Writing up a corpus-linguistic paper

Stefan Th. Gries and Magali Paquot

Abstract In this chapter, we provide a brief characterization of what we consider the best and most common structure that empirical corpus-linguistic papers can and should have. In particular, we first introduce the four major parts of a corpus linguistics paper: "Introduction", "Methods", "Results", and "Discussion". Since the nature of corpus data and corpus techniques makes the two sections very field-specific, we then focus more particularly on the "Methods" and "Discussion" sections of a typical quantitative corpus linguistic paper. We provide recommendations that span the research cycle from data description to analyzing the dataset and reporting the results of statistical tests.

26.1 The Structure of an Empirical Paper

As shown in this volume, conducting an empirical analysis of language using corpus data requires that corpus linguists first make informed decisions related to the type/s of corpus, variables and methods needed to answer their research question/s. Next, they need to analyze (quantitatively) corpus data in a scientific and transparent way. Importantly, all the steps taken and decisions made will need to be reported in a corpus linguistic paper. To report a quantitative/empirical study, researchers in a variety of disciplines typically adopt the overall structure represented in Table 26.1.

The 'Introduction' aims at motivating the research question(s); typically, this is based on previous (published or presented) research and/or relevant observations of the phenomenon. For instance, previous studies may have come to results that are difficult to reconcile or previous studies may have not covered a certain part of the relevant population (in the statistical sense), or certain real-life observations do not appear to be explainable with the current state of the art in the field, etc. Ideally, therefore, the introduction leads the reader to expect the author to address any of these scenarios with the present study.

The 'Methods' section is concerned with which corpora are used, why and especially how variables, or factors or predictors of interest, are operationalized in the corpus data, how the relevant data points are extracted from the corpus and annotated as required by the questions/hypotheses outlined in the intro, and how they were statistically (or otherwise) analyzed.

The 'Results' section contains all results of all steps of the analysis. This might begin with the number of hits from a corpus query using regular expressions to find matches of the phenomenon in question, how these were winnowed down by disregarding false positives, the result of sampling procedures or data transformation procedures, etc.; other results might include (co-occurrence) frequencies of annotated features. Most importantly, the 'Results' section will contain all results from the statistical exploration (in the case of exploratory/hypothesis-generating studies) or one's evaluation of one's hypotheses (in the case of hypothesis-testing studies). Ideally, one would not just report the results of significance tests,

Stefan Th. Gries
University of California, Santa Barbara and Justus Liebig University Giessen
Magali Paquot
FNRS - Université catholique de Louvain

but also all relevant statistics such as effect directions, effects sizes (raw and/or standardized), indices of model/classifier quality and classification/prediction accuracies, as well as the results pertaining to model/classifier diagnostics and validation; also for most advanced analyses, this is the part where the main results should be visualized in a way that facilitates their comprehension even, but also especially, for readers whose statistical knowledge is more limited.

Finally, the 'Discussion' section interprets the results against the background of the questions/hypotheses discussed in the introduction and contextualizes the results in the light of their bigger-picture implications for subsequent studies of the current or related phenomena, but also for the future development of data, theories, and methods.

Although we will not discuss this further, this template can easily be extended to papers that report more than one empirical case study. Typically, after a general introduction, each case study would have its own 'Methods', 'Results', and 'Discussion' sections. The case studies would then be followed by a 'General Discussion' section that puts everything together and answers the main research questions on the basis of the combined results.

Table 26.1 Structure of a quantitative corpus-linguistic paper (based on Gries, 2017:174)

Part	**Content**
Introduction	What is the question? Motivation of the question: Why is this problem important? Overview of previous relevant work Formulation of hypotheses
Methods	Choice of method (e.g. diachronic vs. synchronic corpus data; tagged vs. untagged corpora; etc.) Source of data (which corpus/corpora?) Operationalization of variables Retrieval algorithm or syntax Software that was used Data filtering/annotation (e.g., how were false hits identified? What did you do to guarantee objective coding procedures? How did you annotate your data? etc.) Choice of statistical test(s) and how they are implemented
Results	Summary statistics Graphic representation Significance test: test statistic, degrees of freedom (where available), and *p* Effect size: the difference in means, the correlation, etc.
Discussion	Implications of the results for your hypotheses Implications of the results for the research area

