Originalveröffentlichung in: Beißwenger, Michael/Lemnitzer, Lothar/Müller-Spitzer, Carolin (Hrsg.): Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium. – Paderborn: Wilhelm Fink, 2022. S. 162-178. (UTB 5711)





**Sascha Wolfer**
**Carolin Müller-Spitzer**

# 9. Sind Wörterbücher wirklich nützliche Werkzeuge beim Überarbeiten von Texten? Ein experimenteller Zugang

Wir stellen eine empirische Studie vor, die der Frage nachgeht, ob und in welchem Ausmaß Wörterbücher und andere lexikographische Ressourcen die Ergebnisse von Textüberarbeitungen verbessern. Studierende wurden in unserer Studie gebeten, zwei Texte zu optimieren und waren dabei zufällig in drei unterschiedliche Versuchsbedingungen eingeteilt: 1. ein Ausgangstext ohne Hinweise auf potenzielle Fehler im Text, 2. ein Ausgangstext, bei dem problematische Stellen im Text hervorgehoben waren und 3. ein Ausgangstext mit hervorgehobenen Problemstellen zusammen mit lexikographischen Ressourcen, die zur Lösung der spezifischen Probleme verwendet werden konnten. Wir fanden heraus, dass die Teilnehmer*innen der dritten Gruppe die meisten Probleme korrigierten und die wenigsten semantischen Verzerrungen während der Überarbeitung einführten. Außerdem waren sie am effizientesten (gemessen in verbesserten Textabschnitten pro Zeit). Wir berichten in dieser Fallstudie ausführlich vom Versuchsaufbau, der methodischen Durchführung der Studie und eventuellen Limitationen unserer Ergebnisse.

## 1. Einleitung

Es ist ein alltäglicher Bestandteil des Schreibprozesses, Texte zu überarbeiten und zu verbessern. Man schreibt einen Text meist zunächst als Rohtext und im Anschluss daran versucht man ihn zu glätten, Fehler auszubessern, Wiederholungen zu streichen, Formulierungen abwechslungsreich zu gestalten etc. Schreibassistenzsysteme können diesen Prozess unterstützen, indem sie zum einen die potenziell fehlerhaften Stellen im Text markieren und zum anderen möglichst auch Verbesserungsvorschläge aufzeigen. Textverarbeitungsprogramme leisten bislang meist nur Unterstützung bei fehlerhafter Rechtschreibung und bei Kongruenzfehlern (z. B. „einen blinde Fisch"). Denkbar sind aber weiterreichende Unterstützungen, z. B. Ersetzungsvorschläge für unpassende Kollokationen (z. B. „Hund an der Schnur führen" statt „Leine") oder Hinweise zu potenziellen Registerproblemen (z. B. „Klamotten" statt „Kleidung"). Solche Informationen, aus denen diese Hinweise extrahiert werden können, finden sich in Wörterbüchern. Aus ihnen können gezielt bestimmte Angaben herausgegriffen werden, die für die Verbesserung von einzelnen Textstellen hilfreich sind. Eine solche Schreibumgebung zu entwickeln ist aber natürlich eine große Herausforderung. Auch das gezielte Extrahieren von Informationen aus Wörterbüchern ist keine triviale Aufgabe. Deshalb ist es wichtig zu wissen, ob sich dieser Aufwand lohnt. Denn die Aussage, dass Informationen aus Wörterbüchern oder anderen Nachschlagewerken in diesem Kontext hilfreich sind, ist natürlich erst einmal nur das: eine Aussage, die es zu überprüfen gilt – und diese Überprüfung können wir wissenschaftlich mit einer empirischen Untersuchung angehen, wie wir sie Ihnen hier vorstellen möchten. Diese Studie war ein Kooperationsprojekt zwischen den Universitäten Mannheim und Darmstadt, der Eurac Research in Bozen und dem Leibniz-Institut für Deutsche Sprache in Mannheim. Eine ausführliche Beschreibung der Studie findet sich in Wolfer et al. (2018).

## 2. Fragestellung

Die übergeordnete Forschungsfrage, die wir aus den obigen Ausführungen ableiten, lautet: Helfen lexikographische Ressourcen bei der Überarbeitung von Texten? Zur Untersu-

chung dieser Frage entschieden wir uns, Menschen vor sprachliche Probleme zu stellen, die sie dann mithilfe von lexikalischen Ressourcen, die wir Ihnen ebenfalls bereitstellten, bearbeiten sollten, genauso wie es in einer Schreibassistenzumgebung der Fall sein könnte. Doch das alleine würde noch nicht ausreichen, um die Frage zu beantworten, ob Wörterbücher wirklich hilfreich sind, denn es fehlt hier ein Vergleich mit anderen Gruppen, die keine lexikographischen Ressourcen verwenden können. Wir können nur dann schließen, dass lexikographische Ressourcen helfen, wenn wir Hinweise darauf finden können, dass Menschen mit Informationen aus Wörterbüchern Texte *besser* überarbeiten, als wenn sie diese nicht bekommen. Deshalb haben wir die Aufgabe in drei Gruppen variiert: Zusätzlich zu der Wörterbuch-Gruppe gab es eine Gruppe, die nur die fehlerhaften Texte bekam ohne irgendwelche Hinweise darauf, wo problematische Stellen liegen könnten. Diese Gruppe musste also Texte ohne die Unterstützung von irgendetwas verbessern, was einer oben skizzierten Schreibumgebung ähnlich ist. In einer zweiten Vergleichsgruppe haben wir die Situation simuliert, dass eine Schreibumgebung auf die problematischen Stellen im Text hinweist, sie also markiert, aber keine zusätzlichen Informationen zur Verbesserung geboten werden. Unsere Fokusgruppe war damit die Gruppe, die mithilfe lexikographischer Ressourcen die Texte verbessern konnten, unsere Kontroll- und Vergleichsgruppen waren die beiden anderen Gruppen, die dieses Hilfsmittel nicht bekamen. Wir können die Fragestellungen in zwei Hypothesen präzisieren:

> H1: Markierungen der problematischen Textstellen sind für die Textüberarbeitung hilfreich, d. h. die Revisionsergebnisse der beiden Textversionen mit hervorgehobenen Sprachproblemen liefern eine höhere Qualität der Überarbeitungen als die Ergebnisse der Version, in der die Proband*innen keine Hinweise auf problematische Textstellen erhalten.
> H2: Lexikographische Ressourcen haben zusätzlich positive Auswirkungen auf die Überarbeitungs-Qualität, d. h. Proband*innen, die zusätzlich zu den Hinweisen auf problematische Textstellen auch noch Auszüge aus lexikographischen Ressourcen erhalten, übertreffen die Teilnehmer*innen mit der Version der hervorgehobenen Probleme und der Nur-Text-Version.

Diese Variation in der Aufgabenumgebung und weitere Details zum genauen Aufbau der Studie werden wir Ihnen in Abschnitt 3 darlegen. Danach gehen wir in Abschnitt 4 auf die Ergebnisse der Untersuchung ein. In Abschnitt 5 diskutieren wir die Ergebnisse im Rückbezug auf die Forschungsfrage.

## 3. Material, Methode und Analyse

Unsere Forschungsfrage allein legt noch kein eindeutiges Vorgehen während der Studie fest. Es gilt an mehreren Stellen Entscheidungen zu treffen, wie genau vorgegangen wird. Die Gliederung dieses Abschnitts deckt all diese Bereiche ab, damit Sie sich ein genaues Bild davon machen können, wie die Studie ablief. Wir beginnen damit, Ihnen den logischen Aufbau der Untersuchung, das sog. Studiendesign, nahezubringen.

