. In other words, the effect of VOICE seems to depend on the value for EXTERN.ROLE.

In summary, the results confirm some of the findings of the bivariate analysis (EXTERN.ROLE, EMBEDDING and ANI.HEAD are considered important), but they go

Table 2. Most distinctive variables in the data set (plus response: LANGUAGE).

No	Variable	Scale of measurement				Labeling
3	EXTERN.ROLE	Factor	w/	3	levels	ext.subj, ext.obj, ext.other
4	EMBEDDING	Factor	w/	2	levels	center, right
5	VOICE	Factor	w/	2	levels	act, pass
6	ANI.HEAD	Factor	w/	2	levels	h.ani, h.ina

beyond that result, insofar as they suggest that the variable VOICE should also be taken into account, when describing probabilistic differences between English and German relative constructions. Consequently, we continue our analysis focusing on the variables given in Table 2.

By removing 6/10 irrelevant variables, we have substantially reduced the complexity of our data.

3.3 Using binomial logistic regression to model quantitative differences

We now have good reason to believe that the variables in Table 2 (i.e. external role of the head, animacy of the head, type of embedding and the voice of the RC) are good discriminators of English and German relative constructions. We will now employ a statistical model that — in contrast to the tree-based model — considers all these variables simultaneously. At this point, however, we must exclude one factor from the model due to multicollinearity problems, which we have mentioned earlier. The variables EMBEDDING and EXTERN.ROLE correlate highly, as modification of the subject of the dominating clause almost always leads to center embedding. In consequence, we choose one of the two factors to be included in our model and drop the other. In this case, EXTERN.ROLE was retained and EMBEDDING was dropped.[10] This leaves us with three variables that will be used in the multivariable model (cf. Table 3).

For our final assessment of the probabilistic difference between English and German relative constructions, we will apply a generalized linear model, specifically a binomial logistic regression model, to the data. The goal is to model the probability p of a binary outcome, namely whether a RC type is English or

Table 3. Variables used in regression model.

No	Variable	Scale of measurement				Labeling
3	EXTERN.ROLE	Factor	w/	3	levels	ext.subj, ext.obj, ext.other
5	VOICE	Factor	w/	2	levels	act, pass
6	ANI.HEAD	Factor	w/	2	levels	h.ani, h.ina

German, as a function of the above-mentioned set of explanatory variables. The general form of the model is represented in (3).

(3) logit $p = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_k x_k$

In (3) "logit p" is the logit-transformed probability of some event,[11] the symbols $x_1, \dots x_k$ denote the explanatory variables (or *predictors*), and the expressions $\beta_1, \dots, \beta_k$ are the regression coefficients for each variable, which indicate how a given factor influences the probability of the event we wish to model (i.e. the β-coefficients express the "relative weights" of our factors). In our example, we want to model the probability that a RC type is English as a function of the factors animacy of the head, external grammatical role of the head, voice of the RC and all their potential interactions. Schematically, this is given in (4).

(4) logit Probability(RC-type is English)
     = *intercept* + weighted ANI.HEAD + weighted EXTERN.ROLE + weighted VOICE
     + all interactions between these factors

The β-coefficients (the weights of the factors) are estimated by the method of maximum likelihood, which is a generalization of the method of least squares that is used in ordinary linear regression. The goal is to find the "best" values for each β-coefficient, i.e. the values that make the observed data most likely. Table 4 presents the input data.

Table 4. Input data for logistic regression (ordered by frequency of configuration).

extern.role	ani.head	voice	Freq E	Freq G	Total
do	h.ina	act	76	66	142
subj	h.ina	act	87	19	106
subj	h.ani	act	36	30	66
other	h.ina	act	5	42	47
do	h.ani	act	6	30	36
do	h.ina	pass	25	11	36
other	h.ani	act	0	21	21
subj	h.ina	pass	10	11	21
other	h.ina	pass	3	14	17
other	h.ani	pass	0	4	4
do	h.ani	pass	1	1	2
subj	h.ani	pass	1	1	2
			250	250	500

Table 4 lists the twelve logically possible RC-types and their respective frequencies in our data. We observe that RCs that modify the direct object of the MC and that describe an inanimate entity and that are in the active voice occurred 142 times in the data (66 examples are from German, 76 from English). The next frequent pattern is similar and only differs from the first in that it modifies the subject, rather than the direct object. It occurred 106 times in our sample (19 German, 87 English), and so on and so forth. The maximum likelihood procedure estimates the weights of the investigated factors on the basis of these frequency data. We can see that there are still some cells with small counts in the table, but these need not worry us too much. Our data complexity reduction has certainly done a lot for us. We have dropped 7/10 variables and have good reason to believe that we have not lost much interesting information in doing so.