In this chapter, we focus on how to write the 'Methods' and 'Results' sections of a quantitative corpus linguistic paper since the 'Introduction' and 'Discussion' sections are so phenomenon-dependent that, apart from the general guidelines above, they defy easy discipline-specific characterization. Therefore, for more information about the general content of the 'Introduction' and 'Discussion' sections, we refer to (and strongly encourage students to read)

Chap. 2, "Manuscript Structure and Content", of the *Publication Manual of the American Psychological Association* (2010). However, the 'Methods' and 'Results' sections of many corpus-linguistic papers share some important commonalities that we believe can be usefully summarized for less experienced writers and/or beginning corpus researchers.

26.2 The 'Methods' Section

Given that we prefer to see corpus linguistics as a method rather than a theory (see the special issue of the *International Journal of Corpus Linguistics* 15(3) for a debate of these two views), we believe outlining the methodological details of a corpus study in a way that is comprehensive enough is absolutely central. At a very high level of abstractness, there is really only one rule, which says it all: The characterization of the methods employed in a paper needs to be so precise that the study is reproducible or, more explicitly, that someone who wanted to explore the same thing and had access to the same raw data would be able to follow the description in the methods section such that they end up with the same result (cf. Berez-Kroeker et al. 2017, Branco et al., 2017), and by extension, replicable (cf. Porte 2012).

To meet this objective, the methods section must include a number of compulsory parts. First, it should start with a detailed description of the corpus used. Each corpus type (e.g. diachronic corpora, web corpora, parallel corpora) comes with its own specificities and these should be carefully described. For example, the reader of a learner corpus study needs to know as much as possible about the learners who produced the language samples (what are their language background, proficiency level in the foreign language, age, etc.) and the task settings (what were the learners requested to produce? A timed argumentative essay? A spontaneous dialogue with peers?) (cf. Part I.III for more about the specificities of different corpus types). If relevant, this section should also specify which version of the corpus and what types of annotations available with the corpus were used to answer the research questions. For example, a corpus such as the British National Corpus World XML edition (BNC, BNC Data Consortium 2001, http://www.natcorp.ox.ac.uk/) comes with Part-of-Speech (POS) tags assigned automatically using CLAWS C5 tagset. As far as possible, this description should be accompanied with (a) a proper reference to the corpus used, in the form of a citation to a scientific article or a data paper in which its authors introduced the dataset (corpus compilers often specify how they would like the corpus to be referred to),[1] and, (b) if the dataset is available for research, a permanent link to where to find the corpus.

After a general description of the corpus used, detailed information about corpus pre-processing should be presented. Two major types of pre-treatment can be distinguished, i.e. automatic annotation and sampling. As for the former, if answering the research questions requires an unannotated corpus to be automatically tagged or parsed, details about the tool used will need to be provided. These include:

a. The full name of the tool and its version number.
b. The selected parameters: Depending on the tool, it may be necessary to mention the tagset or the language model used (ideally with a URL). For example, the TreeTagger (Schmid 1994) can be used with two distinct English parameter files that contain different tagsets trained on different corpora, https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/).

[1] This is also a means of bringing credit and recognition to all those involved in corpus compilation.

c. General information on its reliability: If no information is available, it may be necessary to conduct a precision and recall study (see Chap. 2).
d. A full reference and a download link.

Today's corpora can be huge and, depending on the linguistic feature under study, it may often be only possible to analyze a random sample of instances. The method used to select the final dataset should also be carefully described. A word of caution is warranted here: many off-the-shelf tools offer a 'random selection' option that makes it possible to retrieve randomly x instances of the searched item out of the total number of occurrences. While this is a common approach, it may not always be the best solution. Depending on the research question, it may be necessary to be able to statistically control for autocorrelation or priming effects. These notions refer to the fact that often the dependent, or response, variable is not just correlated with a variety of independent, or predictor, variables, but also with previous values of itself. For instance, speakers who have used one of a set of functionally very similar constructions are, all other things being equal or at least very similar, more likely to use that construction again than they would be if they had not used it before. Such effects can be quite strong and predictive on their own: Gries (2016) shows that future choices (*will* vs. *going to* vs. *shall*) can be predicted with more than 80% accuracy just on the basis of the last choice of a speaker, which means that most of the time the sampling unit should not be the individual usage event but the conversation, the (newspaper) article, the corpus file, etc. so that information from previous choices can be accounted for.