### 3.1 Studiendesign

Bei der vorgestellten Studie handelt es sich um eine reaktive Querschnittsstudie (vgl. → Kapitel 2 [Grundlagen] in diesem Band). Das heißt, dass es *einen* Messzeitpunkt gibt (also keine Entwicklung über die Zeit hinweg gemessen wird) und dass die Teilnehmer*innen über die Studie Bescheid wussten und sich somit bewusst waren, dass sie Teil einer Datenerhebung sind. Es handelt sich darüber hinaus um ein faktorielles Versuchsdesign, denn zwei Faktoren wurden gezielt gekreuzt, um die Auswirkung der Manipulation dieser Faktoren (oder unabhängiger Variablen) auf die Ergebnisse (die abhängigen Variablen) zu testen. Diese beiden Faktoren waren 1. der Text, den die Teilnehmer*innen bearbeiteten und 2. die Art der Hilfestellung, die den Teilnehmer*innen bei der Überarbeitung der Texte gegeben wurde. Der Faktor 1 (Text) umfasste zwei Ausprägungen, nämlich einen

Schülertext über das Thema „Jugend" und einen Text über Phraseologismen, der von einer/-m Studierenden geschrieben wurde. Alle Teilnehmer*innen bearbeiteten beide Texte nacheinander, die Abfolge der Texte war immer zufällig. Faktor 2 (Hilfestellung) umfasste drei Ausprägungen: „Nur Text", „Markierung" und „Markierung+Wörterbuch". Das bedeutet, dass die Teilnehmer*innen entweder nur die Texte vorgelegt bekamen (also keine Hilfestellungen bei der Überarbeitung bekamen) oder Texte bearbeiten sollten, in denen kritische Stellen (dazu mehr in Abschnitt 3.3.) hervorgehoben waren. Die dritte Ausprägung des Faktors umfasste ebenfalls diese Markierungen, doch zusätzlich wurden unterschiedliche lexikographische Ressourcen eingeblendet, die bei der Lösung der Probleme helfen konnten. Jede Person wurde zufällig einer dieser Faktorausprägungen zugelost, d.h., eine Person sah immer nur eine Version der Texte. Da diese beiden Faktoren gekreuzt wurden, spricht man in diesem Fall von einem 2x3 mixed-design[1]. Tabelle 1 gibt einen Überblick über das Versuchsdesign (s. auch Abbildung 1 zur Illustration der Versuchsbedingung „Markierung+Wörterbuch"). Bei experimentellen Studien bietet sich eine solche Darstellung aus mehreren Gründen an: Erstens bekommen die Leser*innen einen schnellen Überblick über den Studienentwurf, zweitens kann im weiteren Verlauf des Artikels auf die entsprechenden Bezeichnungen und Kombinationen verwiesen werden.

### 3.2 Teilnehmerinnen und Teilnehmer

Alle Teilnehmer*innen waren Studierende im Grundstudium der Germanistischen Linguistik an der Universität Mannheim. Die Teilnahme an der Studie war Bestandteil einer einführenden Vorlesung in die Linguistik. Das bedeutet, dass die Gruppe der Teilnehmenden relativ homogen hinsichtlich ihrer fachlichen Ausrichtung war, was wir in dieser Studie zunächst als Vorteil ansehen, da dadurch inter-individuelle Variation zumindest teilweise eingeschränkt wird. Insgesamt sammelten wir Daten von 105 Teilnehmer*innen, davon gaben 26 an, dass Deutsch nicht ihre Muttersprache sei. Die Daten dieser Teilnehmer*innen wurden aus den Analysen ausgeschlossen, um die Ergebnisse nicht zu verzerren. Ein weiterer Fall wurde von der Analyse ausgeschlossen, da sie/er weniger als fünf Minuten mit der Bearbeitung der Aufgabe zugebracht hat (dies war eine arbiträre Grenze, die wir zuvor festgelegt hatten). Der Datensatz, der in die Analysen einging, umfasst somit Daten von 78 Teilnehmer*innen. 71 (91 %) dieser Personen gaben an, dass sie im ersten Semester Linguistik studierten, sechs Personen befanden sich im dritten Semester und eine Person im achten Semester. Die 78 Personen verteilten sich wie folgt auf die Versuchsbedingungen A/B, C/D und E/F (für die Bezeichnungen siehe Tabelle. 1). Nur Text: 26 Teilnehmer*innen; Markierung: 25 Teilnehmer*innen; Markierung+Wörterbuch: 27 Teilnehmer*innen. Wir fragten von

*Tab. 1:* Designtabelle für den Aufbau der Untersuchung. Die Buchstaben A bis F bezeichnen die unterschiedlichen Versuchsbedingungen, d.h. Kombinationen von Faktoren

| | | Faktor 2: Hilfestellung (between-participants) | | |
|---|---|---|---|---|
| | | Nur Text | Markierung | Markierung+Wörterbuch |
| Faktor 1: Text (within-participants) | Jugend | A | C | E |
| | Phraseologismen | B | D | F |

1 „2x3" deshalb, weil der erste Faktor „Text" zwei Ausprägungen und der zweite Faktor „Hilfestellung" drei Ausprägungen hat. „Mixed" deshalb, weil der erste Faktor „within-participants" variiert wird, d.h. jede*r Teilnehmer*in beide Texte sieht und der zweite Faktor „between-participants", d.h. dass jede*r Teilnehmer*in nur eine Hilfestellungsvariante bekommt.

den Teilnehmer*innen außerdem ab, wie oft sie einsprachige Wörterbücher verwenden. 17 (21,8 %) gaben an, „mindestens einmal pro Woche" einsprachige Wörterbücher zu verwenden. 23 (29,5 %) verwenden diese „mindestens einmal pro Monat", 24 (30,8 %) „mindestens einmal im halben Jahr" und 14 (17,9 %) „seltener oder nie". In den experimentellen Bedingungen „Nur Text", „Markierung" und „Markierung+Wörterbuch" zeigt sich eine gleichmäßige Verteilung dieser Antwortkategorien. Somit lässt sich kein Effekt der experimentellen Bedingung, der unten berichtet wird, auf die individuelle Erfahrung mit einsprachigen Wörterbüchern zurückführen.

### 3.3 Text- und Hilfsmittel-Material

Der Text zum Thema „Jugend" ist dem KoKo-Korpus (vgl. Abel et al. 2014) entnommen. Er wurde von einer Person in der zwölften Klasse am Gymnasium verfasst und umfasst 260 Wörter. In dem Text wurden im Vorfeld der Studie 20 problematische Stellen identifiziert, die wir im Folgenden als „Stolpersteine" bezeichnen. Diesen Terminus wählten wir, weil es sich nicht im strengen Sinne um eindeutige Fehler handelt, sondern eben um Textstellen, die verbesserungswürdig sind. Die Stolpersteine umfassten alle sprachlichen Ebenen und enthielten Probleme wie die Wahl eines unangemessenen sprachlichen Registers („bis der Arzt kommt" im Schulaufsatz), regionale Ausdrücke („Buben" statt „Jungen"), den fehlenden Einsatz des Konjunktivs, unpassende Kollokationen („die Fragestellung beläuft sich auf"), den Einsatz des unbestimmten Artikels, wo ein bestimmter Artikel angebracht wäre (und umgekehrt), die Wahl einer unpassenden Abstraktionsebene, den problematischen Einsatz von anaphorischen Personalpronomen, Probleme bei der Argumentstruktur von Verben („sich sein eigenes ‚Ich' besser kennen lernen"), der Wiederholung von Wörtern in kurzem Abstand usw. Der Text zum Thema „Phraseologismen" wurde der Einleitung einer studentischen Hausarbeit von der Uni Dortmund entnommen und umfasst 204 Wörter. Dort wurden 15 Stolpersteine identifiziert.

Beide Texte wurden auf zwei Bildschirmseiten aufgeteilt, um sowohl den Ausgangstext als auch ein Textfeld zur Bearbeitung im Browser auf der Seite unterzubringen. Mit Abbildung 1 können Sie sich ein Bild davon machen, wie das für die Versuchsteilnehmer*innen mit vollen Hilfestellungen aussah. Für die Gruppe, die nur die hervorgehobenen Stolpersteine sah, fielen die rechte Spalte mit den lexikographischen Ressourcen sowie die Verweise im Text (fettgedruckte Zahlen) weg, die gelben Markierungen blieben. In der „Nur-Text"-Bedingung fielen auch diese Markierungen weg.

Die Aufgabe der Versuchsteilnehmer*innen war es, in dem Textbearbeitungsfeld (in Abbildung 1 unten links) eine überarbeitete bzw. verbesserte Version des Textes einzutragen. Zu Beginn des Versuchs war dort lediglich der Text von oben identisch enthalten.

Die Hilfsmittel wurden – in den Bedingungen E und F – immer auf der rechten Seite des Bildschirms dargeboten. Wir haben diese Hilfsmittel in ihrem generellen Erscheinungsbild aneinander angeglichen, den Inhalt, der auf der jeweiligen Ressource präsentiert wurde, jedoch nicht verändert. Die Hilfsmittel wurden anhand der folgenden Ressourcen erstellt:

- canoonet: Eine Online-Ressource zu Wörterbüchern, Wortbildung und Grammatik (http://www.canoonet.eu/).[2]
- E-Valbu: Ein elektronischen Valenzwörterbuch deutscher Verben (https://grammis.ids-mannheim.de/verbvalenz).
- DWDS-Wortprofile: statistische Auswertungen des Digitalen Wörterbuchs der Deutschen Sprache zu typischen Wortverbindungen (Kollokationen) (https://www.dwds.de/wp).
- GermaNet: Ein lexikalisch-semantisches Netz zur deutschen Sprache (http://www.sfs.uni-tuebingen.de/GermaNet/).

