

On relativizer use in French learners of English: a corpus-based study

Paula Lissón, Nicolas Ballier, Kim Gerdes

Universität Potsdam, Université Paris Diderot, Université Sorbonne Nouvelle

paula.lisson@uni-potsdam.de, nicolas.ballier@univ-paris-diderot.fr, kim.gerdes@sorbonne-nouvelle.fr

In standard English, the use of an overt relativizer in object relative clauses (ORCs) is optional (Quirk, 1957). In sentences such as “That’s the music that I like”, the speaker may use an overt relativizer (*which/that*), or a null relativizer ($\emptyset$, also known as *zero*). There are some psycholinguistic theories that address relativizer omission in L1 speakers of English, such as ambiguity avoidance (Temperley, 2003) the Predictability Hypothesis (Wasow, Jaeger, & Orr, 2011), or entrenchment (Wiechmann, 2015), among others. These theories have been essential for many corpus-based studies dealing with the detection of features that condition relativizer use, such as definiteness of the antecedent, complexity, lexical density, or priming. In the learner corpora community, however, relativizer omission literature is scarce: as far as we know, only a few studies deal with relativizer omission/alternation in a quantitative way (e.g. Lester, forthcoming; Olofsson, 2009).

The present work reports new insights on learner use of relativizers. In particular, our study focuses on whether there is a preference in the use of overt vs. null relativizer, and whether different overt relativizers are used with the same frequency and within the same contexts. We also investigate which features correlate with the use of the null relativizer ($\emptyset$) vs. two overt relativizers (*that/which*). We address three major questions:

- which are the factors or patterns correlated with the absence or presence of a relativizer?
- which are the factors or patterns correlated with the use of *which* and the use of *that*?
- are the factors or patterns distinctive enough to be different from the ones previously attested in L1 use?

In order to investigate these questions, we downloaded a subset of the EFCAMDAT corpus² (Geertzen, Alexopoulou, & Korhonen, 2013). The texts were parsed with SpaCy (Honnibal & Johnson, 2015), and ORCs were extracted using a Python script³. The final dataset contained 1,675 ORCs from 560 learners, across 24 tasks. We adapted part of the methodology followed in previous studies on relativizer use (e.g. Fox & Thompson, 2007; Grafmiller, Szmrecsanyi, & Hinrichs, 2016; Hundt, Denison, & Schneider, 2012). Each of the extracted RC was coded with the following features: definiteness, number, length of the antecedent, length of the sentence containing the RC, task, type-to-token ratio (TTR), number of sentences of the text, mean length of the sentences, and learner ID.

A random forest model was fit to the data (see Table 1), with all the aforementioned features as predictors. The model shows that the most relevant features related to the variability in use of the relativizers *which*, *that*, and *zero* are related to external predictors, i.e., stylistic factors of the text where the relative clause was found. A closer analysis of the most important variables within the classification model revealed that TTR⁴ scores and mean sentence length values were higher for most of the texts where *zero* relative clauses were found. This, contrary to the complexity principle by Rohdenburg (1996), indicates that learners tend to produce the *zero* relativizer in complex contexts. Interestingly, this was also found in Lester (forthcoming) for learners’ spoken use of relativizers.

The analysis of the properties of the antecedents will be also presented and compared with previous studies on English L1 (see Table 2), and the role of the task will be discussed: the frequency and features of the antecedents could be the result of the elicited tasks. We will also discuss the difficulties in the automatic retrieval and annotation of ORCs from a learner corpus, as well as the specific details of the random forest model.

Table 1: Confusion matrix

	that	which	zero
that	223	0	0
which	0	35	0
zero	29	7	1381

² Productions corresponding to the B2 level of the CEFR from the French learners of the corpus.

³ <https://github.com/kimgerdes/SUD/tree/master/tools>

⁴ Used as a proxy, however questionable, for lexical complexity.

Table 2: Comparison of
with relativizers in Wasow et al.

Head-noun	Expected relativizer	Results in this study (number of occurrences)
people	overt	that (1) zero (16)
stuff, little	overt	not enough occurrences
all	zero	that (4) zero (20)
way	zero	that (1) zero (53)
time	zero	that (2) zero (38)
first, last, every	zero	not enough occurrences

antecedents highly correlated
(2011)

References

- Fox, B. A., & Thompson, S. A. (2007). Relative clauses in English conversation: Relativizers, frequency, and the notion of construction. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language"*, 31(2), 293–326.
- Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadia Proceedings Project.
- Grafmiller, J., Szmrecsanyi, B., & Hinrichs, L. (2016). Restricting the restrictive relativizer. *Corpus Linguistics and Linguistic Theory*. <https://doi.org/10.1515/cllt-2016-0015>
- Honnibal, M., & Johnson, M. (2015). An Improved Non-monotonic Transition System for Dependency Parsing. (pp. 1373–1378). Presented at the EMNLP.
- Hundt, M., Denison, D., & Schneider, G. (2012). Relative complexity in scientific discourse. *English Language & Linguistics*, 16(2), 209–240.
- Lester, N. (forthcoming). That's hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research*.
- Olofsson, A. (2009). The Gift of the Gap: A Study of Dutch and Swedish Learners' Use of the English Zero Relativizer. *English Studies*, 90(3), 333–344.
- Quirk, R. (1957). Relative clauses in educated spoken English. *English Studies*, 38(1–6), 97–109.
- Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics (Includes Cognitive Linguistic Bibliography)*, 7(2), 149–182.
- Temperley, D. (2003). Ambiguity avoidance in English relative clauses. *Language*, 464–484. <https://doi.org/10.1353/lan.2003.0189>
- Wasow, T., Jaeger, T. F., & Orr, D. (2011). Lexical variation in relativizer frequency. *Expecting the Unexpected: Exceptions in Grammar*, 175–195. <https://doi.org/10.1515/9783110219098.175>
- Wiechmann, D. (2015). *Understanding relative clauses: A usage-based view on the processing of complex constructions* (Vol. 268). Walter de Gruyter GmbH & Co KG.