MATERIALS & METHODS REPORT

Inter-rater reliability in Learner Corpus Research

Insights from a collaborative study on adverb placement

Tove Larsson[1], Magali Paquot[2,3] and Luke Plonsky[4]
[1] Uppsala University | [2] FNRS | [3] UCLouvain | [4] Northern Arizona University

In Learner Corpus Research (LCR), a common source of errors stems from manual coding and annotation of linguistic features. To estimate the amount of error present in a coded dataset, coefficients of inter-rater reliability are used. However, despite the importance of reliability and internal consistency for validity and, by extension, study quality, interpretability and generalizability, it is surprisingly uncommon for studies in the field of LCR to report on such reliability coefficients. In this Methods Report, we use a recent collaborative research project to illustrate the pertinence of considering inter-rater reliability. In doing so, we hope to initiate methodological discussion on instrument design, piloting and evaluation. We also suggest some ways forward to encourage increased transparency in reporting practices.

Keywords: inter-rater reliability, coding errors, reporting practices, study quality, Fleiss' kappa

1. Introduction

There is error in all we measure. Inconsistencies in manual coding and inaccuracies in measurement introduce a threat to the internal validity of our research by obscuring the signal that we seek to detect in our data (Plonsky & Derrick, 2016). To be clear, the kind of error referred to here is not the kind that is produced by learners of a language (e.g. developmental errors such as incorrect subject-verb agreement). While researchers in the field of Learner Corpus Research (LCR) often come across such learner errors, the present report deals

https://doi.org/10.1075/ijlcr.20001.lar | Published online: 10 December 2020
International Journal of Learner Corpus Research 6:2 (2020), pp. 237–251. issn 2215-1478 | e-issn 2215-1486
Available under the CC BY-NC 4.0 license. © John Benjamins Publishing Company

with a very different kind of error, one that has received far less attention in LCR, namely measurement error. Specifically, this report discusses measurement error of the kind that stems from the manual coding and annotating of linguistic features that we, as fallible human researchers, conduct (i.e. error resulting from a coder miscoding a value).[1]

Manual coding is commonly used in LCR and neighboring subfields to classify tokens into categories. One example of categories that are often coded manually is predictors of particle placement (e.g. TYPE OF DIRECT OBJECT, ANIMACY and SEMANTICS OF THE VERB; e.g. Paquot, Grafmiller, & Szmrecsanyi, 2019). Another example is semantic classifications of constructions (e.g. subject extraposition), using semantic categories such as 'Hedges' and 'Attitude markers' (e.g. Hedge: *it seems that*; Attitude marker: *it is interesting that*; e.g. Larsson, 2018). Manual coding might be compromised for a variety of reasons, both systematic (e.g. due to ambiguity of the coding scheme, inadequate coder expertise or training and/or coder bias) and random (e.g. due to coder fatigue and/or typing errors); all of these will introduce inaccuracies into our results.

Regardless of the source of error, it is necessary to examine and estimate the amount of error present in a coded dataset (Révész, 2012). The degree of consistency for a single coder and the degree of agreement between coders can be assessed through coefficients of intra-rater reliability and inter-rater reliability, respectively. Intra-rater reliability refers to the extent to which a coder consistently assigns categories in the same dataset on a different occasion. Inter-rater reliability is concerned with the degree to which two or more coders arrive at the same categorizations when analyzing the same data with the same coding scheme and procedure (e.g. Loewen & Plonsky, 2015). These two types of reliability are distinct from 'internal consistency' (also called 'instrument reliability'), which refers to the extent to which a set of items in a psychometric scale are correlated (see McKay & Plonsky, in press). The critical distinction between internal consistency and intra/inter-rater reliabilities lies in the source of error; the former attributes inconsistencies to the items themselves and the scale to which they belong, whereas the latter estimates human error (for examples and related discussion, see Kutuk, Putwain, Kaye, & Garrett, 2019, and Morgan, Zhu, Johnson, & Hodge, 2014, respectively).