3.3.1   *Applying the (binomial) logistic regression model*

The next step in our analysis is to apply a binomial logistic regression model that considers the effects of all variables simultaneously. The goal here is to model the probability that a RC-type is English, as a function of the variables

–   external role   (EXTERN.HEIGHT)
–   voice of the RC   (VOICE)
–   animacy of the head   (ANI.HEAD)
–   and all interactions among those factors

The goal in statistical modeling typically is to find an elegant model, i.e. a model that constitutes a good compromise between descriptive adequacy and internal complexity (cf. Crawley, 2007). There are numerous criteria that affect the complexity of the model, but we can illustrate the guiding idea on the basis of a very intuitive criterion, namely the number of explanatory variables in the model. Following the principle of parsimony, a model with k factors is preferable over a model with $k+1$ factors, so long as the inclusion of the additional factor does not yield a statistically significant increase in the predictive power/fit of the model. More generally, every increase in model complexity has to be justified by a noteworthy, i.e. statistically significant, explanatory gain. Consequently, all terms that do not significantly increase the power of the model should be dropped.

Our analysis begins with the application of a maximal model, i.e. a model containing all three main effects of our predictor variables and all possible 2-way and 3-way interactions between them. We then delete all insignificant terms in a stepwise fashion, until we arrive at a model that contains only significant terms. This final, most parsimonious model is called the minimal adequate model. In our example, the minimal adequate model consists of all main effects and the 2-way interaction between the factors EXTERN.ROLE and VOICE. Recall that this interaction

Table 5. Analysis of Deviance table.

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			499	693.15	
voice	1	0.06	498	693.09	0.81
ani.head	1	20.33	497	672.76	0.00 ***
extern.role	2	108.31	495	564.46	0.00 ***
voice:extern.role	2	15.43	493	549.03	0.00 ***

was also suggested by the tree-based model, but we had not tested if that tendency was statistically significant. Table 5 presents the results of an analysis of deviance, which shows whether or not a given factor improves the model in a way that is statistically significant.[12]

The overall statistics of the minimal adequate model are acceptable. The model is clearly significant (Log-likelihood $X^2 = 14.12$, d.f. $= 6$, $p = < 2.22e-16$) and the overall power is not bad (C $= 0.78$, Nagelkerke's $R^2 = 0.34$).[13] For our present purposes, the overall model statistics are not too important though, as our goal is not perfect prediction accuracy. (If we wanted that, we would have used predictors that introduce categorical contrasts). On the other hand, what we are interested in are the estimates of the magnitude of the quantitative contrast associated with the terms in our model. Table 6 shows these estimates of the regression coefficients, their standard errors, and tests for whether each regression can be assumed to be zero (in which case we have no reason to believe that the factor in question discriminates between the languages).

As these estimates are expressed on a logit scale, they are not easily interpreted. For purposes of interpretation, it is convenient to convert the logit coefficients back to expressions of changes of predicted probabilities and present these graphically.[14] Figure 7 presents the final results of the present study.

Table 6. Estimated regression coefficients of minimal adequate model.