The next step is to report on the methods used to retrieve the final set of linguistic item/s. This also requires the author to describe how, for instance, false hits, i.e. instances of something in a corpus that fits the structural description or regular expression used for data retrieval, but that turn out to not actually be instances of the phenomenon in question, were identified. For instance, in a shallowly-parsed corpus, one might search for instances of V NP NP with the goal of retrieving ditransitive constructions such as *He gave him a book*, but that search might also return instances of object complementation such as *He called him a liar*, which would have to be filtered out. Again, each corpus processing tool used at the retrieval stage should be listed, described and properly referred to. Importantly, the exact search expression should be reported and the settings used should be specified (e.g. list of word separators, minimum frequency or dispersion threshold for word lists; cf. Part I.II for the settings typically associated with different corpus methods). If a programming language was used, exact search expression (in particular more complex regular expressions) should always reported; depending on the complexity of all analytical procedures, even providing pseudocode can help readers comprehend the research reported on better.

The 'Methods' section of a corpus-linguistic paper also needs to contain information on how the notions/concepts considered relevant for the analysis were operationalized as variables and how annotations were added with regard to the variables that may affect the linguistic phenomenon under study. Following Wilkinson et al. (1999:595), a "variable is a method for assigning to a set of observations a value from a set of possible outcomes." For example, a variable called "NP length" might assign to each NP one value quantifying it length. But if one considers the length of constituents (e.g. for a study of a syntactic alternation), the readers needs to be told how constituents' lengths were measured: actual time in ms (from an audio/video corpus), length in characters, phonemes, syllables, morphemes, words, phrases, … If one considers the animacy of referents of constituents for the same study, the readers needs to be

told how many and which different levels of animacy were distinguished: two (animate vs. inanimate), three (human, non-human animate, inanimate), etc.

The result of any annotation process should virtually always be a spreadsheet in the so-called case-by-variable, or long, format in which

- every row is one case, i.e. measurement of the dependent variable under investigation;
- every column is one variable – independent variable or otherwise – for which each case was annotated. (See Gries 2013: Sect. 1.3.3 for additional discussion).

This is because, as repeatedly mentioned in the chapters on statistical testing (Part II), most statistical tests are easiest done on data in this format.

In many cases, it might be necessary to explain in the 'Methods' section how elements were treated whose classification is not obvious: If one measures the length of a subject in characters, are spaces or punctuation marks included? If one distinguished the above three animacy levels, how was *God* classified or *This virus* or *The cabinet*? Things can be even more complicated when the notion to be explored is harder to operationalize. For instance, the degree of givenness, or cognitive accessibility, of the referent of an NP is a graded notion and can be variously (and only imperfectly) operationalized via, for example, the number of times the referent has been mentioned, if at all, in the preceding 10 or 20 clauses or by the distance to the last mention of the referent, if any, in the same preceding context. But even then one has to consider tricky questions such as whether the word *flower* is an antecedent for the word *rose* (because it is a superordinate term and therefore arguably evokes *rose* to at least some degree), or whether the word *car* is an antecedent for the word *tire* (because a car is a whole of which a tire is a part), etc.[2] If multiple annotators are involved, one provide at least an indication of interrater reliability and/or how differences in annotation decisions were resolved (cf. Spooren & Degand 2010; Fuoli & Hommerberg 2015). If only one annotator was involved, it is still recommended to maintain a coding book/logs and report an intra-rater reliability score, i.e. a score that measures the reliability of the coding by a single researcher based on the repeated coding of the same set of data at a later time (cf. Loewen & Plonsky 2015:164). Although this is still too rarely often done in the field (including by us), reporting intra-rater reliability would appear good practice given that linguistic data are typically annotated by just one researcher, especially in M.A. and Ph.D. dissertations. The methods section is also the place to mention whether any instrument/material/coding scheme developed for the purposes of the study has been made publicly available (either in the form of an appendix to the article or uploaded on the author's website or onto an online repository such as IRIS, i.e. a collection of instruments, materials, stimuli, data coding and analysis tools used for research into second languages (Marsden et al. 2016, https://www.iris-database.org).