2 Diese Ressource ist inzwischen nicht mehr in der Form vorhanden, wie wir sie zum Zeitpunkt der Studienerstellung genutzt haben. Unter der angegebenen URL findet sich ein entsprechender Informationstext.

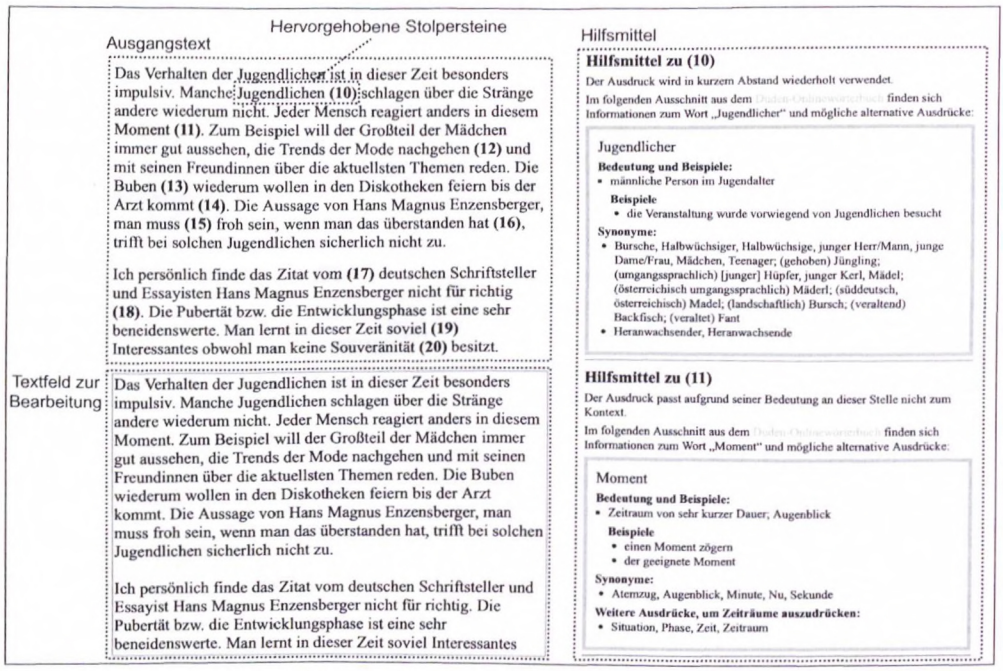


*Abb. 1:* Beispiel-Stimulus aus der Bedingung F (volle Hilfestellung, Schüler-Text). Die Annotationen sind nur zu Illustrationszwecken enthalten

- Duden online: Onlinewörterbuch des Bibliographischen Instituts[3].
- Grammis: Informationssystem zur deutschen Grammatik des IDS[4].

Wie Sie an dieser Aufzählung schon sehen können, ist die Bandbreite an Ressourcen recht hoch. Nicht nur klassische Wörterbuchressourcen gingen in die Studie ein, sondern auch lexikologische Nachschlageressourcen im weiteren Sinne.

## 3.4 Datenerhebung

Alle Daten wurden während der Zeit einer Vorlesungseinheit (1,5 Stunden) gesammelt. Die Teilnehmer*innen wurden zunächst zufällig auf zwei Vorlesungssäle an der Universität Mannheim aufgeteilt. Beim Eintritt in den jeweiligen Saal mussten die Teilnehmer*innen einen Zettel mit einer von drei URLs ziehen. Unter den abgedruckten URLs war jeweils eine Version des Experiments zu erreichen – eine der drei Hilfsmittelbedingungen „Nur Text", „Markierung" oder „Markierung+Wörterbuch". Die Teilnehmer*innen wurden mit mindestens zwei Plätzen Abstand in den Sälen platziert. Nachdem alle Teilnehmer*innen ihren Platz gefunden hatten, sollten sie mit ihren eigenen Geräten die URL aufrufen, die sie zuvor gezogen hatten. Die Studierenden wurden gebeten, ruhig an der Aufgabe zu arbeiten und während des Experiments nicht miteinander zu interagieren. In jedem der beiden Hörsäle waren mindestens drei Aufsichtspersonen anwesend. Es war den Teilnehmer*innen nicht erlaubt, andere Fenster außer des Browserfensters zu öffnen, andere Internetressourcen zu verwenden oder andere Geräte zu nutzen. Dies wurde von den herumgehenden Aufsichtspersonen überprüft.

3 https://www.duden.de/.

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

*Tab. 2:* Beispiele für die angewendeten Annotationskategorien anhand überarbeiteter Textausschnitte aus der Studie

| *Ausgangstext mit Markierung des Stolpersteins (Ausschnitt)* | *Überarbeitete Texte, die als „verbessert" annotiert wurden (bzgl. Stolperstein 14)* | *Überarbeitete Texte, die als „semantisch verzerrt" annotiert wurden (bzgl. Stolperstein 14)* |
|---|---|---|
| Die Buben (13) wiederum wollen in den Diskotheken feiern bis der Arzt kommt (14) | Die Jungen wiederum wollen ohne Einschränkung in den Diskotheken feiern. | Die Buben sind häufiger in den Diskotheken anzutreffen und stellen gesetzliche Grenzen in Frage, was beispielsweise den Alkohol angeht. |
| | Die Jungen wiederum wollen in den Diskotheken feiern ohne Grenzen. | Die männliche Jugendliche wiederum wollen in den Diskotheken feiern, oftmals mit fatalem Ende im Krankenhaus. |
| | Jungen wiederum wollen in den Diskotheken ungehalten feiern. | Den Jungs wiederum ist das Feiern in Diskotheken wichtiger. |

Das Experiment, das in der Online-Software QuestBack Unipark[5] implementiert war, begann mit einer detaillierten Instruktion zum Ablauf des Experiments. Diese Instruktion war selbstverständlich auf die jeweilige Version zugeschnitten. Alle Studierenden – egal in welcher Versuchsbedingung – wurden instruiert, sich eine Situation vorzustellen, in der sie den Text eines Kommilitonen bzw. einer Kommilitonin überarbeiten sollten. Sie sollten dabei nicht den Inhalt des Textes verändern, sondern lediglich auf Formulierungen achten. In den Hilfestellungsbedingungen „Markierung" und „Markierung+Wörterbuch" haben wir die Teilnehmer*innen außerdem in der Instruktion darauf hingewiesen, dass sie nicht für jede hervorgehobene Stelle unter allen Umständen eine alternative Formulierung finden müssen. Das Ziel war, die sprachlich beste Version des Textes zu finden.

Wir baten die Teilnehmer*innen, nach der Bearbeitung der Studie ruhig an ihrem Platz sitzen zu bleiben, um die anderen nicht zu stören. Damit es nicht zu attraktiv war, schnell mit der Studie fertig zu werden, kündigten wir anfangs an, dass in der verbleibenden Zeit noch Grammatikübungen zu lösen wären.

### 3.5 Aufbereitung der Überarbeitungen

Als Studienresultat erhielten wir die überarbeiteten Texte der Teilnehmer*innen als Fließtexte. Im nächsten Schritt mussten diese Texte nach der Studie annotiert werden, um zu sehen, welche Stolpersteine überhaupt bearbeitet worden waren (die möglichen Werte pro Stolperstein waren hier „ja" oder „nein"), welche davon tatsächlich verbessert worden waren und ob sich die Bedeutung des Texts durch die Überarbeitungen verändert hat (die möglichen Werte waren somit pro Stolperstein „verbessert", „unverändert", „verschlechtert" und „semantisch verzerrt", Letzteres ist eine Unterkategorie von „verschlechtert", s. illustrierende Beispiele in Tabelle 2). Zwei Personen annotierten diese Informationen unabhängig voneinander und wir prüften, wie gut die Annotationen übereinstimmten. Detailliertere Informationen über den Annotationsprozess bieten Wolfer et al. (2018). Wie Sie aus den obigen Ausführungen entnehmen können, haben wir nur jene Überarbeitungen berücksichtigt, die sich auf die vorher von uns identifizierten Stolpersteine bezogen, d.h., wir haben nicht alle Veränderungen gegenüber dem Ausgangstext analysiert. Dies gilt auch für alle weiteren Analysen.

## 4. Ergebnisse und Diskussion

Wir werden in den nächsten drei Abschnitten 4.1. bis 4.3. die annotierten Variablen „Veränderung", „Verbesserung" und „semantische Verzerrung" auf Basis der Stolpersteine ana-

5 https://www.unipark.com/.