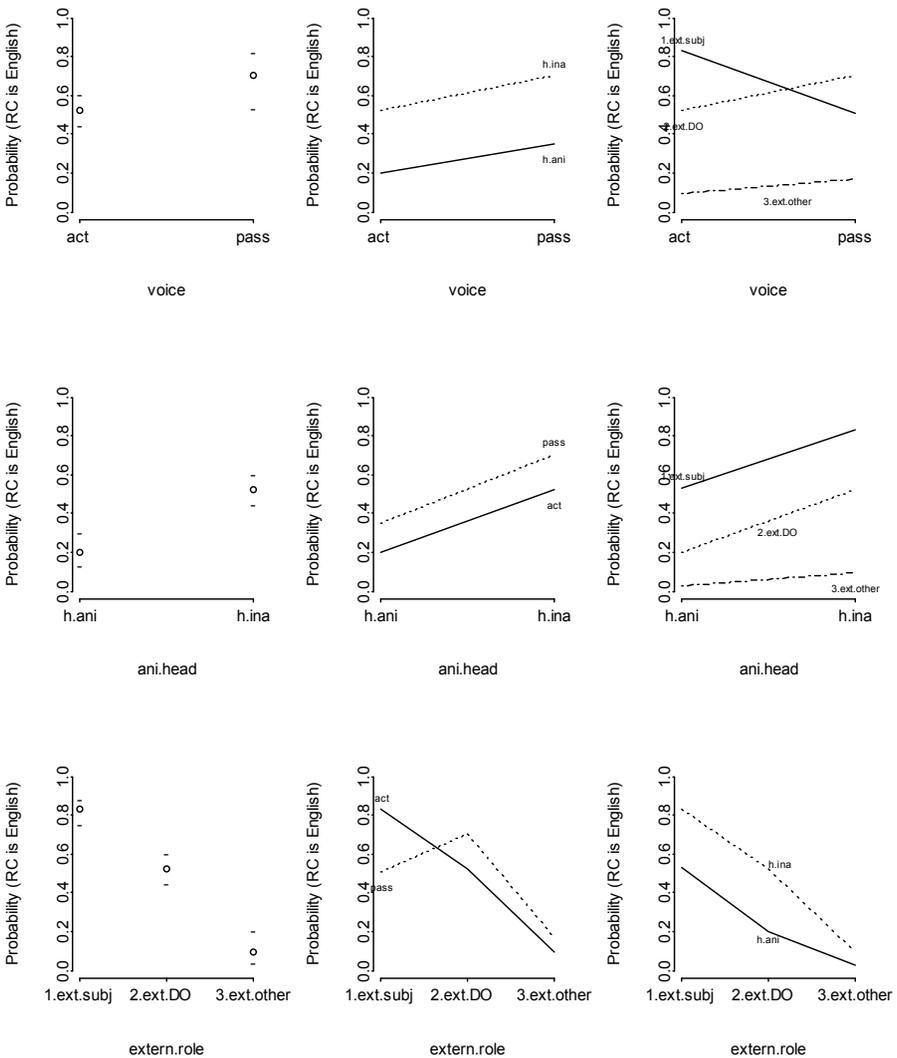|	Estimate	SE	z	Pr(>|z|)
(Intercept)	−0.12	0.22	−0.54	0.59
voice=pass	1.57	0.48	3.29	0.00 **
ani.head=h.ina	−1.47	0.27	−5.54	0.00 ***
extern.role=ext.DO	1.49	0.26	5.76	0.00 ***
extern.role=ext.other	3.84	0.51	7.47	0.00 ***
voice=pass:extern.role=ext.DO	−2.32	0.61	−3.81	0.000139 ***
voice=pass:extern.role=ext.other	−2.20	0.92	−2.399	0.016421 *

Figure 7. Regression coefficients (turned back into probabilities) for each level of each factor.

Figure 7 is best read in a row-wise fashion. Each plot presents the changes in predicted probability that an RC is English that go along with the level-changes of a given factor. The factors whose influences are not shown in a given graph are adjusted to some specific default level. This adjustment allows us to disentangle effects attributable to different variables in the model. The default levels for the adjustments are as follows: for EXTERN.ROLE the adjustment is "ext.do", for ANI.HEAD the adjustment is "inanimate" and for VOICE it is "active". The left-hand column

of plots in Figure 5 shows the effects of a given single variable when the two re-maining ones are adjusted to their reference levels. The middle and right-hand columns then add information pertaining to one of the two adjusted variables. Consequently, the full array of plots allows us to inspect all possible (two-way) interactions between the three variables.

The first plot (in the upper left-hand corner) shows the changes in predict-ed probability that an RC is English when we change the value for VOICE from ACT(ive) to PASS(ive), with the remaining variables in our model set to their ad-justed values. Note that the plot does not represent the overall probabilities (that an RC is English) for the two levels of voice. Rather, it shows us these probabilities for inanimate heads that function as main clause DO. For these cases we get to see that passive RCs are more typical of English constructions. The difference is not very pronounced, though. The bars around a given estimate indicate the respective 95% confidence intervals, i.e. the range in which the true values are likely to fall. We can see that these intervals overlap (the upper bound of "act" overlaps with the lower bound of "pass"), which means that we do not have a lot of reasons to believe that the VOICE marks an interesting contrast here.