---

1. As the present report focuses on measurement error resulting from manual coding (rather than semi or fully automatic coding) of a linguistic category (in our case adverb placement), the report will not cover assessment of measurement error resulting from automated taggers. We refer readers who are interested in those types of measurement error to studies such as Lu (2010) and Rosen, Hana, Stindlova, & Feldman (2014).

Given the status of reliability as a pre-requisite for validity and, therefore, study quality, interpretability and generalizability (Plonsky, 2013; Purpura, Brown, & Schoonen, 2015), it is unfortunate, to say the least, that there is no established tradition of reporting or interpreting reliability coefficients in LCR. One area where there is some discussion of annotation reliability is Computer-aided Error Analysis (CEA) (cf. Andreu-Andrés, Astor-Guardiola, Boquera-Matarredona, Macdonald, Montero-Fleta, & Pérez-Sabater, 2010; Díez-Bedmar, 2015; Viyatkina, 2016) but, to the best of our knowledge, even in CEA studies, the reporting and interpretation of inter-rater reliability coefficients, as well as the discussion of resolution methods, still remain the exception rather than the rule. While such estimates are not relevant to all types of studies, Paquot and Plonsky (2017) nevertheless expressed concern at the exceedingly small portion of learner corpus studies that measured and reported inter-rater reliability coefficients: 11% (K=378).[2] We hasten to add, though, that rates of reporting were also found to be increasing in the sample over time.

The main objectives of this report are to stress the pertinence of reliability as a prerequisite for internal validity and to offer some suggestions for how to move forward. In order to do so, we illustrate the importance of considering (inter-rater) reliability and we report on some of the methodological steps taken in a recent collaborative research project (Larsson, Callies, Hasselgård, Laso, van Vuuren, Verdaguer, & Paquot, 2020), as detailed below. In sharing insights from this project, we hope to initiate methodological discussion on instrument design, piloting and evaluation, and to encourage increased transparency in reporting practices in LCR.

2. Working towards increased reliability in a study on adverb placement

This section reports on relevant methodological aspects from a larger collaborative project between seven researchers who participate in the *Varieties of English for Specific Purposes Database* (VESPA) corpus collection initiative (Paquot, Hasselgård, & Oksefjell Ebeling, 2013). Larsson et al.'s (2020) study provides an exploratory account of adverb placement in the spoken and written production of native speakers and learners of English with different first-language (L1) backgrounds. In an attempt to revisit and then move beyond accounts of

2. The same troubling conclusion has been drawn in other corpus-based domains of enquiry such as discourse studies (Spooren & Degand, 2010). See also Polio and Shea (2014) for a discussion of inter-rater reliability applied to linguistic accuracy in second language writing research.

syntactic L1 transfer common to studies of the positional distribution of adverbs, the study included both extralinguistic (e.g. register, the students' L1) and linguistic factors (e.g. clause type, number of auxiliaries) with the aim of investigating their relative association with the distributional tendencies of adverbs. For the full account of this study and its results, readers are referred to Larsson et al. (2020).

In the present report, however, only the coding scheme that provides the foundation for the other analyses reported on in Larsson et al. (2020) will be discussed. For this part, all the data (a total of 12,814 adverbs) were manually classified into positional categories.[3] Specifically, we extracted the concordance lines and coded the data in Excel. The initial plan was for each author to code the data from their own subcorpus (i.e. the adverbs produced by learners with the same L1 as the coder). However, this approach immediately raised issues related to reliability: how were we to ensure that the different datasets, each to be coded by a different researcher, would be coded in the same way? Repeatedly being forced to address questions pertaining to reliability served as an eye-opener, and in sharing this experience, we hope to encourage researchers in the field to consider these issues.