*Tab. 3:* Ergebnistabelle für die abhängige Variable Überarbeitungen. Die Werte in den Zellen geben an, wie viel Prozent der Stolpersteine in den jeweiligen Bedingungen verändert wurden. Die Randmittelwerte erlauben einen Vergleich von Zeilen bzw. Spalten. Ein Randmittelwert wird immer für eine komplette Spalte oder Zeile berechnet. Für die erste Zeile gehen die Werte 48,5 %, 83,6 % und 89,8 % ein. Der Randmittelwert beträgt so 74,0 %. Der Wert ganz unten rechts ist der Gesamtmittelwert, d. h. dass über alle Bedingungen hinweg zwei Drittel aller Stolpersteine überarbeitet wurden. Die Werte sind jeweils auf eine Nachkommastelle gerundet

| | | Faktor 2: Hilfestellung (between-participants) | | | *Rand-mittel-werte* |
|---|---|---|---|---|---|
| | | Nur Text | Markierung | Markierung+Wörterbuch | |
| Faktor 1: Text (within-participants) | Jugend | 48,5 % | 83,6 % | 89,8 % | *74,0 %* |
| | Phraseologismen | 21,5 % | 64,5 % | 88,1 % | *58,4 %* |
| *Randmittelwerte* | | *36,9 %* | *75,4 %* | *89,1 %* | *67,3 %* |

lysieren. In den Abschnitten 4.4 und 4.5. werden wir ein Punkte-basiertes Maß einführen und auf dieser Grundlage die Performanz und Effizienz der Versuchsteilnehmer*innen in den verschiedenen Experimentalbedingungen vergleichen. Die Rohdaten können wir aufgrund der Einverständniserklärung, die die Teilnehmer*innen unterschrieben haben, nicht offen zur Verfügung stellen[6].

## 4.1 Überarbeitungen

Tabelle 3 zeigt den Anteil der veränderten Stolpersteine in den verschiedenen Versuchsbedingungen. Diese Tabelle ist der Design-Tabelle (Tabelle 1) sehr ähnlich, denn wir schreiben die Ergebnisse einfach in die entsprechenden Zellen.

Die Ergebnisse dieser Tabelle sind in Abbildung 2 visualisiert. Dort sehen Sie außerdem, wie viele Stolpersteine maximal in der jeweiligen Versuchsbedingung überarbeitet werden konnten (angegeben mit *n* unten in den Balken). Hierzu eine kurze Erläuterung: Im ersten Balken ist „n = 520" vermerkt. Diese Zahl ergibt sich aus 20 (Anzahl der von uns identifizierten Stolpersteine im Text „Jugend") multipliziert mit 26 (Anzahl der Versuchsteilnehmer*innen, die der Gruppe „Nur Text" zugelost wurden). Es konnten somit maximal 520 Stolpersteine überarbeitet werden. 252 davon wurden tatsächlich überarbeitet, wodurch sich der Prozentsatz der überarbeiteten Stolpersteine für diesen Balken (252 / 520*100 = 48,5) ergibt. Für die anderen Balken gelten diese Berechnungen entsprechend.

Es ist zu sehen, dass die Anzahl der überarbeiteten Stolpersteine von der „Nur Text" bis hin zu Bedingung „Markierung+Wörterbuch" hinweg kontinuierlich steigt. Außerdem wurden im „Jugend"-Text konsequent mehr Stolpersteine überarbeitet. Allerdings schrumpft dieser Unterschied zwischen den beiden Texten merklich in der Hilfestellungsbedingung „Markierung+Wörterbuch".

Mit einem Regressionsmodell konnten wir diese Ergebnisse statistisch absichern. Dabei wird der Einfluss von unabhängigen Variablen (hier die Faktoren „Hilfestellung" und „Text") auf abhängige Variablen (an dieser Stelle die Überarbeitung von Stolpersteinen) geprüft. Wir berechneten ein gemischtes logistisches Regressionsmodell in R (R Core Team 2019) mit dem Paket lme4 (Bates et al. 2015). In solch einem Modell können sogenannte Zufallseffekte beachtet werden, um inter-individuelle Unterschiede zwischen Teilnehmer*innen und innerhalb des Stimulusmaterials (hier den Texten) zu kontrollieren. Dieses Modell zeigt uns die Unterschiede

6 Sie können aber Sascha Wolfer unter wolfer@ids-mannheim.de oder die Herausgeber*innen dieses Bandes kontaktieren, wenn Sie einen anonymisierten Auszug der Daten einsehen möchten.

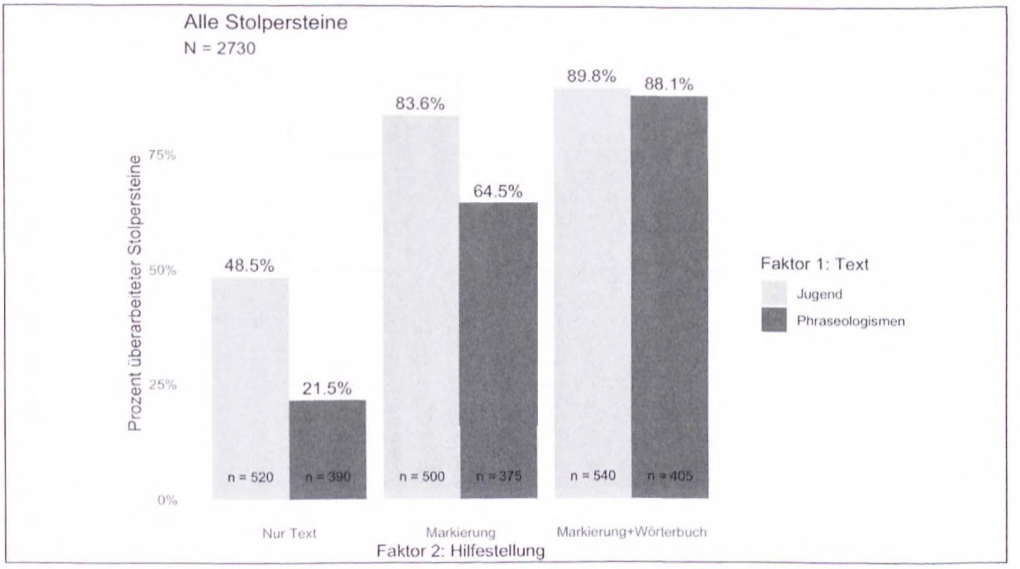


*Abb. 2:* Ergebnisdarstellung für den Anteil überarbeiteter Stolpersteine. N (im Untertitel des Diagramms) gibt die Gesamtanzahl der Stolpersteine an. Die unten in den Balken abgetragenen Stichprobengrößen (n) geben an, wie viele Stolpersteine in der jeweiligen Bedingung maximal überarbeitet werden konnten

zwischen den beiden Texten ($\beta$ = -1,72, SE = 0,38, z = -4,49, $p < 0{,}0001$)[7] sowie zwischen allen Ausprägungen des Hilfestellungsfaktors an (Markierung vs. Nur Text: $\beta$ = 2,50; SE = 0,47; z = 5,35; $p < 0{,}0001$; Markierung+Wörterbuch vs. nur Text: $\beta$ = 3,48; SE = 0,48; z = 7,21; $p < 0{,}0001$; Markierung+Wörterbuch vs. Markierung: $\beta$ = 0,98; SE = 0,49; z = 1,99; p = 0,047). Auch die Interaktion, also die Beobachtung, dass der Unterschied der beiden Texte über die Hilfestellungsbedingungen hinweg schrumpft, ist statistisch bedeutsam ($\beta$ = 1,43; SE = 0,33; z = 4,40; $p <$ 0,0001).

## 4.2 Verbesserungen

Im nächsten Schritt widmen wir uns der Frage, wie viele Stolpersteine in den verschiedenen Versuchsbedingungen nicht nur verändert, sondern tatsächlich verbessert wurden. Selbstverständlich konnten die Teilnehmer*innen nur dann eine problematische Textstelle verbessern, wenn sie diese auch verändert haben. Daher gehen in diese Analyse nur jene Stolpersteine ein, die verändert wurden. Die Gesamtzahl der analysierten Stolpersteine sinkt daher von 2730 auf 1838 Beobachtungen. Wir verzichten hier auf die tabellarische Darstellung, da alle relevanten Informatio-

7 $\beta$ ist der Effektschätzer, der im Regressionsmodell angibt, wie groß der Effekt ist. SE ist der Standardfehler dieses Effektschätzers. z ist die sog. Prüfgröße und zeigt an, wie viel größer der Effekt gegenüber dem Standardfehler ist ($\beta$ / SE = z). Der p-Wert gibt das Signifikanzniveau an. Allgemein wird ein Effekt als signifikant angenommen, wenn $p < 0{,}05$ ist. Bitte beachten Sie aber auch, dass die Berechnung von Signifikanzniveaus in der neueren statistischen Literatur durchaus umstritten ist. Man sollte bei der Interpretation von statistischen Ergebnissen nicht „blind" einem Signifikanzniveau folgen. Siehe hierzu u. a. ein Beitrag von der Online-Seite des Magazins „Spektrum der Wissenschaft": https://www.spektrum.de/news/statistik-wenn-forscher-durch-den-signifikanztest-fallen/1224727. Aus diesem Grund geben wir in diesem Beitrag alle relevanten Größen der statistischen Tests an und betonen die p-Werte nicht übermäßig.