The middle plot in the top row shows the effects of VOICE for both levels of ANI.HEAD (thus controlling for only one remaining variable). Crucially, the plot tells us that the changes in predicted probability are nearly identical across both levels of ANI.HEAD. In other words, there is no interaction between the two factors. Note, however, that we do observe an interaction between VOICE and EXTERN. ROLE in the right-most plot: the probability that an RC is English rises when we move from active to passive RCs if EXT.ROLE is either "ext.do" or "ext.other", but it drops considerably if EXT.ROLE=SUBJ. So the effect of VOICE depends on the value of EXT.ROLE. The second row shows the effect of head animacy. Here the story is more straightforward: the probability that an RC is English rises significantly when we change the level of ANI.HEAD from animate to inanimate and this effect is observed for all levels of all other factors. So English has a greater inclination to modify inanimate heads by way of a tensed RC than German has. German, too, has a preference towards modifying inanimate head over animate ones, but this preference is less pronounced.

Finally, the third row informs us about the factor EXT.ROLE. We had already seen that English has a tendency to modify the subject of the dominating clause, whereas German makes stronger use of modifications of the direct object or "low-er" grammatical roles. But again, as already mentioned above, the changes in the probability that an RC is English for the variable EXT.ROLE depend on the value of VOICE. So, while there is a greater overall tendency for English to modify more main clause subjects than objects, this tendency is significantly lower when the RC is in the passive voice.

4.   Discussion

The study disclosed a number of significant quantitative differences between English and German RCs, which are in need of explanation. While both languages prefer to employ relative clauses to modify inanimate heads, which seems plausible as inanimate entities are more likely to require extra information to fix the reference of the term, English is characterized by a significantly smaller proportion of modifications of animate heads. How can this difference be accounted for? One possibility is that English tends to employ different modification strategies for animate and inanimate entities. While this may seem improbable at first glance, it can be motivated by the fact that the relative positioning of the modifier can signal semantic contrasts in English. According to Langacker (2008: 319) "[t]here is a definite tendency in English that a modifier that directly precedes the head to specify an intrinsic or permanent property, whereas post head modifiers tend to be used for properties of a contingent or temporary character".

Now, if (p) this tendency is in fact empirically true for English, and if (q) speakers tend to ascribe different types of properties (intrinsic vs. temporary ones) to different types of objects (animate vs. inanimate objects), then we would expect there to be distributional differences of modifier types for animate and inanimate entities. And it is, furthermore, certainly conceivable that German may simply not exhibit this proposed tendency (to the same degree). Such a situation would then give rise to the probabilistic difference reported here. Building on the present findings, future research may pursue questions about differences in modification of animate head nouns between the two languages.

Secondly, concerning the external syntactic properties of the RC, the present study has demonstrated that English and German RCs differ significantly with respect to the modified grammatical role of the dominating clause: English shows a strong preference to modify subject constituents, relegating modifications of the direct object to a secondary role. Roles below that of DO on the accessibility hierarchy hardly ever receive modification by way of a RC in the English data. German, in contrast, employs RCs mainly to modify the referent of the DO, which is more expected from the perspective of information flow: direct objects should be in stronger need of modification, as they are more likely to be part of new information in the discourse (whereas subjects are usually given).

This distribution can be taken to suggest that the RCs in formal written language investigated here are not used to restrict the set of possible referents of the subject constituents, but they are rather used for purposes of adding a proposition to the sentence. In fact this has been proposed in recent investigations of English academic writing and newspaper texts. Biber and colleagues have reported that English academic writing has become increasingly more compact and compressed,

showing growing proportions of inserted dependent clauses adding extra information over the last 100 years (Biber, 2003; Biber and Gray, 2010). It appears that German does not employ RCs for this function to a comparable degree leading to a significant probabilistic difference (at least in the register domains studied here).

Finally, we have reported an interaction between the variables VOICE and EXTERN.ROLE. It was shown that English subject modification by way of a RC is much less likely when the RC is in the passive voice. This is likely due to the fact that (in formal registers) passive clausal modifiers are often realized as reduced participial constructions (cf. "the method applied here", instead of "the method that|which was applied here"). The relativized role of such RCs invariably is that of subject causing the observed interaction. So part of the reason for the observed distributions is that English can make use of alternative constructions that are unavailable in German. With these differences being revealed, future research can now engage in targeted investigations of distributional differences in the modification of subject constituents between the two languages.