Often, the next kind of information the reader needs to learn about is some descriptive statistics of the data of the type discussed in Chap. 17. This might involve frequency tables of categorical variables as well as box plots and/or ecdf plots for numeric variables. The purpose of these descriptive summaries is that readers get a better overview of the data (including information about missing or unmeasurable/unclassifiable data: how many such cases there were, how they were dealt with, etc.). This also means that care has to be taken to make sure the right kinds of statistics are reported because even descriptive statistics sometimes come with

[2] See Gries (in press) for more information about how to carry out the tasks of retrieval and annotation discussed above.

some assumptions that need to be borne in mind: For instance, (i) it does not make much sense to report one overall mean for a Zipfian or a bimodal distribution of a numeric variable and (ii) it does not make sense to report any measure of central tendency without a measure of dispersion. For categorical dependent variables, frequencies/percentages should be reported (as they constitute the baseline against which any models will be evaluated).

Also, it is often helpful to discuss what, if any, other kinds of exploratory steps were undertaken and what, if any consequences they had for the analysis subsequently reported. For instance, in many studies, numeric variables have to be transformed to make them more 'well-behaved' in a subsequent statistical analysis so readers need to know which transformations were applied (logging, square-root, inverse, centering, *z*-standardizing, logit, etc.), why they were applied, how outliers were dealt with and so on. In the cases of categorical variables, readers should be told if certain categories that were distinguished at an earlier stage were then conflated for conceptual/theoretical or statistical reasons (e.g., when one or more categories are so rare that their rarity would cause problems for subsequent statistical analyses); Zuur, Ieno, & Elphick (2010) provide a nice overview of many such exploratory steps.

If more than just descriptive statistics are computed, i.e. statistics of the kinds discussed in Chap. 20 to 25, then it is necessary to discuss how it was made sure that the data meet the assumptions of the method that was ultimately employed: If a chi-squared test for independence was computed, were all data points independent of each other and were the expected frequencies large enough? If a *t*-test for independent samples was computed, were the data checked for normality and what was the result? If a regression model with multiple predictors was computed, how was collinearity diagnosed (and addressed)? See Chap. 20 and following for more info about the assumptions of statistical tests. Also, the reader needs to get a precise explanation of all the often many steps of the statistical analysis. For example, for regression modeling,

- did the analysis involve fitting and testing just a single model? If so, what was that model and why did it look the way it did – i.e., how did it test which hypotheses?
- if the analysis involved fitting multiple models, was that a stepwise model selection process or a model amalgamation process? If it was the former, what was the direction of the selection process (forwards, backwards, hybrid) and which criterion was used (p, $AIC_{(c)}$, BIC, …)?
- did analyses have to be redone or changed because of problems emerging during the analysis or from initial results? For instance, did the analysis reveal that 0.5% of the data exhibit a degree of leverage on the results that distorted the general trend so it was decided to re-do everything without these 0.5% of the data?

It is also important to explain how many tests were performed on one and the same data set to test which/how many hypotheses and which, if any, corrections for multiple testing were employed.

As is obvious from the above, a lot of corpus-linguistic work does not discuss all their methodological aspects in sufficient detail, but in order to at least begin to approach the ideal of reproducibility, it is essential that all this information be provided at a sufficient level of detail. This also applies to the results, whose presentation we discuss in the next section.