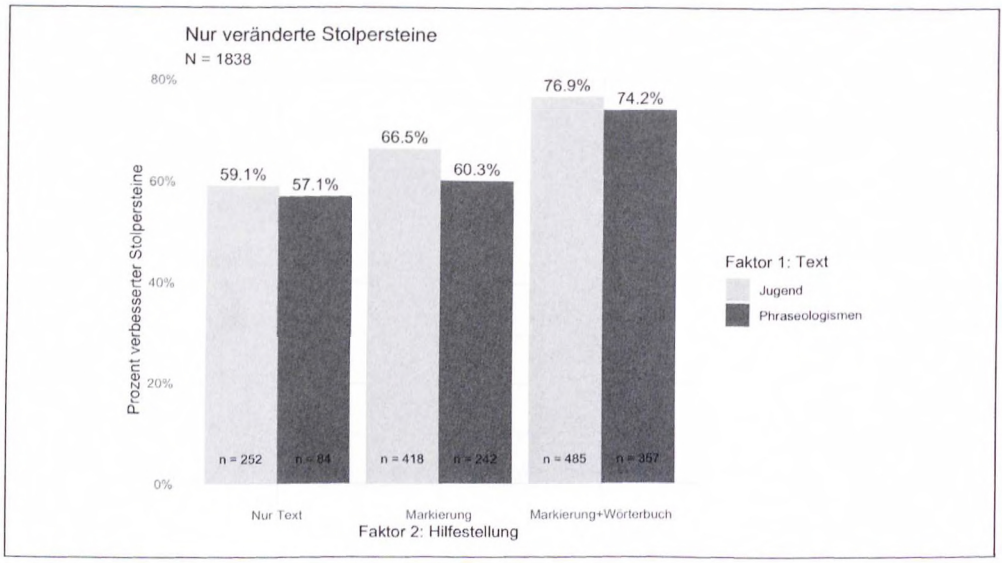


*Abb. 3:* Ergebnisdarstellung für den Anteil verbesserter Stolpersteine. N gibt die Gesamtanzahl der veränderten Stolpersteine an. Die unten in den Balken abgetragenen Stichprobengrößen (n) geben an, wie viele Stolpersteine in der jeweiligen Bedingung maximal verbessert werden konnten

nen auch im Diagramm abzulesen sind (siehe Abbildung 3).

Die Unterschiede zwischen den Hilfestellungsbedingungen sind ähnlich, aber deutlich kleiner als bei den Veränderungen. Es fällt außerdem auf, dass sich die beiden Texte nun kaum noch unterscheiden, obwohl bei den Veränderungen der Text „Jugend" noch deutlich höhere Werte aufwies (zumindest für die Gruppen „Nur Text" und „Markierung"). Der statistische Test, der der gleiche war wie für die Analyse der Veränderungen, zeigt ein leicht unterschiedliches Effektmuster. Die Interaktion, also das Zusammenwirken der beiden Faktoren (hier würde sich eine Interaktion bspw. dadurch abbilden, dass sich der Faktor Hilfestellung in den beiden Texten unterschiedlich auswirkt), bringt keine zusätzliche Information, daher haben wir sie nicht mehr berechnet. Ein deutlicher Unterschied zwischen den Texten kann auch inferenzstatistisch nicht mehr nachgewiesen werden. Einzig die Unterschiede zwischen den verschiedenen Versuchsbedingungen zeigen sich noch immer. Wir können davon ausgehen, dass der Prozentsatz an verbesserten Stolpersteinen in der Bedingung „Markierung+Wörterbuch" höher ist als in der Bedingung „Nur Text" ($\beta = 1{,}06$; $SE = 0{,}17$; $z = 6{,}11$; $p < 0{,}0001$) und in der Bedingung „Markierung" ($\beta = 0{,}74$; $SE = 0{,}15$; $z = 5{,}01$; $p < 0{,}0001$). Der Unterschied zwischen „Markierung" und „Nur Text" bleibt über der gemeinhin angenommenen Signifikanz-Schwelle von $p = 0{,}05$ ($\beta = 0{,}32$; $SE = 0{,}17$; $z = 1{,}87$; $p = 0{,}061$) und sollte daher nur mit äußerster Vorsicht bzw. gar nicht interpretiert werden.

Statistisch gesehen gibt es also keinen Unterschied hinsichtlich der Verbesserung von Stolpersteinen zwischen dem Schüler- und dem Studierenden-Text. Gleichzeitig können wir schließen, dass nur in der Versuchsbedingung mit Hilfsmitteln (Markierung+Wörterbuch) der Prozentsatz verbesserter Stolpersteine höher ist als in beiden anderen Bedingungen.

### 4.3 Semantische Verzerrungen

Wie wir in Abschnitt 3.5. schon erwähnten, haben wir die überarbeiteten Stolpersteine auch daraufhin überprüft, ob die Versuchs-

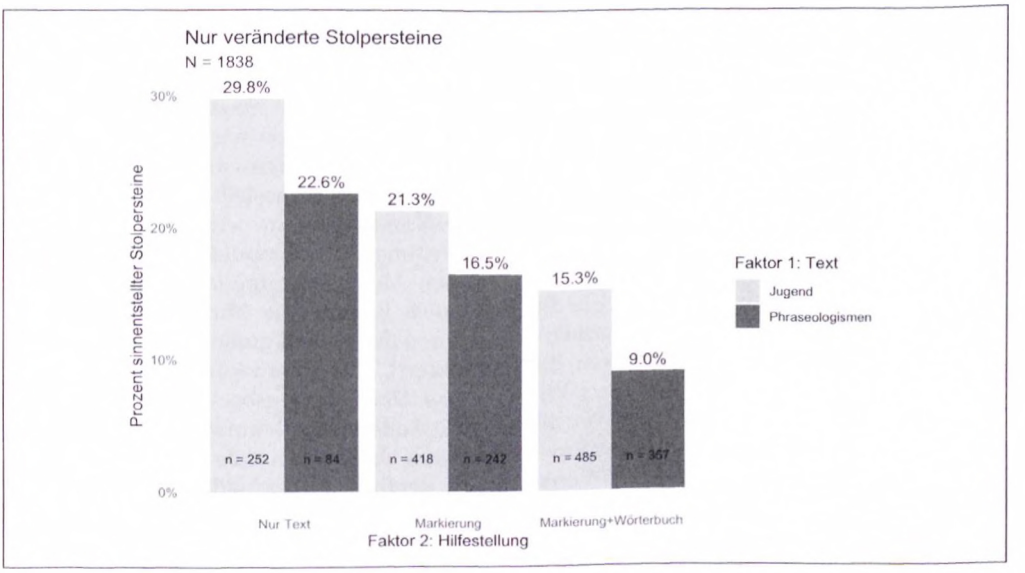


*Abb. 4:* Ergebnisdarstellung für den Anteil sinnentstellter Stolpersteine. Die Bedeutung von N und n ist analog zu den vorherigen Abbildungen

teilnehmer*innen während der Verarbeitung den Sinn der entsprechenden Textstelle veränderten oder verzerrten, d.h. den Text im Grunde ‚verschlimmbesserten' und damit verschlechterten. Tatsächlich haben die Versuchsteilnehmer*innen bei 329 von 1838 Überarbeitungen an Stolpersteinen solche semantischen Verzerrungen vorgenommen, also in 17,9 % aller Fälle. Wenn lexikographische Ressourcen wirklich bei der Textüberarbeitung helfen sollen, dann sollten die Teilnehmer*innen in der Gruppe mit voller Hilfestellung nicht nur mehr Stolpersteine verbessern, sondern auch weniger semantische Verzerrungen durch die Überarbeitungen hervorrufen. Auch hier haben wir nur jene Stolpersteine beachtet, die tatsächlich verändert wurden. Abbildung 4 gibt einen Überblick über die Ergebnisse.