5.   Conclusion

The goal of this paper was to contribute to the further development of the methodological repertoire of corpus-based contrastive linguistic research. On the basis of a (preliminary) analysis of 500 English and German relative clause constructions, the paper proposed a two-step corpus-based methodology for the investigation of non-categorical contrasts between genetically related languages. It was proposed that an exploratory data analysis via conditional inference trees combined with subsequent regression modeling can reveal and describe interesting quantitative contrasts between the languages. It was argued that applying a logistic regression model right away is problematic in cases where we wish to explore small and unbalanced data sets that include correlating variables, i.e. in cases which are considered typical for contrastive linguistic studies.

It was further shown that the iterative nature of the tree-based procedure commends itself as a first step in the analysis, as it circumvents problems of techniques that consider all variables simultaneously, like logistic regression. The tree-based model allows us not only to investigate an arbitrary number of variables, but also to feed into the algorithm variables that highly correlate with each other. The result is an easy-to-use and easy-to-interpret first step towards weeding out uninteresting variables, which is preferable to a series of simple bivariate analyses because it can also identify variables that figure in some interaction. The variables that were retained in the tree-based model were then further analyzed by way of regression modeling, so as to confirm the results of the tree-based technique and to better

understand the nature of the discriminatory effects of the investigated variables and their interactions. I hope that the methodology described in the present paper is of some utility for future work that aims at uncovering probabilistic contrasts between genetically related languages.

Notes

* I would like to thank Michael Cysouw, Arne Lohmann, Karsten Schmidtke-Bode and an anonymous referee for their helpful suggestions and comments. All remaining errors are, of course, my own.

1. The factor level EXTERN.OTHER subsumes all grammatical roles that assume positions lower than that of direct objects on Keenan and Comrie's Accessibility Hierarchy (cf. Keenan and Comrie, 1977).

2. All statistical procedures were carried out using R 2.13 (R Development Core Team, 2009).

3. The term "interaction" commonly describes a situation when the effect of one predictor variable on the response variable depends on the value of another predictor variable.

4. In cases of multicollinearity the regression estimates are prone to error and may change erratically in response to small changes in the model or the data.

5. The initial version of this paper employed only a single tree-based model. One reviewer rightly pointed out that the iterative nature of the algorithm underlying tree-based models introduces its own set of problems: it may overlook important aspects of the structure of the data just because it focuses on one factor at a time. This criticism is acknowledged and remedied by the use of many trees (cf. Section 3.2).

6. Cramer's V is a measure of effect size, i.e. a measure of the strength of the relationship between two variables in a statistical population. It yields a value between 0 and 1, with larger values indicating stronger association. The measure was chosen here because it can handle contingency tables of arbitrary complexity, whereas other effect size measures, for example odds ratios, are restricted to 2 by 2 tables.

7. The ensemble of conditional inference trees (random forest) was grown using the function cforest in the party library.

8. The tree model was built using the function ctree(party). The minimum number of observations that must exist in a node resulting from a split before a split will be performed (= min split) was set to 100, which is rather conservative for the present data. The minimum number of observations allowed in any leaf was set to 20. This is a bit lower than what is often considered the default, namely to set this value to roughly a third of the min split value (here ~ 33). The consequence of this decision is that splits will be made which may not be as robust. In this study, this is countered by the use of many trees and the re-assessment of a factor's importance via logistic regression in Section 3.3.

9. The ranking of the ctree algorithm is based on the value of adjusted p-values (but cf. Strasser and Weber, 1999 for alternatives). Here, the ctree function was set *testtype = "Bonferroni"*.

10. Model comparisons via ANOVA revealed that a model that uses EMBEDDING instead of EXTERN.ROLE performs significantly worse.

11. For introductions to generalized linear models and logistic regression with *R*, the reader is referred to Crawley, 2007; Gelman and Hill, 2007; Dalgaard, 2008; Baayen, 2008; Gries, 2009.

12. The factor VOICE is retained in the model because it figures in an interaction.

13. The C index is a rank correlation between the predicted probability of response under the applied model and the actual response. A value for C exceeding 0.80 can be taken to imply useful predictability of the model. Nagelkerke's R2 is a pseudo R2 measure (there is no true R2 value in logistic regression like there is in O(rdinary) L(east) S(quares) regression). Like C, it is used to express how well the statistical model performs. It assumes values between 0 and 1, with higher values indicating better performance.

14. We can convert freely between probabilities p and odds o for an event as $o = p/1−p$ and $p = o/1+o$. The logit y of a probability p is defined as $y = \log (p/1−p)$, where "log" represents the natural logarithm. We can transform the logits back to expressions of changes in predicted probability by applying the inverse of the logit function, which is defined by $p = \exp(y)/(1+\exp(y))$.

