Replicating a Data-Driven Corpus Analysis: The Example of Academic Language

Andresen, Melanie

melanie.andresen@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Pichler, Axel

axel.pichler@ts.uni-stuttgart.de
Universität Stuttgart, Germany

Replication studies are generally considered an important way of checking the generalizability of research results (see for instance Berez-Kroeker et al. 2018; Schöch et al. 2020). However, in the humanities they are still comparatively rare. We argue that data-driven studies in particular are in need of replication. While theory-driven studies derive a hypothesis based on a specific theory and thus can provide good reasons for their hypothesis prior to the empirical study, data-driven studies start with the data and relate the results to theory and previous research post-hoc (Kitchin 2014). In doing so, they work abductively and infer the best explanation for the patterns revealed (Wadepuhl 2016, Gerstorfer 2020). This way, they are especially prone to inadequate explanations and overgeneralization.

As an example, we present a (partial) replication of the data-driven study in (Andresen 2022). This study looks into the differences in the German academic languages of literary studies and linguistics. Two corpora of 30 PhD theses per discipline are compared to each other, considering tokens, parts of speech and syntax as well as several types of sequences of these features. For the replication, we focus on the comparison of the parts-of-speech distribution. The comparison is performed by training a machine learning classifier based on the support vector machine algorithm (SVM) that learns to distinguish the two disciplines based on their parts-of-speech distribution. The SVM provides coefficients that express how useful each feature was for the classification task. All parts of speech are ranked by these coefficients and the most helpful features, i.e., those with the largest differences between the disciplines, are presented and related to possible reasons for the differences.

Following in the typology of (Porte 2012: 8), we present an approximate replication study as we change the variable text type: Instead of PhD thesis we analyze a corpus of journal papers from the same two disciplines, literary studies and linguistics. We use a corpus of 45 papers per discipline with 626,316 tokens in total. Except for this change, we stick to the methodology of the original study and compare the results.

When comparing the coefficients of the two studies, there is only a moderate correlation with a Pearson's r of 0.55. This indicates that there are some substantial differences between the results of the two studies. Table 1 shows the ranking of the most helpful features in the replication study with the rank in the original study in parentheses. Some of the results are quite stable: Literary studies uses more proper nouns (NE), possessive pronouns (PPOSAT) and personal pronouns (PPER). This is related to the high importance of humans in the discipline, be it as authors or literary characters. Also, the high frequency of numbers in linguistics is very stable and explainable by more quantitative research in the discipline. However, other features display very different trends in the replication study: For instance, finite modal verbs (VMFIN) and infinite full verbs (VVINF) have even changed sides and are now indicators of literary studies. This indicates that the conclusions drawn in Andresen (2022) regarding verbal patterns in the disciplines might only apply to the text type PhD thesis or not generalize beyond the specific corpus at all.

Table 1: The 15 part-of-speech tags (tagset: STTS, Schiller et al. 1999) with the highest SVM coefficients (C=500). Rank in the original study indicated in parentheses. Tags highlighted in bold have changed sides compared to the original study.

Rank	**Literary Studies**	Score	Linguistics
1		-40.11	CARD (↑3)
2	NE (↓1)	36.95	
3		-29.41	FM (↑25)
4		-28.46	APPRART (↑29)
5		-24.14	**APPR (↑35)**
6		-22.14	VVPP (↑12)
7	ART (–)	20.83	
8	PRELS (↑13)	20.53	
9	ADV (↑16)	19.96	
10	PPOSAT (↓4)	18.97	
11	**VVINF (↑19)**	18.50	
12		-18.10	XY(↑20)
13	PPER (↓5)	16.66	
14		-15.99	TRUNC (↑22)
15	**VMFIN (↑18)**	14.31	

In sum, our replication study shows that by changing the text type that represents academic language, some results of the original study can be confirmed while others are in fact inverted. This raises the question of the status of abductively obtained (hypo)theses in data-driven research. Is such a (hypo)thesis to be discarded as a consequence of a replication study that does not have the same experimental prerequisites or to be specified according to the new experimental conditions? We argue that replication studies should be performed more frequently in the digital humanities (and beyond) if the aim of a study is to allow for generalizations. Especially the status of data-driven results as generating hypotheses should be taken more seriously. If our results are not rooted in a theory that justifies them, gives them plausibility and is formulated in a way that makes it possible to be falsified, we might be tempted to overgeneralize interpretations. It would therefore be desirable for data-driven results to be backed up by additional studies which on the one hand evaluate these hypotheses deductively and on the other hand test their range.

Bibliography

Andresen, Melanie (2022): Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 10) Tübingen: Narr Francke Attempto.

Berez-Kroeker, Andrea L. et al. (2018): "Reproducible research in linguistics: A position statement on data citation and attribution in our field", in: Linguistics 56 (1): 1–18. 10.1515/ling-2017-0032.

Gerstorfer, Dominik (2020): "Entdecken und Rechtfertigen in den Digital Humanities" in: Reiter, Nils / Pichler, Axel / Kuhn, Jonas (eds.): Reflektierte algorithmische Textanalyse: Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt. De Gruyter 107–124. 10.1515/9783110693973.

Kitchin, Rob (2014): "Big Data, new epistemologies and paradigm shifts", in: Big Data & Society 1 (1): 1–12. 10.1177/2053951714528481.

Porte, Graeme (2012): "Introduction" in: Porte, Graeme (ed.): Replication research in applied linguistics (= The Cambridge applied linguistics series). Cambridge, New York: Cambridge University Press 1–17.

Schiller, Anne / Teufel, Simone / Thielen, Christine / Stöckert, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Schöch, Christof / Dalen-Oskam, Karina van / Antoniak, Maria / Jannidis, Fotis / Mimno, David (2020): Replication and Computational Literary Studies. in: Digital Humanities 2020 (DH2020), Book of Abstracts. https://doi.org/10.5281/zenodo.3893427 [letzter Zugriff 25. August 2020].

Wadephul, Christian (2016): "Führt Big Data zur abduktiven Wende in den Wissenschaften?", in: Berliner Debatte Initial 27 (4): 37–49.