Statistical tests for the analysis of learner corpus data

Stefan Th. Gries

This paper is an overview of several basic statistical tools in corpus-based SLA research. I first discuss a few issues relevant to the analysis of learner corpus data. Then, I illustrate a few widespread quantitative techniques and statistical visualizations and exemplify them on the basis of corpus data on the genitive alternation – the *of*-genitive vs. the *s*-genitive from German learners and native speakers of English. The statistical methods discussed include a test for differences between frequencies (the chi-squared test), tests for differences between means/medians (the *U*-test), and a more advanced multifactorial extension, binary logistic regression.

1. Introduction

1.1 General introduction

Linguistics as a whole and nearly all of its sub-branches are currently undergoing a change to becoming much more empirical, much more rigorous, and much more quantitative/statistical. While most, though of course not all, of 20th century linguistics was characterized by a reliance on what some have referred to as armchair linguistics, where a linguist develops a theory and at the same time makes up the data – usually acceptability judgments of decontextualized isolated sentences – this situation is very different now. In many, maybe most, linguistic fields, we now routinely find studies that use experimental designs and/or sophisticated analyses of corpus data. In tandem with this development to more objective and rigorous processes of data gathering, there is also a development towards more rigorous data analysis: statistical analysis of various levels of complexity have become a mainstream component of linguistic analysis.

This is a good development: results of quantitative studies often afford us with higher degrees of comparability, objectivity, replicability, and precision. Consider the following hypothetical discussion of data on the genitive alternation:

> The correlation between different semantic roles and possessors and the choice of an *of*- or an *s*-genitive changes as foreign language learners become more advanced. For beginners, the semantic role of the possessor does not play much of a role, but as soon as they reach an intermediate stage, possessors' semantic roles become much more important. By contrast, the transition from intermediate to advanced learners does not make much of a difference anymore for how possessors correlate with genitive choices.

Even if we leave aside for now how 'beginners', 'intermediate', and 'advanced learners are defined, this statement is still too imprecise to be useful. What does 'play not much of a role mean'? How much is 'much more important'? And how little is 'not much of a difference'? If I replicated that study and found a 10% difference between intermediate and advanced learners – is that finding compatible with the one reported above or not? Or would a difference of 20% be? And is a change of 10% (or 20%) significant or not, i.e. probably not due to chance or a sampling accident?

Proper statistical analysis addresses these and many other problems. In this paper, I can obviously not provide a full-fledged introduction to quantitative methods in linguistics (cf. Section 4.2 for references) or second/foreign language learning research, but a few first introductory steps are nonetheless possible. In Section 1.2 I will discuss a variety of caveats regarding the use of corpus data in SLA research. In Section 1.3, I will exemplify how to set up corpus data for statistical analysis and then present the data I will use to exemplify some statistical methods. In Section 2, I will explain the logic and application of several frequent and simple statistical tools to analyze quantitative learner corpus data. In Section 3, I will provide two short examples of binary logistic regression as a primer to more complex, but also more interesting multifactorial methods (i.e. methods involving the impact of several causes on an effect). Section 4 will conclude.

1.2 A very brief view on caveats regarding learner corpus research

In the last 20 years or so, the area of learner corpus research has been among the most booming areas in corpus linguistics. In particular the research undertaken and the corpora compiled at the Centre of English Corpus Linguistics at the *Université Catholique de Louvain* led by Sylviane Granger have inspired a whole field of learner corpus researchers. Resources such as the International Corpus of Learner English (ICLE), the Louvain International Database of Spoken English Interlanguage (LINDSEI), and, for the purpose of comparison, the Louvain Corpus of Native English Essays (LOCNESS) and the Louvain Corpus of Native English Conversation (LOCNEC) have literally transformed the field into a thriving empirical discipline.

In spite of the constantly growing number of resources, there are still many caveats to consider, nearly all of which have to do with the variability of the data. Some threats to the reliability and validity of our studies have to do with the degree to which we can conflate and compare different learner corpora and/or native speaker comparison corpora. The evaluation of data from such corpora involves a larger number of dimensions to be taken into consideration, some of which involve the compilation and annotation of the corpora *per se*, while others involve the retrieval and analysis of examples from the corpora:

- dimensions related to the speakers: their first language (and maybe the dialect they are speaking in that first language), other second/foreign languages they have learned and/or speak, their overall academic proficiency, ...;
- dimensions related to the circumstances of collection: the medium/register in which the data are produced, constraints on the topic and the time of production (e.g. in the typical kind of essay collections), the possibility (or lack thereof) of using dictionaries or other resources (e.g. the internet) during production, whether or not the data are tainted by feedback from instructors, whether or not the software that, say, the learner used to write a text featured a spell- and/or grammar checker, ... (cf. Lozano & Mendikoetxea this volume);
- dimensions involving annotation: annotation is already difficult and far from uncontroversial in native speaker data – what part-of-speech tags to use, whether to try and impose a syntactic parse on the data, etc. – and things are even more complex with learner data where automatic lemmatizers, taggers, and parsers may not be able to handle, say, the effect of misspellings on POS-tagging and subsequent parsing or the creative syntactic choices learners may use, and where somewhat subjective decisions may be called for in the tagging of errors. In addition, much of what constitutes non-native expression by learners may only unidiomatic, but not real errors, ... (cf. Reznicek et al. this volume);
- dimensions involving retrieval: misspellings, etc. can of course not only affect annotation but also the mere retrieval of data. For instance, the study of unannotated learner corpora would be impacted by learners' confusions of, say, *there* and *their* or *lose* and *loose*, or learners problems with *accept<u>a</u>ble* vs. *corrupt<u>i</u>ble*, *teach<u>e</u>r* vs. *act<u>o</u>r* vs. *li<u>a</u>r*, or *beli<u>e</u>ve* vs. *rec<u>e</u>ive* because searches based on exact character strings may fail to receive misspelled target structures.

All these very real problems notwithstanding, it is clear that the growing availability of learner corpus resources has a tremendously positive impact on the field and is a prerequisite for the also growing number of rigorous quantitative studies in this field.