In Sections 2.1–2.4, we report on the different steps taken to address these issues, from piloting the coding scheme on a small dataset that was coded and discussed by the seven collaborators, to revising the instrument and estimating inter-rater reliability (IRR) at different points of the coding procedure. For ease of reporting (and because including more variables would not enrich the discussion), we focus on the coding of just one variable, namely ADVERB POSITION.

2.1 The coding scheme

To get a better understanding of what the distributional tendencies of adverbs are in the data, the first step was to see where in the clause they occurred. For this purpose, a slightly adapted version of Hasselgård's (2010) coding scheme (based on Quirk, Greenbaum, Leech, & Svartvik, 1985: 490ff) was used. Six main adverb positions were coded for: clause-initial (I), clause-medial 1 (M1), clause-medial 2 (M2), clause-medial 3 (M3), indeterminate intital/medial (IM) and end (E) position. In addition to these main syntactic positions, we added one category for

3. While semi-automatic classification was considered for the adverb project, we decided against it, given in particular the somewhat messy nature of the spoken data (e.g. false starts, pauses marked by periods in the transcriptions); see Larsson et al. (2020) for a more detailed discussion.

tokens to be excluded (NA). Definitions and examples of these categories can be found below.

The first positional category, I, comprises any tokens where the adverb (bolded in the examples) comes before any obligatory element in the clause, as in (1), which most often means "a position before the subject, or before the verb in cases of S–V inversion or subject ellipsis" (Hasselgård, 2010: 42). The first of the three medial categories coded for, M1, is the position "between the subject and any part of the verb phrase", as shown in (2); the second one, M2, is the position "after the (first) auxiliary, but before the main verb", as in Example (3), and the third one, M3, is the position "between the verb phrase and some other obligatory element, viz. an object, a predicative, or an obligatory adverbial" exemplified in (4) (Hasselgård, 2010: 42). In some cases, for example when the subject is omitted, as in (5), it is not possible to distinguish between the clause-initial and clause-medial position; these instances make up the category IM. The final position, E, is the position "following all obligatory elements", see (6) (Hasselgård, 2010: 42).

(1) **Of course**, the *'s* is not repeated. (UCL0035-LING-01)

(2) [...] it is possible that some students **really** do have an unusually good understanding of the language. (UPP0133-LING-01)

(3) [...] a lexeme belonging to one class can **simply** be "converted" to another [...]. (UCL0017-LING-01)

(4) [...] one of those was not **really** part of the discourse of turn taking [...]. (STO0023-LING-01)

(5) [...] words will always develop new senses, or **simply** become obsolete. (UPP0002-LING-01)

(6) [...] they also had snow **actually** [...]. (SW040)

Tokens in the written data that were part of linguistic examples (so-called 'mentioned items') or quoted speech were excluded, along with tokens in the spoken data that occurred in false starts; these made up the NA category. In addition, as we define the position of the adverb relative to the verb phrase (cf. Hasselgård, 2010: 40), any instances where the adverb was found inside a noun phrase or where it modifies an adjective were excluded. Examples can be found in (7) and (8).

(7) [...] thank God cos I don't know I'd **probably** (erm) I don't know I'd probably start giggling or something like that [...]. (SW034)

(8) [...] the Jerry Springer show is **really** famous [...]. (UCL1306-LING-01)

Although learner errors (i.e. developmental errors) were not the focus of Larsson et al. (2020), we note that the very few unorthodox uses of adverbs found in the learner data (e.g. (9)) did not pose a problem for the coding scheme, as we took a descriptive approach (i.e. in this case, the adverb comes in-between the verb phrase and some other obligatory element in the clause and is thus coded as M3, even though the token technically includes a learner error). These tokens were subsequently analyzed manually in relation to our research questions.