Die statistischen Tests legen keine Interaktion zwischen Text und Hilfestellung nahe. Und obwohl die Teilnehmer*innen im Phraseologismen-Text durchweg weniger semantische Verzerrungen während der Überarbeitung eingebaut haben, ist auch dieser Effekt statistisch nicht signifikant. Allerdings sind alle Unterschiede zwischen Hilfestellungsbedingungen statistisch bedeutsam. Die Mittelwerte für die verschiedenen Bedingungen (über beide Texte hinweg) sind wie folgt. In der „Nur Text"-Bedingung sind 28 % aller Überarbeitungen semantische Verzerrungen des Inhalts. In der Bedingung „Markierung" sind es noch 20 %, während in der Bedingung mit Markierungen und lexikographischen Hilfsmitteln („Markierung+Wörterbuch") nur noch rund 13 % aller Überarbeitungen semantische Verzerrungen hervorrufen.

Alle Unterschiede zwischen den Hilfestellungsbedingungen sind statistisch signifikant (Markierung vs. Nur Text: $\beta = -0{,}57$; $SE = 0{,}21$; $z = -2{,}76$; $p = 0{,}006$; Markierung+Wörterbuch vs. Nur Text: $\beta = -1.30$; $SE = 0{,}21$; $z = -6{,}22$; $p < 0{,}0001$; Markierung+Wörterbuch vs. Markierung: $\beta = -0{,}73$; $SE = 0{,}18$; $z = 4{,}04$, $p < 0{,}0001$). Das überrascht kaum, halbiert sich doch der Anteil an semantischen Verzerrungen von „Nur Text"- zu „Markierung+Wörterbuch"-Bedingung.

## 4.4 Punkte-basierte Auswertung

In allen zuvor vorgestellten Analysen haben wir einen bearbeiteten Stolperstein als einen Fall, also als eine Zeile in unserem Datensatz, behandelt. Wir wollen die Perspektive nun etwas verändern und die Teilnehmer*innen der Untersuchung noch direkter in den Blick nehmen. Die Teilnehmer*innen waren natürlich bisher auch bereits in der Untersuchung enthalten, weil sie diejenigen waren, die die Stolpersteine überarbeitet haben. Die im Folgenden vorgestellten Analysen rücken die Teilnehmer*innen direkter in den Fokus. Wir werden hierzu eine Analyse durchführen, die sozusagen ein Scoring-System für die einzelnen Teilnehmer*innen bildet, d. h. auf Punkten basiert. Für jeden Stolperstein, den eine Person verbesserte, wurde ein Punkt vergeben. Für jeden Stolperstein, der verschlechtert oder semantisch verzerrt wurde, wurde hingegen ein Punkt abgezogen. Die Einzelpersonen in den Fokus zu rücken ist auch deshalb sinnvoll, weil natürlich auch immer einzelne Personen einen Text schreiben. Insofern rückt es die Analysen näher an eine alltägliche Schreibsituation, wenn die einzelnen Textüberarbeitungen nicht nur als Einzelfälle betrachtet werden, sondern nach teilnehmenden Personen gruppiert werden.

Jede*r Teilnehmer*in konnte maximal 35 Stolpersteine überarbeiten (20 aus dem „Jugend"-Text und 15 aus dem „Phraseologismen"-Text). Die Maximalpunktzahl von 35 Punkten bekam ein*e Teilnehmer*in somit, wenn sie*er alle Stolpersteine verbessert hat. Die Minimalpunktzahl beträgt -35, die dadurch zustande käme, wenn ein*e Teilnehmer*in alle Stolpersteine verändert, aber alle dabei verschlechtert hätte. Soviel vorweg: Diese Extremwerte kamen nicht vor. Eine Punktzahl von 0 kann mehrere Dinge bedeuten: Eine Person, die keinen Stolperstein bearbeitet, kann auch nichts falsch machen – die Folge wäre eine Punktzahl von 0. Das gleiche gilt für Teilnehmer*innen, die bspw. zwölf Stolpersteine bearbeitet haben und fünf davon verbesserten, fünf davon verschlechterten sowie zwei nicht in der Qualität veränderten. Wir haben dieses Maß entwickelt, um (Gruppen von) Teilnehmer*innen untereinander vergleichbar zu machen und gleichzeitig alle Überarbeitungen, die sie vorgenommen haben, zu beachten.

Die Punkte der einzelnen Versuchsteilnehmer*innen haben wir in Abbildung 5 in Form eines „Bienenschwarm-Diagramms" (*beeswarm plot*) dargestellt. Mit einem Bienenschwarm-Diagramm wird die tatsächliche Verteilung der Datenpunkte in den verschiedenen Versuchsbedingungen sichtbar. Zusätzlich können die Mittelwerte abgelesen werden (hier durch große graue Punkte symbolisiert). Man sieht zwar einerseits, dass es große Überlappungsbereiche der Gruppen gibt. Andererseits können Sie aber auch erkennen, dass die zentrale Tendenz der Gruppen deutlich unterschiedlich ist. Am besten schneiden im Durchschnitt die Teilnehmer*innen aus der Gruppe „Markierung+Wörterbuch" ab (18,6 Punkte). Die Gruppe, die zwar hervorgehobene Stolpersteine aber keine lexikographischen Ressourcen als Hilfestellung bekam liegt mit einem Mittelwert von 10,4 Punkten in der Mitte. Am schlechtesten schneidet die „Nur Text"-Gruppe ab (Mittelwert von 3,6 Punkten). In dieser Gruppe waren auch die einzigen Teilnehmer*innen, die im negativen Bereich abschnitten (zweimal -3 Punkte und einmal -4 Punkte).

Eine Nebenbemerkung zu den Arten der Visualisierung, die wir in Abbildung 5 gewählt haben: Die Höhe der Säule im rechten Diagramm steht für den Mittelwert und die Fehlerbalken für Standardfehler oder Konfidenzintervalle (hier 1 Standardfehler). An dieser Visualisierung ist aus rechnerischer Sicht auch nichts auszusetzen. Allerdings haben Sie zwei konzeptionelle Eigenschaften, die in wissenschaftlicher Hinsicht bedenkenswert sind: 1. Oft werden die Unterschiede zwischen Gruppen in solchen Diagrammen visuell überbewertet. Die Überlappungsbereiche zwischen experimentellen Gruppen, die praktisch immer vorhanden sind, treten ziemlich in den Hintergrund. 2. Die Rezipient*innen dieser Diagramme können nicht einschätzen, wie ein bestimmter Mittelwert zustande kommt. Handelt es sich um breit gestreute Messwerte oder variieren die einzelnen Messwerte sehr dicht um den Mittelwert? Trennt sich die Gruppe evtl. gar in zwei Un-

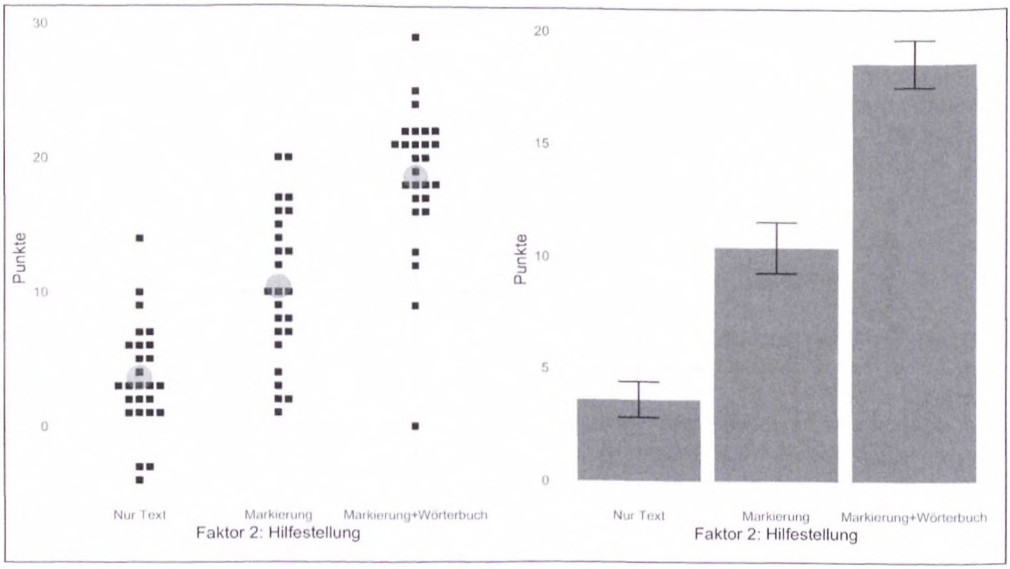


*Abb. 5:* „Bienenschwarm-Diagramm" (links) und Säulendiagramm mit Fehlerbalken (rechts) der Punkteverteilungen der Teilnehmer*innen für die drei Hilfestellungsbedingungen. Jedes schwarze Quadrat im linken Diagramm steht für eine*n Teilnehmer*in, die großen grauen Punkte symbolisieren die Mittelwerte der drei Gruppen. Die Fehlerbalken im rechten Diagramm symbolisieren 1 Standardfehler. Wenn sich hier zwei Fehlerbalken nicht überlappen, deutet das auf einen signifikanten Unterschied zwischen den Gruppen hin

tergruppen, wo weitere Untersuchungen interessant wären, woher diese Spaltung kommt? Das Bienenschwarm-Diagramm mag in der Forschungscommunity zwar noch nicht sehr verbreitet sein, aber es löst diese beiden Probleme, indem jeder einzelne Messwert sichtbar wird (siehe hierzu das „erste Gesetz" der Visualisierung von Daten von Tufte (2001: 92): „Above all else show the data").