References

Baayen, H. 2008. *Analyzing Linguistic Data — A Practical Introduction to Statistics using R*. Cambridge: CUP.

Baker, M. 1993. "Corpus linguistics and translation studies: Implications and applications". In *Text and technology. In honour of John Sinclair*, M. Baker, G. Francis and E. Tognini-Bonelli (eds), 233–250. Philadelphia/Amsterdam: John Benjamins.

Biber, D. 2003. "Compressed noun phrase structures in newspaper discourse: The competing demands of popularization vs. economy". In *New media discourse*, J. Aitchison and D. Lewis (eds). New York: Routledge.

Biber, D. and Gray, B. 2010. "Challenging stereotypes of academic writing: Complexity, elaboration, explicitness". *Journal of English for Academic Purposes* 9: 2–20.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1): 5–32.

Crawley, M. 2007. *The R Book*. West Sussex: Wiley and Sons.

Dalgaard, P. 2008. *Introductory Statistics with R*. Berlin: Springer.

Filipovic, R. 1974. "The use of a corpus in contrastive studies". In *Trends in kontrastiver Linguistik*, H. Raabe (ed), 1: 51–66. Tübingen: Gunther Narr.

Gelman, A. and Hill, J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: CUP.

Gries, S.T. 2009. *Statistics for linguistics with R*. Berlin/New York: Mouton De Gruyter.

Granger, S. 1994. "The Learner Corpus: A revolution in applied linguistics". *English Today* 39 10 (3): 25–29.

Granger, S. 2004. "Computer learner corpus research: Current status and future prospects". In *Applied Corpus Linguistics: A Multidimensional Perspective*, U. Connor and T.A. Upton (eds), 123–145. Amsterdam: Rodopi.

Hartmann, R.R.K. 1996. "Contrastive textology and corpus linguistics: On the value of parallel texts". *Language Sciences* 18: 947–957.

Hothorn, T., Hornik, K. and Zeileis, A. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework". *Journal of Computational and Graphical Statistics* 15(3): 651–674.

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

Johansson, S. and Hofland, K. 1994. "Towards an English-Norwegian Parallel Corpus". In *Creating and using English language corpora*, U. Fries, G. Tottie and P. Schneider (eds), 25–37. Amsterdam and Atlanta: Rodopi.

Keenan, E.L. and Comrie, B. 1977. "Noun phrase accessibility and universal grammar". *Linguistic Inquiry* 8(1): 63–99.

Langacker, R. 2008. *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

Nelson, G. 1996. "The Design of the Corpus". In *Comparing English Worldwide: The International Corpus of English*, S. Greenbaum (ed), 27–35. Oxford: Clarendon Press.

Peduzzi, P., Concato, J., Kemper E., Holford, T.R. and Feinstein, A.R. 1996. "A simulation study of the number of events per variable in logistic regression analysis". *Journal of Clinical Epidemiology* 49: 1372–1379.

Prince, E. 1981. "Toward a taxonomy of given-new information". In *Radical pragmatics*, P. Cole (ed), 223–255. New York: Academic.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org [last accessed 25 August 2011].

Szmrecsanyi, B. 2011. "Corpus-based dialectometry: a methodological sketch". *Corpora* 6(1): 45–76.

Stoll, S. and Bickel, B. 2009. "How deep are differences in referential density?" In *Crosslinguistic Approaches to the Psychology of Language: Research in the Traditions of Dan Slobin*, E. Lieven, J. Guo, N. Budwig, S. Ervin-Tripp, K. Nakamura, and Ş. Özçalişkan (eds), 543–555. London: Psychology Press.

Strasser, H. and Weber, C. 1999. "On the asymptotic theory of permutation statistics". *Mathematical Methods of Statistics* 8: 220–250.

Strobl, C., Boulesteix, A.M., Kneib,T., Augustin, T., and Zeileis, A. 2008. "Conditional Variable Importance for Random Forests". *BMC Bioinformatics* 9: 307.

Strobl, C., Malley J., and Tutz, G. 2009. "An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests". *Psychological Methods* 14(4): 323–348.

Author's address

Daniel Wiechmann
Institut für Anglistik, Amerikanistik & Romanistik (IFAAR)
RWTH Aachen University
Karmanstraße 17/19
52074 Aachen
Germany

wiechmann@anglistik.rwth-aachen.de