1.3 The corpus data: The genitive alternation (*of*- vs. *s*-genitives)

This chapter will exemplify several statistical tests on the basis of a sample of data from a large study conducted with Stefanie Wulff (University of Florida); cf. Gries & Wulff (2013) for results from a larger data set. Here, I will use a small random sample of our data on the genitive alternation, i.e. the choice between *of*- and *s*-genitives as exemplified in (1), by native speakers of English vs. German learners of English (from the German part of ICLE):[1]

(1) a. the speech of the President — *of*-genitive: possessed *of* possessor
b. the President's speech — *s*-genitive: possessor*'s* possessed

Previous research on native speaker English has identified many different variables that are correlated with speakers' constructional choices. Some of these variables directly involve a particular genitive choice, whereas others are more general preferences regarding characteristics of the speech stream and may favor different constructions at different times. The former include, but are not limited to:

- the number of the possessor: plural possessors prefer *of*-genitives and irregular plurals prefer *s*-genitives (cf. Altenberg 1982; Plank 1985);
- the animacy of the possessor and the possessed (Anim_Possor and Anim_Possed): human possessors prefer *s*-genitives and non-human possessors prefer *of*-genitives (cf. Leech et al. 1994; Biber et al. 1999);
- meanings and functions of the genitives: for example, prototypical possession as in *Peter's car* prefers *s*-genitives whereas depiction as in *the pictures of the party* prefers the *of*-genitive (cf. Stefanowitsch 2003);
- the lengths of the possessor (Len_Possor) and the possessed (Len_Possed) (cf. Cooper & Ross 1975; Bock 1982) come together to yield a general short-before-long preference;
- the related criterion of syntactic-branching direction: post-modified possesseds as in *the study on attention of Nick* would actually prefer an *s*-genitive whereas post-modified possessors as *the study of Nick, who is at the U of M* prefer *of*-genitives, etc. (cf. Rosenbach 2002).

The latter include some well-known factors but also several somewhat understudied variables such as:

- rhythmic alternation: the dispreference of having two stressed syllables or three or more unstressed syllables follow each other (cf. Selkirk 1984); accordingly,

1. I am using *genitives* to refer to both constructions, and *possessor* and *possessed* as convenient cover terms; of course, both constructions can be used with many more diverse semantic roles.

the stress clash in *Emile's portrait* would make this dispreferred compared to *the portrait of Emile*;
- segment alternation: the preference for CV alternations especially at word boundaries (cf. Hayes 2008); accordingly, *Mary's idea* would be preferred compared to *the idea of Mary*;
- horror aequi: formally identical structures in very close succession as in *Steffi's brother's dog* are dispreferred (cf. Brugmann 1909).

For reasons of space, I can only focus on a very small set of variables, namely Anim_Possor, Len_Possed, and, crucial in an SLA context, a variable called Speaker, which has two levels, learner and native, reflecting whether a particular genitive in the corpus data was used by a second/foreign language learner of English or a native speaker.

Trivially, before any statistical analysis of (corpus or experimental) data can be undertaken, two steps are necessary. First, the data to be analyzed statistically have to be gathered and then organized in a suitable format. Second, they must be saved in a way that allows their import into statistical software. As for the first step, it is absolutely essential to store the data to be analyzed statistically in a spreadsheet software application such that they can be easily evaluated both with that software as well as with statistical software. There are three main rules that need to be considered in the construction of the required so-called *case-by-variable format*:

- each data point, i.e. count or measurement of the dependent variable(s), is listed in a row on its own;
- every variable with respect to which each data point is described is recorded in a column on its own;
- the first row contains the names of all variables.

In our example involving genitives, the raw data should be organized as in Table 1. The column Match contains the matches from the concordance lines; the column Genitive contains the dependent variable (*of* vs. *s*); the columns Anim_Possor and Anim_Possed contain the categorical independent variables related to number (animate vs. inanimate); the columns Len_Possor and Len_Possed contain the lengths of the possessors and possesseds in words.

Once the data have been organized in this way, the second step before the statistical analysis is to save them such that they can be easily loaded into a statistics application. To that end, one should save the data into a format that makes them maximally readable by a wide variety of programs. The simplest way to do this is to save the data into a tab-separated file, i.e. a raw text file in which different columns are separated from each other with tabs. In LibreOffice Calc, one first

Table 1. Example of the format of a raw data table

Match	Genitive	Anim_ Possor	Anim_ Possed	Len_ Possor	Len_ Possed	...
the ball *of* our dog	of	animate	inanimate	2	2	...
the problems *of* poverty	of	inanimate	inanimate	3	3	...
People*'s* worries	s	animate	inanimate	2	2	...
the cars *of* all those folks	of	animate	inanimate	3	2	...
...	...	...	...	...	...	...

chooses *File: Save As...*, then chooses *Text CSV (.csv)* as the file type, and chooses {Tab} as the *Field delimiter*.[2]

To load the data into a statistical software, one must first of all decide on which software to use. From my point of view, the best statistical package currently available is the programming language and software environment R (cf. R Development Core Team 2013). R is extremely powerful – in fact, since R is a programming language, it can do whatever a user is able to program. In addition, R's graphical facilities are nearly unlimited and as an open source project, it is freely available and has extremely fast bugfix-release times. For these and many other reasons, R is used more and more widely in the scientific community, and I will use it here, too.

When R is started, by default it only shows an empty console and expects user input from the keyboard. The input to R consists of what are called *functions* and *arguments*. Just like in a spreadsheet software, functions are commands that tell R what to do; arguments are specifics for the commands, namely what to apply a function to (e.g. a value, the first row of a table, a complete table, etc.) or how to apply the function to it (e.g. what kind of logarithm to compute, a binary log, a natural log, etc.). A companion file available from the author's website at <http://tinyurl.com/stgries> contains all the R code that would be necessary to conduct the statistical tests and generate the plots discussed in this paper; to run the code below, read and then copy and paste the relevant functions from the code file into R.