(9) [T]here have been **apparently** initial observations as well as conceptual and theoretical considerations [...]. (GE_rpa1.g.fr.031)

2.2 Piloting the coding scheme and estimating inter-rater reliability

In order to assess the reliability of the coding scheme, a random sample of 100 tokens (50 tokens from the written data and 50 tokens from the spoken data) was sent to all seven coders at the beginning of the project. The initial approach was intuitive and straightforward: We estimated reliability across the team by calculating the overall percentage of cases on which all coders agreed. Doing so revealed a somewhat disheartening outcome, at least at first glance, as all seven raters agreed on only 41 percent of the cases.[4] However, a closer look at the results showed that in almost all cases, only one or two raters had diverging views for any given token. In other words, in this case, raw percentages of unanimous agreement appear to underestimate agreement.[5] Based on these and other considerations, percentages are often misleading when estimating inter-rater reliability (cf. Cohen, 1960).

Instead, we employed a commonly used index, Fleiss' kappa (Fleiss, 1971), to test the agreement between the coders (for an example of this index in the context of L2 research, see Collentine, 2009). Fleiss' kappa is one of several coefficients that can be used to assess the reliability of agreement; others include Cohen's kappa, the use of correlation coefficients, such as Kendall's τ, and the intra-class correlation coefficient (ICC)[6] (Hallgren, 2012). Somewhat simplified, Fleiss' kappa assesses inter-rater reliability by calculating the extent to which there

4. In the present project, we opted for a conservative approach where only tokens for which all seven raters were in agreement were considered having reached "full agreement". Another, less ambitious (and arguably a somewhat less rigorous) approach would have been to use a cut-off point for what is considered "full agreement", such as a majority vote (4 out of 7 raters, etc.).

5. It should be noted, however, that there are also cases where percentages tend to overestimate agreement, such as when there are fewer raters and only two or three categories to code. In addition, percentages "do not correct for agreements that would be expected by chance and therefore overestimate the level of agreement" (Hallgren, 2012: 25).

6. ICC is the statistic most often used for ordinal, interval and ratio variables (Hallgren, 2012: 29).

is agreement between raters divided by that which could be expected by chance (Fleiss, 1971). By doing so, the researcher avoids presenting an overly optimistic view of agreement among raters. The number of categories included also has an effect on the kappa value (Sim & Wright, 2005: 264). That is, kappa accounts for the fact that chance agreement is less likely when raters are forced to choose among a larger number of categories for each item (i.e. token) to be rated. This index differs from the commonly used measure Cohen's kappa in that Fleiss' kappa can be applied to categorical data assessed by more than two raters.

The kappa values range from 0 to 1, where 0 denotes "agreement no better than that expected by chance, as if the raters had simply 'guessed' every rating" (Sim & Wright, 2005: 259). While there is some disagreement in the literature with regard to what constitutes sufficiently high scores (cf., e.g., Sim & Wright, 2005), we have used the commonly-used scale created by Landis and Koch (1977), where scores between 0.41–0.60 are considered to suggest "moderate agreement", 0.61–0,80 "substantial agreement" and 0.81–1.00 "almost perfect agreement". The software environment *R* (R Core Team, 2018) and the R package *irr* (Gamer, Lemon, Fellows, & Singh, 2012) were used to carry out the test. When applied to our initial 100-token sample, we obtained a kappa score of 0.64 ($z = 61.5$), suggesting "substantial agreement" in Landis and Koch's terminology (see also a set of field-specific benchmarks for interpreting reliability estimates in Plonsky & Derrick, 2016). While the results were far from disastrous, we believed that it would be preferable to try to improve them.

2.3 Revising the coding scheme

In an attempt to increase the IRR and, thus, simultaneously both reduce the error in our data and increase the reliability of our results, we used the outcome of this first test to improve and fine-tune the coding scheme. The tokens for which we had not reached full agreement helped highlight decisions that had not been described in sufficient detail in the coding scheme. Two examples of such improvements pertain to the treatment of split infinitives (10) and subject extraposition (11). While there are several possible ways of treating such instances, we made the following decisions: for split infinitives, the *to*-clause should be treated as a clause with a null subject, which means that the adverb will be classified as M1 (as in (10)). For subject extraposition, the introductory subject *it* is to be treated as the subject, which means that the token will be categorized as M1, M2, M3 – or E as in (11).