26.3 The 'Results' Section

If good 'Methods' sections help ensure reproducibility, good 'Results' sections ensure comprehensibility and go a long way towards making a reader accept one's conclusions. If only simple descriptive overview statistics (Chap. 17) are computed, then they might be all that is required for a results section, but chances are that more than such overview statistics are computed. In such cases, quite a few kinds of results are required. In the case of monofactorial statistics of the kind discussed in Chap. 20 or the regression modeling approach discussed in Chap. 21 and Chap. 22, it is usually a good idea to begin with some overall numeric summary statistics. For many monofactorial tests, these would be overall test statistics, degrees of freedom, and one or more *p*-values; for many regression models or similar kinds of predictive models or classifiers, these would be the overall significance test (e.g., an *F*- or a G^2-test with their degrees of freedom and *p*), overall (adjusted) R^2-values, and, for methods involving categorical response variables, precision/recall/accuracy statistics on either the modeled data or, even better, data from cross-validation (see, e.g., Kuhn & Johnson 2013:20-26, 69-71). In addition, readers should be presented with the following kinds of statistics for regression models or equivalents of those for other statistical methods:

- significance values for each variable (often, these result from comparing a model with a predictor in question against one without it);
- coefficients and significance tests for each predictor in the model (often, these reflect the change in prediction from a reference level to another treatment level or a planned contrast. Ideally, the contrasts that are represented by a coefficient are also provided to the reader for each such coefficient as exemplified in Table 26.2 for an independent variable called Animacy with three levels (*human*, the reference level, *animate*, and *inanimate*) so as to make understanding regression results easier especially for complex models with (many)categorical predictors(with many levels).

Table 26.2 Reporting regression coefficients

Predictor	coefficient	*Se*	*t*	*P*
$\text{Animacy}_{\text{human} \rightarrow \text{animate}}$	0.272	0.111	2.45	0.171
$\text{Animacy}_{\text{human} \rightarrow \text{inanimate}}$	3.142	0.577	5.445	0.017

(One should use reasonable numbers of decimals: just because R can provide 10 does not mean the readers needs 10 …). Given that regression coefficients correspond to raw/unstandardized effect sizes, it follows that for any other kinds of test – mono- or multifactorial – the actual effects should be provided: the differences in means/medians, the slopes of a correlation, etc. Ideally, these effects should come with confidence intervals/bands or some other indication of their certainty (see the various chapters in Part II for more info about reporting specific statistics) so as to paint a clearer picture of the reliability/robustness of the numerical point estimates.

Finally, the more complex the statistical analysis, the more important it is to provide a proper visualization of the results; the purpose of visualization is to represent/explain what would be harder to represent/explain in prose. That means, one does not need a bar plot of two percentages: this is a statistical result simple enough to not require visualization. On the other

hand, the numerical results of a multifactorial multinomial regression model are likely to be virtually incomprehensible without any visual aids. Visualization comes with its own set of guidelines, the maybe most important of which involves the notion that a graph should contain all the information it aims to present but no more. In this regard, the following plot in Fig. 26.1 is lacking.