2. Elementary statistical tests

The first step towards statistical analysis is to read the data into R. One way do this involves the function read.table, which, if the raw data table has been created as

2. I recommend using only word characters (letters, numbers, and underscores) within such tables and steer clear of spaces, dollar signs, asterisks, hyphens and other non-word characters. While this is strictly speaking not necessary to guarantee proper data exchange between different programs, it is my experience that simple works best.

outlined above and in note 2, requires only a few arguments specifying which file to load, whether the first row contains names for all columns, and how columns are separated from each other.

To check whether the data have been read in successfully, it is always useful to look at the structure of the imported data first, using the function str, which provides all the column names together with some information on what the columns contain, namely their kind of data (integer numbers, character strings as factors, etc.) as well as the first few values. If you read in a file of the kind shown in Table 1, then this is what the output of str looks like:

```
'data.frame':     600 obs. of 8 variables:
$ CASE:           int 1 2 3 4 5 6 7 8 9 10 ...
$ SPEAKER:        Factor w/ 2 levels "learner","native": 2
2 2 2 2 2 2 2 2 ...
$ MEDIUM:         Factor w/ 2 levels "oral","written": 1 1
1 1 1 1 1 1 1 ...
$ GENITIVE:       Factor w/ 2 levels "of","s": 1 1 1 1 1 1
1 1 1 1 ...
$ ANIM_POSSOR:    Factor w/ 2 levels
"animate","inanimate": 2 2 2 2 2 1 2 2 ...
$ ANIM_POSSED:    Factor w/ 2 levels
"animate","inanimate": 2 2 1 2 2 1 2 2 ...
$ LEN_POSSOR:     int 7 5 2 2 1 1 3 1 13 20 ...
$ LEN_POSSED:     int 5 6 1 1 1 1 5 2 2 5 ...
```

To then be able to access every variable by means of its column name, one can use the function attach together with the name of the loaded data, here raw.data.

2.1 Two-dimensional frequency tables: Chi-squared tests

The first application to be discussed here involves two-dimensional frequency tables, i.e. research scenarios in which one wants to explore if/how two categorical variables are related. As an example, we will explore whether the animacy of the possessor is correlated with the choice of genitive separately for the data by learners and by native speakers. Since both these variables involved (Anim_Possor and Genitive) are categorical, the default method for exploring this correlation involves frequency tables. In R, one can use the function table, together with the names of all variables to be cross-tabulated. In this case, three variables are involved: Genitive (the dependent variable), Anim_Possor (the independent variable), and Speaker (a potential moderator variable to explore the question of whether the relationship between Genitive and AnimPossor is different in the

two speaker groups). Ideally, one computes a three-dimensional frequency table and stores it in a variable/data structure; below, I show one possible result of this approach:

```
,, SPEAKER = learner
             GENITIVE
ANIM_POSSOR  of   s
  animate     38  55
  inanimate  190  17

,, SPEAKER = native

             GENITIVE
ANIM_POSSOR  of   s
  animate     16  102
  inanimate  134   48
```

One effect is brought out very clearly by this representation: there seems to be a strong correlation between Anim_Possor and Genitive: For both learner and native speakers, the *of*-genitive is strongly preferred when the possessor is inanimate, and the *s*-genitive is preferred when the possessor is animate. However, it is still unclear whether the above differences are large enough to be significant, i.e. most likely not just due to chance. This question can be addressed by the chi-squared test for independence. This test requires that all observations are independent of each other (e.g. when they have all been produced by different speakers), that 80+% of the frequencies that would be expected by chance are ≥ 5, and that all of the expected frequencies are ≥ 1 (cf. Sheskin 2011: 638ff).

We assume for now that all genitives are independent of each other (and will check the expected frequencies shortly). One can use the function chisq.test, which standardly requires the two-dimensional table to be tested and an argument correct, which can be set to TRUE or FALSE depending on whether one wants to use a correction for continuity, which we here do not want (because the sample size is greater than 20). For reasons that will become clear shortly, it is best to not just compute the test *per se* but also assign the result of the test to another data structure so we compute two chi-squared tests – one for the learners, one for the native speakers – and assign the two tests to two data structures: learners and natives. These are the results, again first for learners, then for native speakers:

```
       Pearson's Chi-squared test
data: contig.table[,, 1]
X-squared = 91.2446, df = 1, p-value < 2.2e-16
```

```
       Pearson's Chi-squared test
data: contig.table[,, 2]
X-squared = 103.3153, df = 1, p-value < 2.2e-16
```

The tests show that there are highly significant effects: the above-mentioned preferences of animate and inanimate possessors are extremely unlikely to occur by chance. One question now is whether the expected frequencies are large enough to allow the chi-squared test in the first place. The chi-squared test in R computes more than the above output and we can access the expected frequencies from the learners and natives; these are the results, again first for learners, then for native speakers:

```
             GENITIVE
ANIM_POSSOR  of      s
  animate     70.68  22.32
  inanimate  157.32  49.68

             GENITIVE
ANIM_POSSOR  of      s
  animate    59      59
  inanimate  91      91
```

Clearly, all expected frequencies are greater than or equal to 5 so the chi-squared test is unproblematic here.

The other central question is what this correlation looks like. The part of the results that is useful to understand the nature of the correlation involves the so-called Pearson residuals, here rounded to two decimals, first for learners, then for native speakers. Pearson residuals are positive and negative when a cell's observed frequency is larger or smaller than expected respectively, and the more the residuals deviate from 0, the stronger the effect they reflect.

```
             GENITIVE
ANIM_POSSOR  of     s
  animate    -3.89   6.92
  inanimate   2.61  -4.64

             GENITIVE
ANIM_POSSOR  of     s
  animate    -5.60   5.60
  inanimate   4.51  -4.51
```

While this result mainly shows that the conclusions we already drew from the observed frequencies are borne out, Pearson residuals are very useful when tables with more than four cells are studied. In addition, the deviations of the residuals from zero indicate that, for instance, the strongest effect in the data of the learners is their strong preference for *s*-genitives with animate possessors (residual = 6.92) – for the native speakers, it is also the animate possessors that exhibit the strongest effects (residuals = ±5.6).

One graphical representation that highlights the results even more clearly is the so-called association plot, which is shown in Figure 1: black boxes on top of the dashed lines and grey boxes below the dashed lines represent cell frequencies that are larger and smaller than expected respectively; the heights of the boxes are proportional to the above residuals and the widths are proportional to the square roots of the expected frequencies (so that one can easily identify cells where very small expected frequencies might skew the results).

