Originally published in: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019. Pp. 269-270

Ein Tool zur Visualisierung des Gebrauchs von Schreibvarianten

Peter M. Fischer
Leibniz-Institut für Deutsche Sprache
R 5 6-13
68161 Mannheim
peter.fischer@ids-mannheim.de

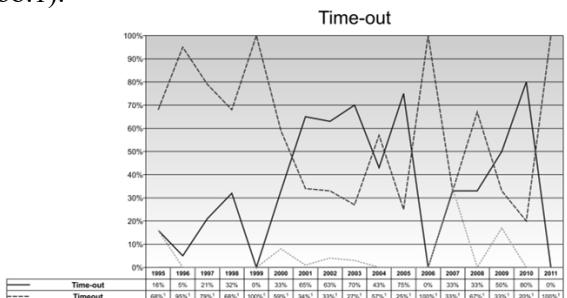
Christian Lang
Leibniz-Institut für Deutsche Sprache
Augustaanlage 32
68165 Mannheim
lang@ids-mannheim.de

Abstract

In unserem Beitrag stellen wir die Entwicklung eines komponentenbasierten Tools zur Abfrage, Auswertung und Visualisierung von Schreibvarianten vor.

1 Einleitung

Die diachrone empirische Untersuchung von Varianz in der Schreibung von Wörtern anhand von Korpusdaten spielt eine zentrale Rolle in der Orthografieforschung und wird auch in Untersuchungen der AG Korpus des Rats für deutsche Rechtschreibung angewendet (vgl. Krome/Roll 2016: 7). Dazu werden in Einzelrecherchen aussagekräftige Grafiken erstellt, die für gegebene Schreibvarianten unter gegebenen Schreibern und in einem gegebenen Untersuchungszeitraum eine übersichtliche Gegenüberstellung entsprechender Schreibgebräuche ermöglichen (vgl. Abb.1).



[1] Keine gültige Schreibung.

Abb. 1: Erhebung der Vorkommen der Schreibungen „Time-out“, „Timeout“, „Time Out“ und „Time out“ (Rat für deutsche Rechtschreibung)

Der Erstellung solcher Grafiken geht ein mehrstufiger Prozess voraus (Auswahl der Korpusgrundlage, Erstellung und Durchführung von Korpusabfragen, Zusammentragen der Abfrageergebnisse, Erstellung der Grafiken), der an verschiedenen Stellen nicht oder nur teilweise automatisiert ist. In Konsequenz wird dieser Prozess daher stark indikations- und bedarfsgeleitet angestoßen, was ein flächendeckendes Monitoring des allgemeinen Schreibgebrauchs ausbremst. Um den Rat für deutsche Rechtschreibung in seiner Arbeit zu unterstützen, wird am Leibniz-Institut für Deutsche Sprache (IDS) eine Software entwickelt, die diesen Prozess durch breitere Automation in den Abläufen fördert.

2 Ziele und Anforderungen

Eine Kernfunktionalität des Tools liegt in der interaktiven Erstellung von Schaubildern, welche die Entwicklung benutzerdefinierbarer Häufigkeitsmaße von Schreibvarianten entlang einer Zeitachse abbilden. Die Benutzeroberfläche begleitet Anwender/-innen dabei im gesamten Erstellungsprozess von der Korpusauswahl bis zum Grafikexport und bietet einen intuitiven und autonomen Zugang zu Untersuchungsergebnissen.

Im Hinblick auf die statistischen Auswertungen und Diagrammtypen greift das hier vorgestellte Tool die Methodologie der auch am IDS entwickelten Programm-Infrastruktur[1] zur Generierung sog. Zeitverlaufsgrafiken (ZVGs) auf, knüpft aber an die ebenfalls am IDS entwickelte universelle Korpusanalyseplattform *KorAP* (Kupietz et al. 2019) an. Da *KorAP* als quelloffenes Projekt entwickelt wird, besteht keine Abhängigkeit zu IDS-Infrastrukturen. Zudem kann *KorAP* auch auf physisch verteilt liegenden Ressourcen operieren, sodass auch der Standort der auszuwertenden Korpora nicht an das IDS gebunden ist, was den Bedürfnissen des Rats für deutsche Rechtschreibung insbesondere in Bezug auf Untersuchungen in bzw. aus den Gebieten seiner sieben internationalen Mitglieder gerecht wird.

Die **einfach** strukturierte und leicht zugängliche Benutzeroberfläche befähigt Nutzer/-innen, auch ohne größere Kenntnis von Korpusrecherche, Datenauswertung und -visualisierung komfortabel vergleichende Grafiken zu erstellen.