Zurück zur aktuellen Fragestellung: Die Teilnehmer*innen aus der Experimentalgruppe mit vollen Hilfestellungen zeigten signifikant bessere Leistungen als die Personen aus den anderen Gruppen. Jedoch: Ihnen mag der Ausreißer bzw. die Ausreißerin aufgefallen sein, die/der in der Gruppe „Markierung+Wörterbuch" 0 Punkte „erreicht" hat. Diese Person hat keinerlei Stolpersteine überarbeitet und war eine der schnellsten Personen bei der Bearbeitung des Experiments. Dies legt nahe, dass sie*er nicht versucht hat, die Texte zu überarbeiten, sondern nur darauf wartete, dass die Experimentalsitzung zu Ende geht. Trotzdem hat sie*er insgesamt länger als fünf Minuten die Texte betrachtet und wurde daher nicht aus der Stichprobe ausgeschlossen. Doch auch mit diesem Ausreißer war die Gruppe, die lexikographische Hilfsmittel zur Lösung der Aufgabe bekam, am besten.

## 4.5 Effizienz

Es wurde aus den vorgehenden Analysen bereits klar, dass die Gruppe mit lexikographischen Hilfsmitteln einen Vorteil beim Bearbeiten der Aufgabe hatte: In dieser Gruppe wurden die meisten Stolpersteine bearbeitet und verbessert. Außerdem wurden in dieser Gruppe die wenigsten semantischen Verzerrungen eingebaut. Berechnet man daraus einen personenbezogenen Punktestand, liegen die Mitglieder dieser Gruppe ebenfalls vor den beiden anderen Gruppen. Wir wollen nun aber ein noch strengeres Kriterium anset-

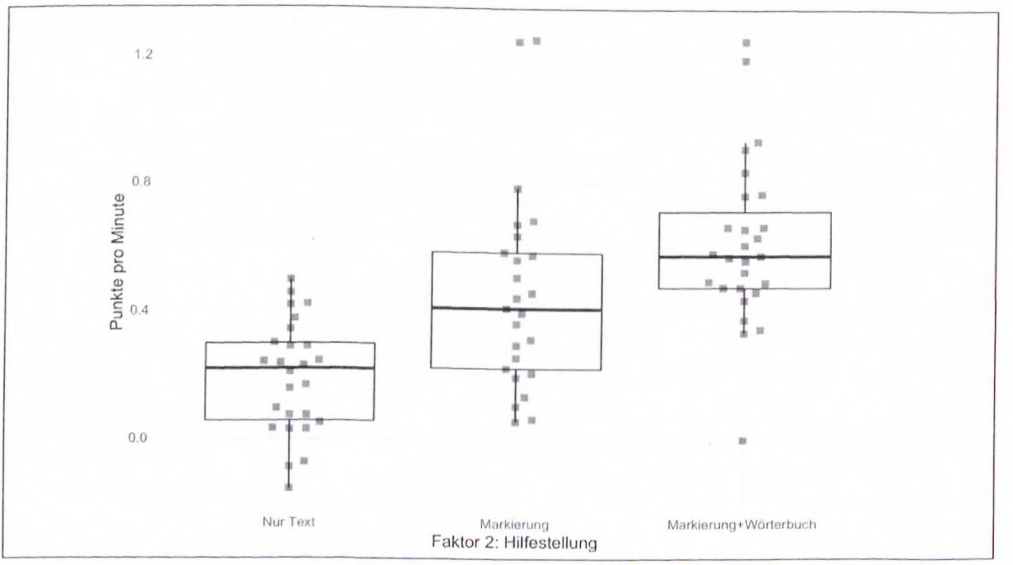


*Abb. 6:* Kombination aus Boxplot und Bienenschwarm-Diagramm für den Vergleich der Effizienz (Punkte pro Minute) über die verschiedenen Hilfestellungsbedingungen hinweg. Jedes Quadrat steht für eine/n Teilnehmer*in. In jeder Box befinden sich die mittleren 50 % der Datenpunkte in der jeweiligen Bedingung. Die Grenzen der Boxen stehen jeweils für das 25 % und 75 % Perzentil (d.h. bspw. für das 75. Perzentil, dass 75 % Prozent aller Datenpunkte niedriger sind als diese Grenze). Die Mittellinien der Boxen geben den jeweiligen Median an. Das ist der Wert, der die Datenmenge in zwei Hälften teilt. Der Median ist, ebenso wie der Mittelwert, ein Wert, der die zentrale Tendenz einer Verteilung beschreibt[8]

zen, um die Gruppen zu vergleichen, nämlich die Effizienz bei der Bearbeitung der Aufgabe. Mit Effizienz meinen wir in diesem Zusammenhang den Erfolg in Zusammenschau mit der benötigten Zeit – oder quantitativ gefasst: Punkte pro Minute. Das ist deshalb besonders interessant, weil es nicht ganz abwegig erscheint, dass die Teilnehmer*innen in der Bedingung „Markierung+Wörterbuch" tatsächlich mehr Zeit bei der Bearbeitung der Aufgabe benötigen, denn sie müssen nicht nur die Hervorhebungen in den Texten verarbeiten, sondern zusätzlich mit den lexikographischen Hilfsmitteln umgehen. Es könnte ja tatsächlich sein, dass der Vorteil, den die lexikographischen Ressourcen bieten, dadurch „aufgefressen" wird, dass die Befragten in dieser Versuchsbedingung viel länger bei der Bearbeitung der Texte brauchen. Wenn man zeitlicher Effizienz einen Stellenwert einräumt, wäre das ein Problem in unserer Argumentation.

Als kurze Randbemerkung: Unser Untersuchungsgegenstand ist die Effizienz des Einsatzes von Hilfsmitteln bei der Textüberarbeitung. In der „wahren Welt", d.h. im Schreiballtag, ist Zeit normalerweise knapp. Deshalb wäre es zwar ein interessantes Resul-

8 Eine Anmerkung noch zu den in diesem Beitrag verwendeten Abbildungen: Normalerweise entscheidet man sich in einem wissenschaftlichen Artikel für eine Diagrammform, wenn gleiche Arten von Daten visualisiert werden. Bspw. würde man sich bei Abbildung 5 und Abbildung 6 für eine Darstellungsart entscheiden, da in beiden Diagrammen eine kategoriale unabhängige Variable (die Gruppeneinteilung) mit einer kontinuierlichen abhängigen Variable (Punkte und Punkte pro Zeit) kombiniert wird. Wir haben in diesem Beitrag aus didaktischen Gründen verschiedene Arten von Diagrammen eingeführt, damit Sie möglichst viele unterschiedliche Arten von Diagrammen kennenlernen.

tat, wenn wir messen können, dass Hilfsmittel die Textüberarbeitung verbessern. Wenn wir jedoch gleichzeitig feststellen würden, dass Hilfsmittel zwar insgesamt gut sind, bei einer Messung pro Bearbeitungsminute jedoch die Markierungs-Bedingung besser abschneidet, könnte das für den Schreiballtag auch bedeuten, dass man eine Schreibumgebung nur mit Markierungen einer mit Hilfsmittel-Unterstützung – zumindest im normalen Alltag – vorziehen würde, weil man zwar nicht ganz so gut, aber auf jeden Fall schneller ist.

Tatsächlich können wir zeigen, dass Teilnehmer*innen in der vollen Hilfestellungsbedingung am längsten an den Texten arbeiteten (Mittelwert: 31,6 Minuten), gefolgt von der Gruppe, die nur hervorgehobene Textstellen zur Hilfe nehmen konnten (26,9 Minuten). Die „Nur Text"-Gruppe war durchschnittlich am schnellsten (24,8 Minuten). Wie oben beschrieben, müssen diese Werte aber – um die tatsächliche Effizienz zu messen – mit der erreichten Punktzahl in Beziehung gesetzt werden. Wir messen damit Punkte pro Minute.