The only thing that remains to be done is quantify the size of the effect. Since chi-squared values and *p*-values are correlated with sample sizes, one cannot use them to identify effect sizes or compare them across different studies. Instead, one can use a correlation coefficient called Cramer's *V*, which falls between 0 and 1, and the larger the value, the stronger the correlation. Cramer's *V* is computed as shown in (2). We obtain 0.551 for the learners and 0.587 for the native speakers, i.e. fairly strong correlations (there seem to be no uniformly accepted guidelines for the evaluation of *V*).

$$(2) \quad \text{Cramer's } V = \frac{X^2}{n(min(n_{rows}, n_{columns}) - 1)}$$

While we have now reached a good understanding of the role of Anim_Possor for Genitive, i.e. the interaction of Anim_Possor and Genitive, one important question has remained unclear and can in fact not be straightforwardly tested with the chi-squared test from above. That is the question of whether the effect of Anim_Possor on Genitive is the same for both learners and native speakers: while both speaker groups exhibit a significant effect of Anim_Possor in the above-mentioned direction, it is not obvious that the strength of that effect is identical, too. This question amounts to testing the three-way interaction of Anim_Possor, Genitive, and Speaker, for which a multifactorial approach of the type discussed below is needed. We will see below in 3.1 and 4 that this three-way interaction is indeed not significant: the learners' genitive choices with regard to Anim_Possor are not significantly different from those of the native speakers.

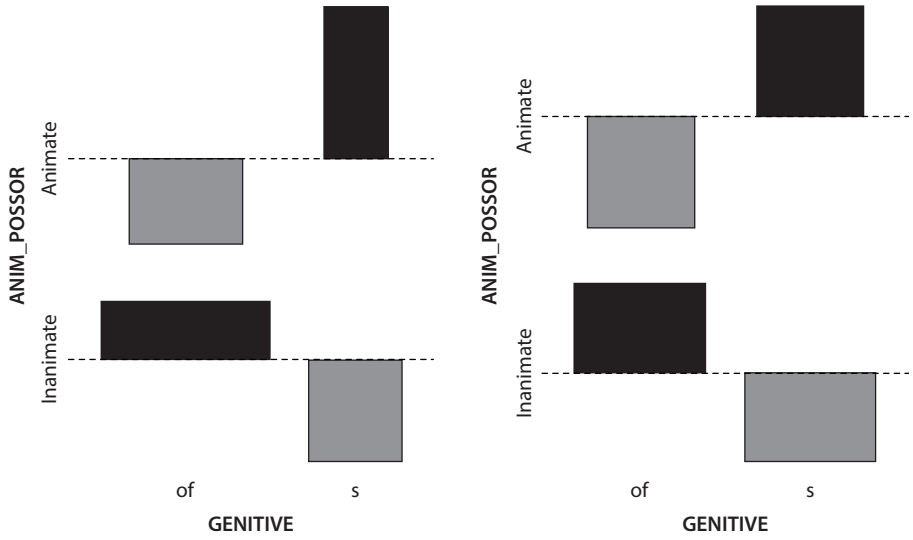

Figure 1. The relation between Genitive and Anim_Possor in an association plot

2.2 Measures of central tendency

The second application to be exemplified involves how to test whether measures of central tendency – means or medians – in two groups differ significantly from each other. As an example, we will consider the question of whether the possessed elements are differently long in both genitives for, again, both the learners and the native speakers.

As a first step, one can compute the means of the possesseds in both genitives across both speaker groups:

```
of          s
3.814815    3.531532
```

Apparently, the two average lengths are rather close to each other but we also need to include the different speaker groups, which changes the picture considerably in that the learners seem to use the genitives differently: In the learner data, the possessed of *s*-genitives is longer whereas in the native speaker data, the possessed of *of*-genitives is longer.

```
        learner    native
of      3.82       3.80
s       4.40       3.11
```

However, one should never compare means without corresponding measures of dispersion (e.g. a standard deviation or a confidence interval), and all these measures are only useful when the data averaged across are approximately normally distributed and have maximally very few outliers. Standard deviations are easy to compute and show quite some variation here, but Figure 2 reveals that none of the lengths of the possesseds are normally distributed at all: the four panels show historgrams with the frequencies of all possessed lengths in the four groups one obtains by crossing Genitive and Speaker.

```
        learner    native
of      2.03       2.31
s       2.85       2.43
```

It is therefore more prudent to compute medians and interquartile ranges: medians are the values one obtains by sorting all values from small to large and choosing the middle one, and the interquartile range is the range of the central 50% of the values around the median. Apparently, there is still the above-mentioned difference between learners and native speakers, but the interquartile ranges are quite high so these differences may not be significant.

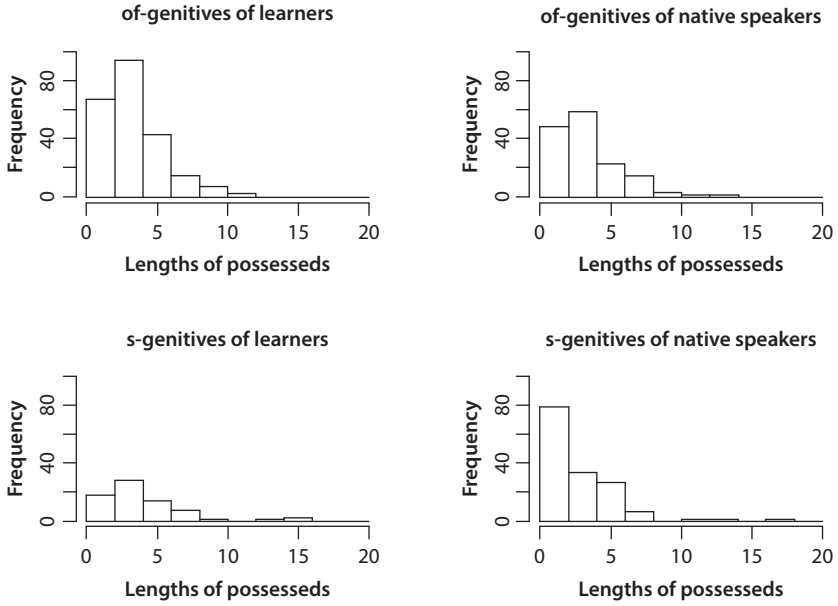

Figure 2. Histograms of Len_Possed for all combinations of Genitive and Speaker

```
        learner    native
of      3          3
s       4          2

        learner    native
of      3.00       3
s       3.25       3
```