(10) [...] it was nice to to **actually** see it [...]. (SW010)

(11) [...] so it's better **maybe** to have it done by an (er) somebody who knows you [...]. (SW040)

We also added a category for tokens that the coder was not sure how to treat due to, for example, structural ambiguity (the "I do not know" category: IDK). Tokens that were coded as IDK were revisited and discussed with the other coders; this category will not be addressed further here. For more details about the coding, we refer readers to the IRIS database (https://www.iris-database.org) where the final coding scheme is available.

2.4 From a single-coder to a double-coder approach

Once the coding scheme had been adjusted and improved, a second IRR test was carried out. The results of this test were better: the kappa score had now increased from 0.64 to 0.79 ($z = 74.7$), meaning that our results were at the upper end of what Landis and Koch (1977) would refer to as "substantial agreement" (0.81 being the cut-off point for "almost perfect agreement"). An overview of the kappa scores for the two IRR tests plotted on Landis and Koch's (1977) scale is provided in Figure 1.

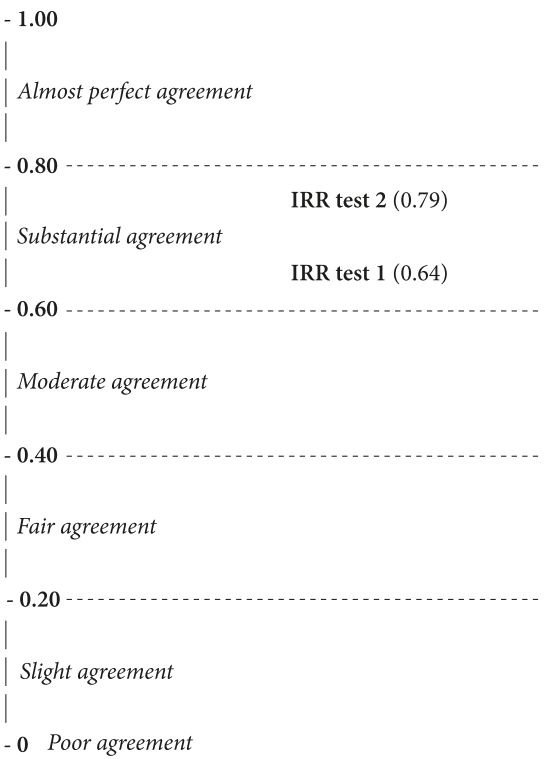

Figure 1. The kappa score for the IRR tests plotted on Landis and Koch's (1977) scale

This estimate was also quite a bit closer to the norms for inter-rater agreement found in L2 research. We base this claim on the results of Plonsky and Derrick (2016) who aggregated (i.e. meta-analyzed) 2,244 reliability estimates from a sample of 537 studies. Their results showed a median IRR for a closely related index, Cohen's kappa, of 0.87 with the 25th percentile approximately equal to the estimate observed in Larsson et al. (2020).

The remaining disagreement after the second IRR mainly resulted from different value judgments of structurally unclear tokens in the spoken data; an example can be found in (12) where some coders had interpreted this instance of *actually* to be a false start (which would result in the token being coded as NA), whereas others had taken this to be clause initial (I).

(12) [...] it's too complicated and (er) well **actually** (erm) .. what really: impressed me in in China is that they are . well of course the number of people but everybody knows about that [...]. (FR040)

While the level of agreement for the second IRR test was considerably higher than that of the first one, the remaining disagreement led us to implement some further resolution methods to improve reliability of our coding to as great an extent as possible. The first thing we did was to change from a single-coder approach to an independent double-coder approach (Spooren & Degand, 2010; see also Rose & MacWinney, 2014, on a version of this approach developed for research in phonology). The new approach meant that each L1 subset was coded by two independent raters: an expert for the respective L1s from the learner groups examined and a research assistant (a native-speaker MA student majoring in linguistics who was trained by the first author). The coding was subsequently checked for consistency by the first author; discrepancies were resolved with the help of the coding scheme, Hasselgård (2010), Quirk et al. (1985) and in discussion with the collaborators.