[1] http://www1.ids-mannheim.de/kl/projekte/methoden/mdca/zvgs.html

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-93129

Ein- und Ausgabedaten können **flexibel** festgelegt und an das jeweilige Auswertungsszenario angepasst werden. Aufseiten der Eingabe sind die Auswahl zur Verfügung stehender Korpora und ihre weitere Eingrenzung zu bestimmen (siehe virtuelle Korpora, Diewald et al. 2016) und eine Liste zu untersuchender Schreibvarianten anzugeben. Aufseiten der Ausgabe können die Art der Darstellung in Form unterschiedlicher statistischer Maße wie absolute Frequenz, relative Frequenz, Häufigkeitsklassen (vgl. Perkuhn et al. 2012:80), prozentuale Verteilung der Varianten etc. sowie beim Export der erstellten Visualisierung das Dateiformat (u.a. png, jpg, pdf, svg, xml oder html) gewählt werden.

Die Visualisierung wird bei anwenderseitig nachträglich eingebrachten Änderungen der Darstellungskriterien **dynamisch** angepasst. Dies erlaubt eine rasche und intuitive Justierung der Grafik bis zum gewünschten Bild.

3 Software-Architektur

Das Tool sieht sich als stark spezialisierte Abfrage-, Auswertungs- und Darbietungsplattform und knüpft daher auf der einen Seite an bestehende Schnittstellen zur allgemeinen Korpusabfrage an und bildet auf der anderen Seite selbst eine Schnittstelle zur Benutzerinteraktion.

Zur Erlangung der **Datengrundlage** bedient sich das Tool skriptgesteuert einer Programmierschnittstelle (*API*) der Korpusanalyseplattform *KorAP* (Kupietz et al. 2019), die für die Verarbeitung sehr großer Korpora mit mehreren Annotationsebenen, mehreren Abfragesprachen und komplexen Lizenzmodellen optimiert ist und in Modulen als Open Source auf GitHub[2] veröffentlicht wird. Über die am IDS installierte Instanz[3] besteht somit u.a. Zugang zum Deutschen Referenzkorpus *DeReKo* (Kupietz et al. 2018).

Die **Visualisierung** ist als Webapplikation in R (R Core Team 2016) mithilfe des R-Paketes *shiny* (Chang et al. 2019) implementiert, das eine interaktive Datenpräsentation in einer Weboberfläche und damit dynamische Anpassungen gemäß Anwendereingaben ermöglicht.

4 Ausblick

Perspektivisch ist geplant, das Tool auch als Modul von *grammis*[4], dem grammatischen Informationssystem des IDS, der Öffentlichkeit zugänglich zu machen. So soll es die Anwendung *KoGra-R* (Hansen-Morath et al. 2019)[5] komplementieren, die ebenfalls am IDS entwickelt wurde und korpuslinguistische Statistiken und Visualisierungen bereithält, allerdings nicht für den Variantenvergleich konzipiert wurde.

[2] http://github.com/KorAP

[3] https://korap.ids-mannheim.de/

[4] https://grammis.ids-mannheim.de

[5] http://kograno.ids-mannheim.de/index.html

Literaturangaben

Chang, W./Cheng, J./Allaire, JJ/Xie, Y./McPherson, J. 2019. *shiny: Web Application Framework for R.* https://cran.rproject.org/web/packages/shiny/shiny.pdf, R package version 1.3.2.

Diewald, N./Hanl, M./Margaretha, E./Bingel, J./Kupietz, M./Bański, P./Witt, A. 2016. KorAP Architecture – Diving in the Deep Sea of Corpus Data. In: Calzolari, N. et al. (Hrsg.). 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.* Paris: European Language Resources Association (ELRA), S. 3586–3591.

Hansen-Morath, S./Schmitz, H-C./Schneider, R./Wolfer, S. 2019. KoGra-R: Standardisierte statistische Auswertung von Korpusrecherchen. In: Fuß, E./Konopka, M./Wöllstein, A. (Hrsg.). 2019. *Grammatik im Korpus*. Tübingen: Narr. S. 299–357.

Krome, S./Roll, B. 2016. Fremdwörter zwischen Isolation und Integration. Empirische Analysen zum Schreibusus auf der Basis von Textkorpora professioneller und informeller Schreiber. In: *Studia Germanistica 19/2016*. Ostrava: Ostravská univerzita. S. 5–40.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, N. et al. (Hrsg.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), S. 4353–4360.

Kupietz, M./Diewald, N./Margaretha, E./Bodmer, F./Stallkamp, H./Harders, P. 2019. Neues von KorAP. In: Eichinger, L. M./Plewnia, A. (Hrsg.): *Neues vom heutigen Deutsch. Empirisch – methodisch – theoretisch. Jahrbuch des Instituts für Deutsche Sprache 2018.* Berlin/Boston: de Gruyter. S. 345–349.

Perkuhn, R./Keibel, H./Kupietz, M. 2012. *Korpuslinguistik*. Paderborn: Fink, 2012.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.