To test whether the two genitives differ with regard to LEN_POSSED for each speaker group – i.e. to test whether two medians are significantly different – we can compute two *U*-tests, one for learners and one for native speakers. The results show that the learners' genitives do not differ significantly with regard to LEN_POSSED but the native speakers' genitives do. This effect can also be seen easily in a boxplot as represented in Figure 3, where the horizontal lines in the middles of the boxes represent the median lengths in each of the four groups and the long dashed line represents the overall median.

```
      Wilcoxon rank sum test
data: LEN_POSSED by GENITIVE
W = 7485.5, p-value = 0.2529
alternative hypothesis: true location shift is not equal to 0
      Wilcoxon rank sum test
data: LEN_POSSED by GENITIVE
W = 13742.5, p-value = 0.0007662
alternative hypothesis: true location shift is not equal to 0
```

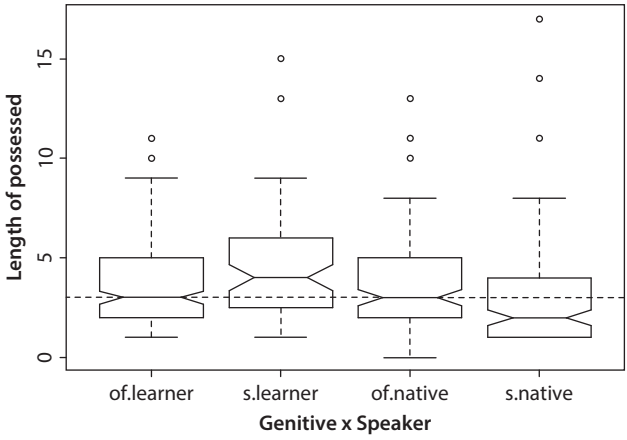

Figure 3. Boxplot of LEN_POSSED as a function of GENITIVE and SPEAKER

The results for the native speakers are somewhat unexpected: in some previous studies, characteristics of the possessed did not play much of a role, and in many studies of constituent order alternations more generally, effects of short-before-long were obtained. Here, however, the native speakers choose the *of*-genitive with longer possesseds than the *s*-genitive, which appears to contradict short-before-long. A more comprehensive study of this phenomenon cannot be conducted here for reasons of space, but it (i) should involve a multifactorial analysis for whether whatever findings one obtains apply to both learners and native speakers – again a three-way interaction, namely of LEN_POSSED, GENITIVE, and SPEAKER, which statistical analysis reveals to be significant (see below 3.2 and 4) – and (ii) could involve pairwise comparisons of LEN_POSSOR and LEN_POSSED for all *of*- and *s*-genitives.

3. A primer on multifactorial methods: Logistic regression

While both examples above were quite simple in their design, they already were more complex than the type of monofactorial tests discussed can handle. Essentially, all linguistic phenomena are multifactorial in nature: there is always more than one cause for any given effect and often we need to take moderator and confounding variables into consideration. It is therefore essential that our methods reflect this fact and can handle the complexities that arise from the combination of many independent variables. Very often, the method of choice is one of the family of regression techniques. Regression models are a statistical technique in which an effect, or the variability of a dependent variable, is explored on the basis of one or more independent variables and (often) their interactions, where an interaction between *n* variables is defined as a non-additive unexpected effect once the *n* variables are considered jointly. To that end, this approach expresses a statistical model in the form of a regression equation. This equation predicts values for the dependent variable which can then be compared to the actually observed values to determine how well the model fits the reality it tries to model.

On one level, regressions can be distinguished depending on the nature of the dependent variable: if the dependent variable is interval-/ratio-scaled, then the typical approach is that of linear regression; if the dependent variable is categorical, then multinomial or polytomous regressions are often used, and if the dependent variable is binary, then one often finds binary logistic regressions (cf. Gries 2013: Ch. 5 for discussion and many worked examples). Correspondingly, in corpus-linguistic studies, linear regressions are fairly rare, but binary logistic regressions are now common and the following two Sections 3.1 and 3.2 exemplify two

applications by following up on Sections 2.1 and 2.2 respectively with logistic regressions.

3.1 Logistic regressions with two categorical independent variables

Above, we saw that the data of both learners and native speakers reveal a strong relationship between Anim_Possor and Genitive, but with the monofactorial approach above it was not easily possible to test whether both speaker groups exhibit the same effect size or not. With a regression approach, this means we have to fit the model exemplified in (3) to the data, which means we want to explore the choice of genitive (Genitive) as a function of (~) (i) whether the possessor is animate or not (Anim_Possor), whether the speaker is a learner or a native speaker (Speaker), and any interaction between Anim_Possor and Speaker (Anim_Possor:Speaker):

(3) Genitive ~ Anim_Possor + Speaker + Anim_Possor:Speaker

One set of results from such a binary logistic regression model are represented below.

```
                          Coef    S.E.    Wald   Z Pr(>|Z|)
Intercept                 0.3697  0.2109   1.75    0.0796
ANIM_POSSOR=inanimate    -2.7836  0.3295  -8.45   <0.0001
SPEAKER=native            1.4826  0.3418   4.34   <0.0001
ANIM_POSSOR=inanimate -0.0955  0.4574  -0.21    0.8347
* SPEAKER=native
```

For all three predictors, the two independent variables and their interaction, we obtain

- a coefficient (and its standard error), which reflects the effect the predictor has on the choice of the *s*-genitive: positive and negative coefficients indicate that the predictor values shown increase and decrease the probability of an *s*-genitive respectively;
- a Wald *z*-score (the quotient of the coefficient and its standard error) and a *p*-value resulting from that *z*-score, which reflects whether the predictor has a significant impact on the dependent variable or not.

