3.2.2 Open Science statement

Another important insight from the methodological part of the literature review (Chapter 2) is that, to the author's best knowledge, no Textbook English study published so far has included (as an appendix or supplementary materials) the data and code necessary to replicate the published results. This means that it is very difficult to evaluate the reliability or robustness of the results reported. Granted, a major issue in (corpus) linguistic research is that it is often not possible for copyright or, when participants are involved, data protection reasons to make linguistic data available to the wider public. However, both research practice and the impact of our research can already be greatly improved if we publish our code or, when using GUI software, methods sections detailed enough to be able to successfully replicate the full procedures. This step can enable others to conduct detailed reviews of our methodologies and conceptual replications of our results on different data.

Aside from data protection and copyright regulations, there are, of course, many reasons why researchers may be reluctant to share their data and code (Berez-Kroeker et al. 2018; McManus 2021). It is not within the scope of this thesis to discuss these; however, it is clear that, in many ways, such transparency makes us vulnerable. At the end of the day: to err is human. Yet, the risks involved in committing to Open Science practices is particularly tangible for researchers working on individual project, like the present author on this doctoral thesis, who have had no formal training in data management or programming and have therefore had to learn "on the job". Nonetheless, the author is convinced that the advantages outweigh the risks. Striving for transparency helps both the researchers themselves and others reviewing the work to spot and address problems. As a result, the research community can build on both the mishaps and successes of previous research, thus improving the efficiency of research processes and ultimately contributing to advancing scientific progress.

It is with this in mind that the author has decided, whenever possible, to publish all the raw data and code necessary to reproduce the results reported in the present thesis following the FAIR principles (i.e., ensuring that research data are Findable, Accessible, Interoperable and Reusable; Wilkinson et al. 2016). For copyright reasons, the corpora themselves and annotated corpus data in the form of concordance lines cannot be made available. However, the outcome of both manual and automatic annotation processes is published in tabular formats in the Online Appendix. These tables allow for the reproduction of all the analyses reported on in the following chapters using the reproducible data analysis scripts also published in the Online Appendix and on GitHub (https://github.com/elenlefoll/TextbookEnglish). In all

chapters of this thesis, full transparency is strived for by reporting on how each sample size was determined and on which grounds data points were excluded, manipulated and/or transformed. Most of these operations were conducted in the open-source programming language and environment R (R Core Team 2020). Most of the data processing and analysis scripts therefore consist of R notebooks. These were rendered to HTML pages (viewable in the Online Appendix) thus allowing researchers to review the procedures followed without necessarily installing all the required packages and running the code themselves. These scripts also feature additional analyses, tables and plots that were made as part of this study but which, for reasons of space, were not reported on in detail in the present thesis. Whenever additional software or open-source code from other researchers were used, links to these are also provided in the Online Appendix (in addition to the bibliographic references in the corresponding sections of the thesis).