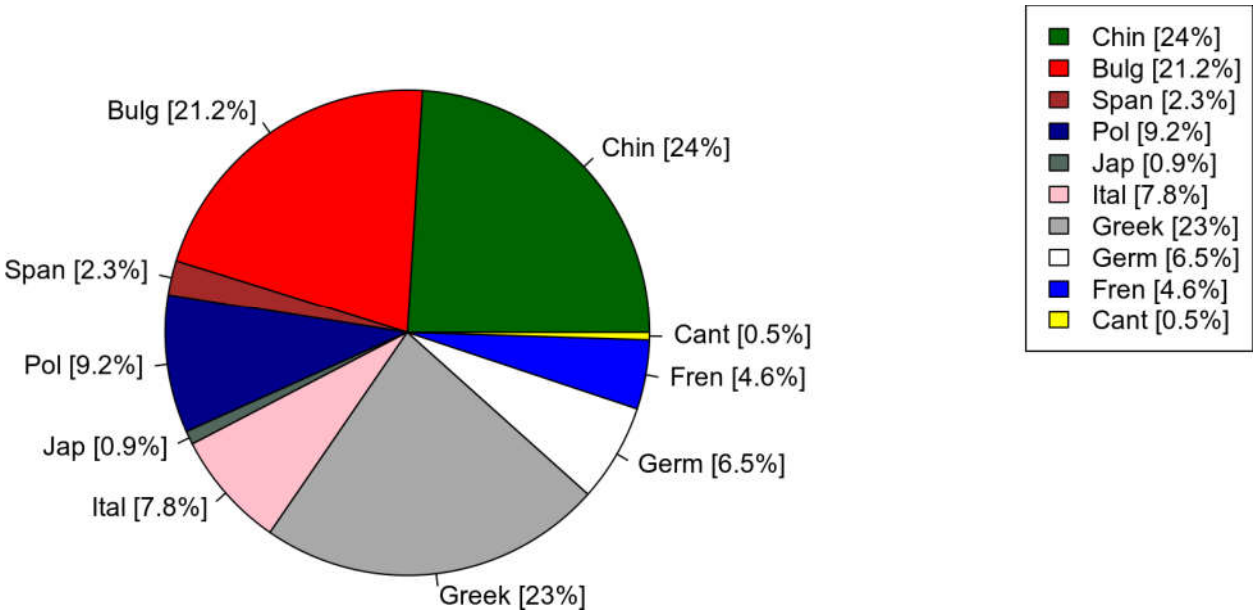

Fig. 26.1 Sub-optimal representation of percentage data

The data that this plot is supposed to represent is nothing more than ten percentages adding up to 100%, i.e. one-dimensional vector/sequence of 10 numbers. However, the original graph that Fig. 26.1 is reproducing, which was in fact a three-dimensional version of Fig. 26.1, represented these data with an extremely bad data-ink ratio:

- the original chart utilized four dimensions (a three-dimensional chart plus different colors) and a non-informative background-shading effect;
- the percentages are represented in a pie chart although humans are much worse at comparing information in angles than they are at comparing locations of points or lengths of lines (see Cleveland & McGill 1985, Tufte 2001:178);
- Figure 26.1 and its original version are extremely redundant, given that the countries and their percentages are listed around the pie and again in a legend.

In other words, this graph contains much more visual 'information' than is merited by the actual data it is supposed to represent and the way that visual 'information' is provided is redundant and does not go well with how humans perceive data well. In terms of data-ink ratio and any other principle of data visualization, Fig. 26.2 is a better representation: Each percentage (i.e., one-dimensional data point) is represented as a point on a line (i.e., a one-dimensional geometric construct), and no redundant information detracts from the message:

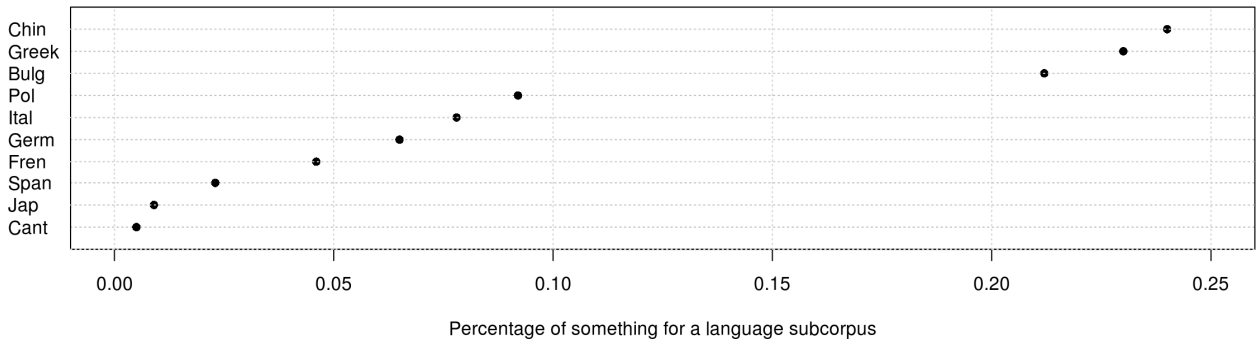

Fig. 26.2 Better representation of percentage data

In the case of regression modeling, the probably most informative way to present results is to have effects plots (cf. Chap. 21; see also Fox 2003 and Fox & Hong 2009) for every predictor to be discussed, which show predictions on the *y*-axis against predictors on the *x*-axis (and maybe with different kinds of points and lines) together with confidence intervals, Fig. 26.3 is an example from a study on *that*-complementation (Wulff, Gries, & Lester 2018), i.e. the question of whether learners of English say *I thought the Borg assimilate other species* or *I thought that the Borg assimilate other species*. The plot shows one effect from a regression analysis on how similar the learner choices are to imputed native-speaker choices; that is, absolute values of the dependent variable Deviation on the *y*-axis indicate how much a learner choice differs from a native speaker choice; the plot represents the effect of an interaction between the length of the complement subject (here, *the Borg*, on the *x*-axis) and the register/mode (speaking vs. writing, represented in colors and with *s*/*w* respectively). More specifically, it shows how the effect of the length of the complement subject is different between speaking and writing: in speaking, there is essentially no effect (because the regression line is horizontal), but in writing there is an effect such that with increasing subject length, deviation scores approach 0 (i.e. the learner choices become more nativelike). The plot contains (raw) effect sizes (the slope of the regression lines), their uncertainty (the confidence bands), labels for everything important (*x*-axis, *y*-axis, positive and negative *y*-value labels etc.), and as a visual representation of fit, the actually observed data points are also included as grey circles. Note that, since this is an effects plot, the effect shown – the interaction of complement subject length and register/mode – is represented *while every other effect in the regression model is controlled for*, which is important because the frequently used plots of *observed* means/correlations do not do that.

Similar recommendations hold for similar kinds of classifiers such as trees and forests (Chap. 25) and other machine-learning algorithms. For some other methods, the resulting visualization might actually be the main result, as in cluster analyses (see Chap. 18) or classification trees, but there, too, it is important to be aware of the data-ink ratio and present everything that is required, but no more.

Lastly, it is typically a good idea to provide some information on the validity of the results. For instance, regression models, but also many other statistical methods, are based on assumptions regarding the data, which means it is important to tell readers that these assumptions were checked (in a process called model validation or diagnostics). This part usually does not need to be long, but, for instance, providing the information that one's model did not suffer from collinearity, overfitting, heteroscedasticity, etc. (see Chap. 22 to 25 for these notions) is strongly encouraged.