This result shows that there is no significant interaction between Anim_Possor and Speaker (note the *p*-value of 0.8347 for the interaction in the grey box above), which means that the predictor Anim_Possor has the same effect on Genitive in both speaker samples. This also means that, according to Occam's razor, that

variable should be deleted and a new model without it should be fit. The results of this second model, model.02, are shown below.

```
                          Coef     S.E.     Wald   Z Pr(>|Z|)
Intercept                 0.3901   0.1874    2.08    0.0374
ANIM_POSSOR=inanimate    -2.8335   0.2283  -12.41   <0.0001
SPEAKER=native            1.4297   0.2274    6.29   <0.0001
```

In this simple case, understanding the results is straightforward: The coefficient for ANIM_POSSOR: INANIMATE is negative, which means that inanimate possessors, compared to animate ones, decrease the probability of *s*-genitives. The coefficient for SPEAKER: NATIVE is positive, which means that native speakers use the *s*-genitives more. In more complex cases, however, visualization is essential to understanding the results. One way to visualize such results is by means of barplots of the probabilities predicted by the model, and it is often good practice to plot different perspectives on the same results, as exemplified by the two panels of Figure 4. Here, both perspectives show, as we already saw above, that animate possessors significantly increase the chances of *s*-genitives for both speaker groups.

The final step to undertake is to assess how well this second model accounts for the genitive choices. The simplest way in which this can be done involves inspecting additional output of the logistic regression. Typically, one obtains

- an R^2-value that can range from 0 to 1, and the higher, the better the model explains the data;
- a *C*-value that can range from 0.5 to 1, and the higher, the better the predictions of the model are (a frequently-cited threshold value for good models is 0.8).

In this case, $R^2 = 0.456$ and $C = 0.841$, which represents a good model fit.

A second way to evaluate the model is to directly compare the genitive choices predicted by the model against the observed choices in the data and compute the resulting classification accuracy. As Table 2 indicates, the model classifies 481 out of 600 genitives correctly, which results in an classification accuracy of 80.17%, which is significantly better than what the least sophisticated model would achieve, the model that simply picks the more frequent genitive all the time and, thus, only gets 378 out of 600 (63%) right.

3.2 Logistic regressions with one categorical and one numeric independent variable

Above, we saw LEN_POSSED had an effect on GENITIVE, but differently in both speaker samples and only significantly so for the native speakers. This seems to

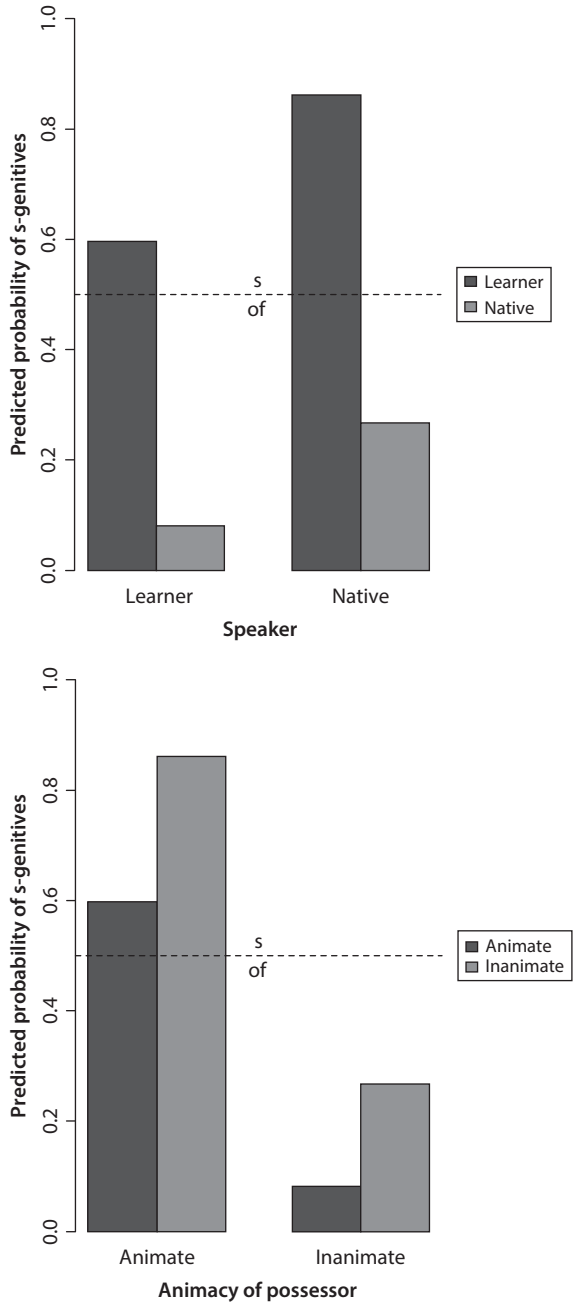

Figure 4. Barplot of the probabilities of *s*-genitives predicted by model 2

Table 2. Classification accuracy of model 2 of Section 3.1

	Predicted: *of*	**Predicted:** *s*	**Totals**
Observed: *of*	324	54	378
Observed: *s*	65	157	222
Totals	389	211	600

indicate an interaction in that the effect of LEN_POSSED on GENITIVE is dependent on SPEAKER and, here, we test whether this interaction is significant using the model in (4).

Let us now return to the effect of LEN_POSSED with a similar regression approach.