While time-consuming, there are several advantages to double coding, the most obvious one being that the number of coding errors is reduced (see Spooren & Degand, 2010). Other advantages include making more explicit the decisions made while coding, thereby increasing the robustness of the categories. However, it is important to note in this context that we do not consider double coding alone as a solution. The IRR tests carried out early on in this project were instrumental to the process, as they not only highlighted problematic tokens and categories but also prompted discussions between all the coders (and not just two coders at a time).

At the outset of the project, adverb placement appeared to be a relatively straightforward topic in that the categorization was based on a well-defined framework. Nonetheless, it was surprisingly difficult to reach agreement for the

categories investigated. Had we not carried out the steps outlined above, we would have lived in blissful ignorance, believing that our coding scheme was sufficiently detailed from the start to ensure high reliability of the results. The procedure reported on above involved critical examination and re-examination of every variable and category used to code the data, which, while time-consuming, ultimately led to more robust categories.

3. Conclusion and ways forward

This report aimed to highlight the importance of considering inter-rater reliability in the context of manual annotations in LCR. As illustrated in this paper, checks on reliability at different stages of the research project not only facilitate more transparent reporting of the reliability of the results but also enable researchers to identify inconsistencies in the categorization, which can be used to adjust and improve the coding scheme. Such additional steps take time and effort, to be sure, but are critical to both (a) minimizing error and (b) reporting research in a manner that is transparent with respect to the process and findings (i.e. reproducible).

Although clearly not unproblematic, the categories used in Larsson et al. (2020) had the advantages that they were syntactic (rather than functional) and came from a well-defined, empirically tested framework. However, many studies in LCR involve investigations of operationally more challenging categories. Examples of categories that are notoriously difficult to classify include discourse functions and lexico-grammatical errors, which, despite arguably being more prone to subjective evaluation and/or interpretation than many other surface categories (e.g. Lüdeling & Hirschmann, 2015; Spooren & Degand, 2010), often are coded by a single researcher without thorough reporting of the procedures involved (Paquot & Plonsky, 2017). As readers of such articles, we are then left to assume that the many related threats to internal validity have been sufficiently addressed. In other words, if we assume that reliability is a prerequisite for study validity, this raises concerns for knowledge formation in the field: "[i]f a coding process is significantly affected by random [or systematic] errors of measurement, it follows that it will not be accurate or meaningful, and hence will not allow for valid interpretations to be made" (Révész, 2012).

Looking beyond the present discussion, we would like to stress the importance of transparency in instrument reporting and formulate a few recommendations for the field (including our own future work). While our discussion above focused on the 'coding scheme' as a special type of instrument, the guidelines below also apply to other types of instruments for data collection and analysis

(e.g. a questionnaire, a proficiency test). When submitting a manuscript for publication,[7] we believe that it is essential to do the following:

- Include information about the origins and development of all instruments used: Where do the instruments (e.g. coding schemes, questionnaires, proficiency tests) come from? How were they built and/or modified?
- Describe the piloting phase in a precise manner: How were instruments piloted? Did piloting lead to identification of problematic construct operationalizations, ambiguous coding categories and/or systematic or random errors? How were these issues remedied or otherwise addressed? What resolution methods were used?
- Provide information about the reliability of the instruments used: How was instrument reliability measured? What type of tests or indices were used? How much of the data were double-coded? By whom? Were the coders trained? How so? How should reliability estimates be interpreted? How does the instrument reliability in your study compare to proposed standards in the field (cf. Plonsky & Derrick, 2016, for a discussion)? What effect – if any – might the error in the measures have led to attenuated study effects? (see Osborne, 2003; Trafimow, 2017; and McKay & Plonsky, in press, for discussion in the context of L2 research).
- Make your instruments available to enable other researchers to benefit from them and/or critically evaluate them: There are several options to provide access to instruments today, from journals' websites to repositories such as the IRIS digital repository of instruments and materials for research into second languages (https://www.iris-database.org/).