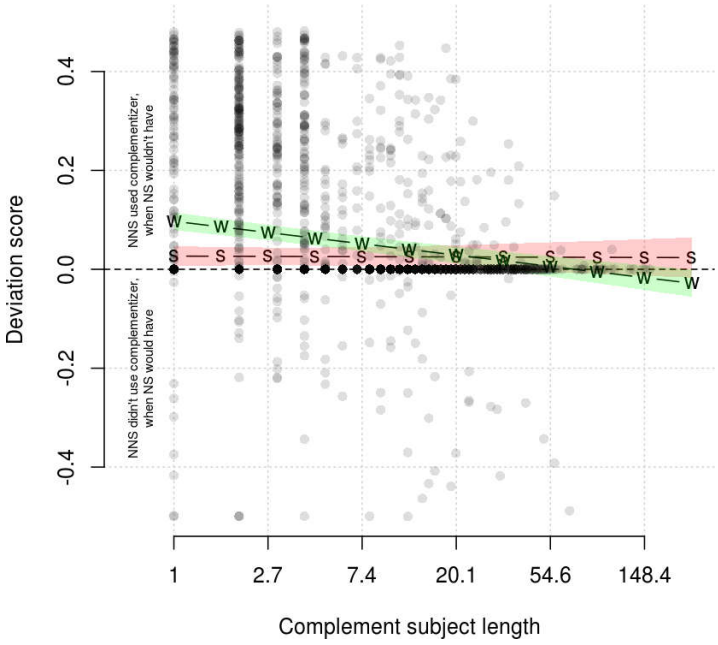

Fig. 26.3 Example of an effects plot

26.4 Concluding Remarks

The level of details we advocate for in this chapter may seem a bit daunting at first. It is however essential for at least the following reasons. First, a complete description of the study design (from data collection to data analysis) enables the reader to evaluate the appropriateness of the data and methods used for answering the research questions as well as the reliability and validity of the results. Second, and as already mentioned in the chapter, detailed information on the data and how it was analyzed is also a prerequisite for reproducibility and replicability. Third, it is only by paying more attention to methodology that the field of corpus linguistics will answer repeated calls for developments in study quality, i.e. "the combination of (a) adherence to standards of contextually appropriate methodological rigor in research practices and (b) transparent and complete reporting of such practices" (Plonsky 2013:657). Scholars have observed a significant number of weaknesses related to sampling practices, data analyses and reporting practices in corpus linguistics. Paquot & Plonsky (2017), for example, provided the first empirical assessment of quantitative research methods and study quality in learner corpus research and reported high rates of both underreported and missing data. Fourth, improving reporting practices will also permit meta-analysts to conduct comprehensive and empirically grounded reviews of previous research, a practice that has only sparsely been adopted in the field of corpus linguistics (see Chap. 27).

References

American Psychological Association. 2010. *Publication Manual of the American Psychological Association* (6th edition). Washington, DC: American Psychological Association.

Berez-Kroeker, A., Gawne, L., Kung, S., et al. 2017. Reproducible research in linguistics: A

position statement on data citation and attribution in our field. *Linguistics* 56(1):1-18.
BNC Consortium. 2001. The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/. Accessed 30 August 2019.
Branco, A., Cohen, K. B., Vossen, P., Ide, N., and Calzolari, N. 2017. Replicability and reproducibility of research results for human language technology : Introducing an LRE special section. *Language Resources and Evaluation*, 51(1), 1-5.
Cleveland, W., and McGill, R. 1985. Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716):828-833.
Fox, J. 2003. Effect displays in R for generalised linear models. *Journal of Statistical Software* 8(15):1-27.
Fox, J., and Hong, J. 2009. Effect displays in R for multinomial and proportional-odds logit models: extensions to the effects Package. *Journal of Statistical Software* 32(1):1-24.
Fuoli, M., and Hommerberg, C. 2015. Optimising transparency, reliability and replicability: annotation principles and inter-coder agreement in the quantification of evaluation expressions. *Corpora* 10(3):315-349.
Gries, S.T. 2013. *Statistics for linguistics with R*. 2nd rev. & ext. ed. Boston & New York: De Gruyter Mouton.
Gries, S.T. 2016. Variationist analysis: variability due to random effects and autocorrelation. In *Triangulating methodological approaches in corpus linguistic research*, eds. Baker, P., and Egbert, J.A., 108-123. New York: Routledge, Taylor and Francis.
Gries, S.T. In press. Managing synchronic corpus data with the British National Corpus (BNC). In *MIT Open Handbook of Linguistic Data Management*, eds. Berez-Kroeker, A.L., McDonnell, B., Koller, E., and Collister, L. Cambridge, MA: The MIT Press.
Kuhn, M., and Johnson, K. 2013. *Applied predictive modeling*. Berlin & New York: Springer.
Loewen, S., and Plonsky, L. 2015. *An A-Z of applied linguistics research methods*. New York, NY: Palgrave.
Marsden, E., Mackey A., and Plonsky, L. 2016. The IRIS Repository: Advancing research practice and methodology. In *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages*, eds. Mackey, A., and Marsden, E., 1-21. New York: Routledge.
Paquot, M., and Plonsky, L. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3(1):61-94.
Plonsky, L. 2013. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition* 35(4):655–687.
Porte, G. 2012. *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.
Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
Spooren, W., and Degand, L. 2010. 'Coding coherence relations: reliability and validity'. *Corpus Linguistics and Linguistic Theory* 6(2):241-266.
Tufte, E. 2001. *The visual display of quantitative information*. 2nd ed. Graphics Press: Cheshire, CT.
Wilkinson, L., and The Task Force on Statistical Inference. 1999. Statistical methods in psychology journals. *American Psychologist* 54(8):594-604.

Wulff, S., Gries, S.T., and Lester, N.A. 2018. Optional *that* in complementation by German and Spanish learners: where and how German and Spanish learners differ from native speakers. In *What does Applied Cognitive Linguistics look like? Answers from the L2 classroom and SLA studies*, eds. Tyler, A., Huan, L., and Jan, H., 97-118. Berlin & Boston: De Gruyter Mouton.

Zuur, A.F., Ieno, E.N., and Elphick, C.S. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1):3-14.