(4) GENITIVE ~ LEN_POSSED + SPEAKER + LEN_POSSED:SPEAKER

The results of the regression are shown below.

```
                          Coef     S.E.     Wald   Z Pr(>|Z|)
Intercept                -1.5838  0.2743   -5.77    <0.0001
LEN_POSSED                0.1053  0.0565    1.86     0.0624
SPEAKER=native            2.0183  0.3461    5.83    <0.0001
LEN_POSSED *             -0.2319  0.0768   -3.02    -0.0025
SPEAKER=native
```

Thus, LEN_POSSED has an only marginally significant effect on the choice of genitive: its coefficient is positive, which means that, as the possessed becomes longer, the *s*-genitive becomes more likely. SPEAKER: NATIVE, on the other hand, is significant and positive again: native speakers use *s*-genitives more compared to learners. However, neither of these effects can be taken at face value because the two predictors' interaction qualifies these interpretations: For native speakers, the positive effect of LEN_POSSED is reversed. However, the exact effect is difficult to assess numerically, which is why a plot of the predicted probabilities of *s*-genitives is more useful. Figure 5 plots *l*s and *n*s for learners and native speakers respectively and illustrates the interaction nicely: for learners, increased values of LEN_POSSED (on the *x*-axis) lead to a higher probability of *s*-genitives (on the *y*-axis), for native speakers, it is the other way round. As above, this means that the learners' choices are more compatible with the short-before-long preference so often observed for English, but now we also know that this difference between learners and native speakers – the interaction – is indeed significant.

It has to be noted, however, that this model does a rather poor job at explaining the data. $R^2 = 0.12$ and $C = 0.675$, i.e. both quite low and we will see below why,

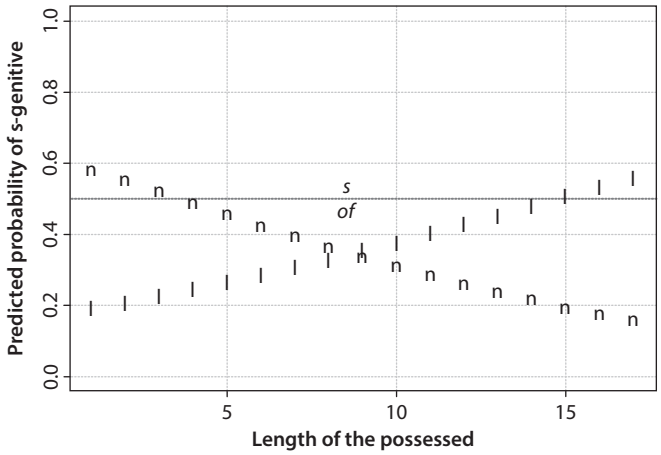

Figure 5. Scatterplot of the probabilities of *s*-genitives predicted by model 1

and the classification accuracy is a mere 65.17% (cf. Table 3), which is not significantly better than the 63% achieved by just picking the more frequent genitive all the time.

An even more advanced analysis would now of course involve a logistic regression with at least both Anim_Possor, Len_Possed, and Speaker and their interactions as predictors or, even better, include more independent variables, such as Anim_Possed, Len_Possor, and many others of the above-mentioned ones. While the detailed discussion of such a model is beyond the scope of the present survey chapter (cf. Gries & Wulff 2013 for that), let me at least represent summary results of a regression modeling process involving Anim_Possor, Len_Possed, and Speaker and their interactions as predictors. Figure 6 reveals that Anim_Possor is significant in the final analysis, as is the interaction between Len_Possed and Speaker. The latter shows that the native speakers do not react much to Len_Possed while the learners are quite sensitive to it in that they prefer genitive choices that preserve short-before-long. This model indicates a fairly strong ($R^2 = 0.47$, $C = 0.85$) and highly significant relationship, with a classification accuracy of 82%.

Table 3. Classification accuracy of model 1 of Section 3.2

	Predicted: *of*	**Predicted:** *s*	**Totals**
Observed: *of*	294	84	378
Observed: *s*	125	97	222
Totals	419	181	600

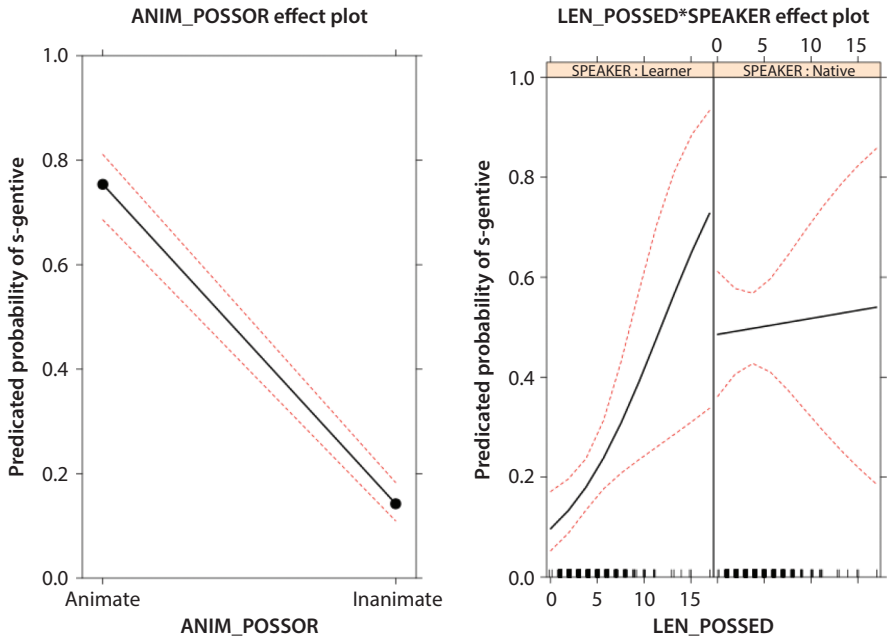

Figure 6. Results of a model selection process involving Anim_Possor, Len_Possed, and Speaker

4. Concluding remarks

4.1 Conditional inference trees: An alternative to regressions

While (linear, logistic, or multinomial) regressions are among the standard tools to handle multifactorial data, other approaches are available. One somewhat popular alternative is the technique of conditional inference trees. The most easily interpretable output of this method when applied to the example of Section 3.1 is exemplified in Figure 7. This technique involves successively splitting up the data into smaller parts that differ most significantly in their patterning of the dependent variable and representing this in a decision-tree like format. To interpret the results, one starts at the top, and the first significant distinction the tree suggests to the analyst is to distinguish between inanimate and animate possessors. If one focuses on inanimate possessors – the left half of the plot – then the next significant distinction is that between learners and native speakers, and the bars show that, when one focuses on the learners, *of*-genitives are *very* much preferred (cf. the small dark grey part of the leftmost bar), whereas when one focuses on the native speakers, *of*-genitives are still preferred, but less so (cf. the larger dark grey part of