In proposing this, we join scholars such as Révész (2012); Derrick (2015); Larson-Hall and Plonsky (2015), and Norris, Plonsky, Ross and Schoonen (2015), who have stressed the importance of piloting instruments and reporting reliability estimates in related fields such as language testing and second language research.[8] As we all know, language is complex and the constructs we are interested in are not always easy to operationalize. Nevertheless, in recognizing potential problems

7. Needless to say, these guidelines also presuppose some good practices in terms of study design.

8. For examples of best practice, we refer readers to Révész (2012) for practical information about how to code second language data validly and reliably; Artstein (2017) for more information about inter-annotator agreement measurement in linguistic annotation; Vyatkina (2016) for a CEA study that details the procedure used to check inter-annotator agreement as well as reports an in-depth analysis of annotator disagreements; and Johnson, Penny and Gordon (2010) for a discussion of the effect of different types of resolution methods on final inter-rater reliability.

related to instrument development, categories or the coding process in general, we can work together as a field towards increased reliability of our instruments, thus helping us move towards a better understanding of the intricate nature of second language use and development. Engagement with transparency in reporting practices will also place LCR in a position to discuss field-specific/oriented best practices (cf. Plonsky & Derrick, 2016) and answer key questions such as: What constitutes acceptable (inter-rater) reliability scores for learner corpus data across linguistic domains (e.g. syntactic vs. discourse variables) and modes (spoken, written)? What constitutes a minimum sample size needed to estimate reliability? What linguistic phenomena or constructs can or cannot be coded reliably by a single coder? We are very much looking forward to following discussions about these highly important methodological questions.

Note

Marcus Callies acted as corresponding editor for this material and methods report to avoid any conflict of interest.

Acknowledgements

We are very grateful to the editorial team and the anonymous reviewers for their very helpful comments and suggestions.

References

Andreu-Andrés, M., Astor-Guardiola, A., Boquera-Matarredona, M., Macdonald, P., Montero-Fleta, B., & Pérez-Sabater, C. (2010). Analysing EFL learner output in the MiLC project: An error it's*, but which tag?. In M.C. Campoy-Cubillo, B. Bellés-Fortuño, & M. Ll. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 167–188). London: Continuum.

Artstein, R. (2017). Inter-annotator agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). New York, NY: Springer. https://doi.org/10.1007/978-94-024-0881-2_11

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. https://doi.org/10.1177/001316446002000104

Collentine, K. (2009). Learner use of holistic language units in task-based synchronous computer-mediated communication. *Language Learning & Technology*, 13, 67–87.

Derrick, D. (2015). Instrument reporting practices in second language research. *TESOL Quarterly*, 50(1), 132–153. https://doi.org/10.1002/tesq.217

Díez-Bedmar, M. B. (2015). Dealing with errors in learner corpora to describe, teach and assess EFL writing: Focus on article use. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in Learner Corpus Linguistics: Research and applications for foreign language teaching and assessment* (pp. 37–69). Bern: Peter Lang.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. https://doi.org/10.1037/h0031619

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr*: Various coefficients of interrater reliability and agreement. *R package version 0.84*.