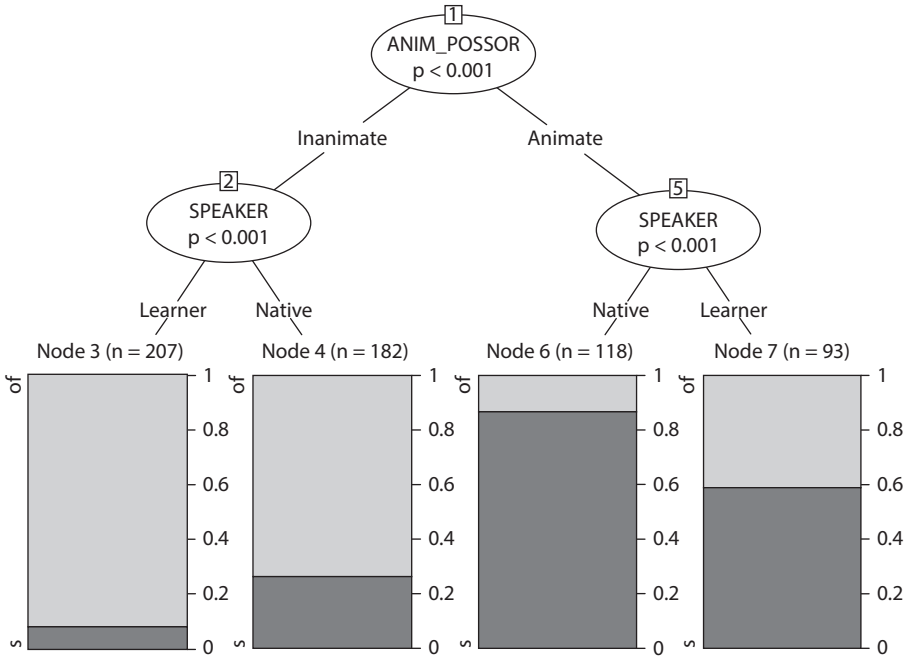

Figure 7. Conditional inference tree for the model discussed in Section 3.1

the second bar). If, on the other hand, one focuses on animate possessors – the right half of the plot – the next significant distinction is again that between learners and native speakers. However, this time the bars show that *s*-genitives are preferred, in particular by native speakers.

This approach can often be a useful alternative to regressions, or an addition if the data one wishes to analyze violate distributional assumptions of the regression.

4.2 Pointers to additional references

As mentioned at the outset of this chapter, rigorous quantitative analyses are not yet as frequent in (applied) linguistics as they could be, but they are on the rise. This chapter could only discuss a few basic methods and even those were only discussed summarily, so this chapter must close with a variety of recommendations for further reference: on R in general, cf. Crawley (2008), and on R for linguists in particular: Baayen (2008), Gries (2009, 2013) and Johnson (2008). In addition, the WWW provides a lot of information on statistical analysis with R, and the following websites are potentially very useful points of reference:

- the website of R: <http://www.r-project.org>;
- the CRAN task views: <http://cran.r-project.org/web/views/>;
- the R-lang mailing list: <https://mailman.ucsd.edu/mailman/listinfo/ling-r-lang-l>;
- the Statistics for Linguists with R mailing list <https://groups.google.com/forum/#!forum/corpling-with-r/web/>;
- an electronic textbook for statistics <http://www.statsoft.com/Textbook>.

As the methodological landscape in linguistics is changing, it is important for the progress within our field(s) that we learn how to handle the kinds of complex and multifaceted scenarios linguistic data pose. I hope this chapter has provided a first overview of what's possible and has whetted the readers' appetites to dive more into such statistical methods.

References

Altenberg, B. 1982. *The Genitive vs. the* Of*-construction: A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bock, J.K. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89(1): 1–47.

Brugmann, K. 1909. Das Wesen der lautlichen Dissimilationen. *Abhandlungen der philologisch-historischen Klasse der königlich-sächsischen Gesellschaft der Wissenschaften* 27, 147–178. Leipzig: Teubner.

Cooper, W.E. & Ross, J.R. 1975. World order. In *Papers from the Parasession on Functionalism*, R.E. Grossman, L.J. San & T.J. Vance (eds), 63–111. Chicago IL: Chicago Linguistic Society.

Crawley, M. 2008. *The R Book*. Chichester: John Wiley.

Gries, St.Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.

Gries, St.Th. 2013. *Statistics for Linguistics with R: A Practical Introduction*. Second edition. Berlin: Mouton de Gruyter.

Gries, St.Th. & Wulff, S. 2013. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics* 18(3): 327–356.

Hayes, B. 2008. *Introductory Phonology*. Malden MA: Blackwell.

Joseph, B.D. 2008. Last scene of all .... *Language* 84(4): 686–690.

Johnson, K. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell.

Leech, G.N., Francis, B. & Xu, X. 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In *Continuity in Linguistic Semantics*, C. Fuchs & B. Victorri (eds), 57–76. Amsterdam: John Benjamins.

Plank, F. 1985. The interpretation and development of form alternations conditioned across word boundaries: The case of *wife's*, *wives*, and *wives*. In *Papers from the 4th International Conference on English Historical Linguistics*, R. Eaton, O. Fischer, W.F. Koopman & F. van der Leek (eds), 205–233. Amsterdam: John Benjamins.

R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org. [Accessed 22.2.2013]

Rosenbach, A. 2002. *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies*. Berlin: Mouton de Gruyter.

Selkirk, E.O. 1984. *Phonology and Syntax: the Relation between Sound and Structure*. Cambridge MA: MIT Press.

Sheskin, D.J. 2011. *Handbook of Parametric and Nonparametric Statistical Procedures*. Boca Raton FL: Chapman and Hall/CRC.

Stefanowitsch, A. 2003. Constructional semantics as a limit to grammatical alternation. In *Determinants of Grammatical Variation in English*, G. Rohdenburg & B. Mondorf (eds), 413–443. Berlin: Mouton de Gruyter.