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hasselgård, H. (2010). *Adjunct adverbials in English*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511676253

Johnson, R. L., Penny, J., & Gordon, B. (2010). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, 13(2), 121–138. https://doi.org/10.1207/S15324818AME1302_1

Kutuk, G., Putwain, D. W., Kaye, L., & Garrett, B. (in press). Development and validation of a new multidimensional language class anxiety scale. *Journal of Psychoeducational Assessment*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. https://doi.org/10.2307/2529310

Larsson, T. (2018). Is there a correlation between form and function? A syntactic and functional investigation of the introductory *it* pattern in student writing. *ICAME Journal*, 42(1), 13–40. https://doi.org/10.1515/icame-2018-0003

Larsson, T., Callies, M., Hasselgård, H., Laso, N. J., Van Vuuren, S., Verdaguer, I., & Paquot, M. (2020). Adverb placement in EFL academic writing: Going beyond syntactic transfer. *International Journal of Corpus Linguistics*, 25(2), 155–184. https://doi.org/10.1075/ijcl.19131.lar

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(Suppl. 1), 127–159. https://doi.org/10.1111/lang.12115

Loewen, S., & Plonsky, L. (2015). *An A–Z of applied linguistics research methods*. New York, NY: Palgrave.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007

McKay, T., & Plonsky, L. (in press). Reliability analyses: Estimating error in L2 research. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing*. New York, NY: Routledge.

Morgan, G. B., Zhu, M., Johnson, R. L., & Hodge, K. J. (2014). Interrater reliability estimators commonly used in scoring language assessments: A Monte Carlo investigation of estimator accuracy. *Language Assessment Quarterly*, 11, 304–324. https://doi.org/10.1080/15434303.2014.937486

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, 65(2), 470–476. https://doi.org/10.1111/lang.12104

Osborne, J. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, & Evaluation*, 8(11). Retrieved from https://pareonline.net/getvn.asp?v=8&n=11

Paquot, M., Hasselgård, H., & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of Learner Corpus Research: Looking back, moving ahead. Proceedings of the first Learner Corpus Research Conference (LCR 2011)* (pp. 377–387). Louvain-la-Neuve: Presses Universitaires de Louvain.

Paquot, M., Grafmiller, J., & Szmrecsanyi, B. (2019). Particle placement alternation in EFL learner vs. L1 speech: Assessing the similarity of probabilistic grammars. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research: Selected papers from the fourth Learner Corpus Research Conference* (pp. 71–92). Louvain-la-Neuve: Presses universitaires de Louvain.

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, 3(1), 61–94. https://doi.org/10.1075/ijlcr.3.1.03paq

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, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, 100, 538–553. https://doi.org/10.1111/modl.12335

Polio, C., & Shea, M. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26(1), 10–27. https://doi.org/10.1016/j.jslw.2014.09.003

Purpura, J., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, 65(Suppl. 1), 37–75. https://doi.org/10.1111/lang.12112

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Révész, A. (2012). Coding second language data validly and reliably. In A. Mackey & S. Gass (Eds.), *Research methods in Second Language Acquisition: A practical guide* (pp. 203–221). Hoboken, NJ: Wiley-Blackwell. https://doi.org/10.1002/9781444347340.ch11

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380–401). Oxford: Oxford University Press.

Rosen, A., Hana, J., Stindlova, B., & Feldman, A. (2014). Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation*, 48, 65–92. https://doi.org/10.1007/s10579-013-9226-3

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Spooren, W., & Degand, L. (2010). Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2), 241–266. https://doi.org/10.1515/cllt.2010.009

Trafimow, D. (2017). The attenuation of correlation coefficients: A statistical literacy issue. *Teaching Statistics*, 38, 25–28. https://doi.org/10.1111/test.12087

Vyatkina, N. (2016). KANDEL: A developmental corpus of learner German. *International Journal of Learner Corpus Research*, 2(1), 102–120. https://doi.org/10.1075/ijlcr.2.1.04vya

Address for correspondence

Tove Larsson
Department of English
Uppsala University
Box 527
75120 Uppsala
Sweden
tove.larsson@engelska.uu.se

Co-author information

Magali Paquot
UCLouvain – FNRS
Centre for English Corpus Linguistics
Collège Erasme
magali.paquot@uclouvain.be

Luke Plonsky
Northern Arizona University
luke.plonsky@gmail.com