Abbildung 6 gibt einen weiteren Einblick in die Ergebnisse. Wieder sind deutliche Überschneidungsbereiche der verschiedenen Datenmengen erkennbar, aber man erkennt eben auch, dass sich die zentralen Tendenzen (in Abbildung 6 gefasst über die Boxplots, die uns den Median[9] und die mittleren 50 % der Datenpunkte in der jeweiligen Bedingung zeigen) deutlich voneinander unterscheiden. In der Hilfsmittelbedingung liegen die Punkte pro Minute deutlich über denen der Markierungsbedingung, die wiederum deutlich höher liegen als in der Nur-Text-Bedingung. Dieser visuelle Eindruck kann auch über einen statistischen Vergleich der Mittelwerte (nur Text: 0,19 Punkte pro Minute; Markierung: 0,46; Markierung+Wörterbuch: 0,62) abgesichert werden. In diesem Fall können wir bspw. einen t-Test für multiple Vergleiche mit der Holm-Korrektur[10] berechnen, der in allen Fällen signifikante Unterschiede zwischen den Gruppen anzeigt (Nur Text vs. Markierung: $p = 0{,}0007$; Nur Text vs. Markierung+Wörterbuch: $p < 0{,}0001$; Markierung vs. Markierung+Wörterbuch: $p = 0{,}028$).

## 5. Methodische Reflexion

Bevor wir auf unsere Forschungsfrage zurückkommen, möchten wir noch auf zwei Dinge hinweisen, nämlich die Untersuchungssituation und die beobachteten Unterschiede zwischen den beiden bearbeiteten Texten.

In Abschnitt 3.4 haben wir beschrieben, dass alle Teilnehmer*innen gleichzeitig in zwei großen Hörsälen der Universität Mannheim das Experiment bearbeiteten. Dies ist im Vergleich zu „klassischen" experimentellen Studien eine außergewöhnliche Situation, denn das prototypische Experiment findet meist in einem Labor statt, in das die Studienteilnehmer*innen einzeln eingeladen werden, um eine bestimmte Aufgabe zu bearbeiten. Das erleichtert es den Forschenden, Störvariablen, die der Umgebung entstammen, gezielt zu kontrollieren bzw. auszuschalten. Dazu gehören bspw. Lärm, Temperaturunterschiede, Lichtverhältnisse, technische Voraussetzungen des Geräts, mit dem das Experiment durchgeführt wird usw. Insofern war die Studie in den Hörsälen ein gewisses Wagnis. Gerade der Fakt, dass die Teilnehmer*innen die Studie auf ihren eigenen mitgebrachten Geräten bearbeiten sollten, stellte ein Risiko dar: Es hätte bspw. sein können, dass technische Probleme zu einem Zusammenbruch der drahtlosen Netzwerkverbindung führen, dass das Experiment auf einzelnen Rechnern nicht dargestellt werden kann oder dass ganz andere, unvorhersehbare Ereignisse die Durchführung erschweren. Diese Faktoren konnten wir nur durch zeitin-

9 Der Median teilt die vorhandenen Datenpunkte in zwei Hälften. Über und unter dem Medien befinden sich also jeweils 50% aller Messwerte.

10 Mit einem t-Test werden immer zwei Gruppen miteinander verglichen. Da jeder Einzeltest mit einer bestimmten Irrtumswahrscheinlichkeit belegt ist, muss man bei multiplen Vergleichen (bei drei Gruppen finden drei Vergleiche statt) eine Korrektur vornehmen.

tensive Vorbereitungen zumindest teilweise kontrollieren. So haben wir bspw. das Experiment auf einer Vielzahl unterschiedlicher Geräte (Betriebssysteme, Browser und Gerätetypen wie Smartphones, Tablets und Laptops) getestet und das Rechenzentrum darum gebeten, für einen reibungslosen Ablauf zu sorgen. Umgebungsvariablen wie Ablenkung durch Lärm oder andere Teilnehmer*innen konnten wir nur durch viel Personal versuchen aufzufangen. Es kam tatsächlich zu keinem Datenverlust durch technische Schwierigkeiten oder zu größeren Ablenkungen, die wir nicht kontrollieren konnten. Der Aufwand hat sich insofern gelohnt und das Wagnis der Untersuchungssituation stellt sich im Rückblick als nicht zu riskant dar.

Die zweite Anmerkung betrifft die Unterschiede zwischen den beiden Texten, die zwar nicht unmittelbar relevant für unsere ursprüngliche Forschungsfrage sind, aber trotzdem interessante Einsichten gewährt. Nehmen wir einmal an, dass der Schüler*innentext zum Thema „Jugend" sprachlich und inhaltlich weniger komplex ist als der studentische Text zum Thema „Phraseologismen" – eine Annahme, die u.E. durchaus gerechtfertigt ist. Wie können wir dann die Unterschiede zwischen den Texten bezüglich unserer abhängigen Variablen interpretieren? Die Stolpersteine im „Jugend"-Text wurden von der Teilnehmer*innen-Gruppe, die nur den Text ohne jegliche Hilfsmittel dargeboten bekam, häufiger bearbeitet als die Stolpersteine im „Phraseologismen"-Text. Dieser Unterschied bestand auch in der Gruppe, für die die Stolpersteine hervorgehoben waren, nicht jedoch in der Gruppe, die zusätzlich die lexikographischen Ressourcen als Hilfestellung bekamen. Interessanterweise schrumpfen bzw. verschwinden diese Unterschiede zwischen den Texten, wenn man die Verbesserungen und Sinnentstellungen betrachtet. Dieses Muster lässt sich so interpretieren, dass die Teilnehmer*innen ohne lexikographische Hilfsmittel eher zögerlich waren, den sprachlich und inhaltlich komplexeren Text zu überarbeiten. Erst, wenn man sprachliche Hilfsmittel hinzuziehen kann, sinkt die Hemmschwelle so weit, dass man sich auch traut, komplexere Texte „anzugehen". Auch zur Förderung von Textüberarbeitungskompetenzen scheinen Hilfsmittel also gut eingesetzt werden zu können. Inwieweit diese Interpretation trägt, lässt sich noch nicht abschließend beantworten – es sind hier zusätzliche Studien notwendig, die direkt auf diese Frage ausgerichtet sein müssten.

Nun zu einer Frage, die wir mit der vorliegenden Studie ziemlich deutlich beantworten können – nämlich die Forschungsfrage, von der wir zu Beginn dieses Beitrags ausgingen: Helfen lexikographische Ressourcen bei der Überarbeitung von Texten? Betrachten wir alle abhängigen Variablen gemeinsam, können wir eine Hierarchie der verschiedenen Gruppen von Versuchsteilnehmer*innen annehmen. Die Performanz der „Nur Text"-Gruppe bei der Überarbeitung der Texte war geringer als jene der Hervorhebungsgruppe. Das Hinzufügen von lexikographischen Ressourcen(Gruppe„Markierung+Wörterbuch") sorgte dann nochmals für einen Anstieg bei der Überarbeitungsleistung. Dafür spricht die Auswertung jeder einzelnen abhängigen Variable: In der Gruppe mit beiden Hilfestellungen (also Hervorhebungen kombiniert mit lexikographischen Ressourcen) wurden mehr Stolpersteine überarbeitet. Von diesen bearbeiteten Stolpersteinen wurden mehr verbessert und es wurden weniger semantische Verzerrungen dabei eingefügt. Darüber hinaus erreichten die Versuchsteilnehmer*innen in dieser Gruppe mehr Punkte und waren auch effizienter als die Teilnehmer*innen in den anderen Gruppen. Wir können unsere Forschungsfrage also beantworten: Ja, lexikographische Ressourcen helfen tatsächlich bei der Überarbeitung von Texten. Im Folgenden möchten wir allerdings noch auf einen Umstand hinweisen, der wichtig ist, um die Relevanz dieser Antwort einzuschätzen.

Der wohl wichtigste Punkt ist, dass wir unseren Teilnehmer*innen einen ganz entscheidenden Schritt abgenommen haben, nämlich das *Auffinden* der relevanten Information in den Ressourcen. Wir haben die Hilfsmittel, die bei der Lösung der Probleme helfen konnten, praktisch „auf dem Silbertablett serviert", indem wir sie direkt neben den Text gestellt haben und mit den Stolpersteinen verknüpf-

ten. Das haben wir bewusst getan, denn wir wollten ein Szenario schaffen, in dem wir uns ausschließlich auf den Effekt des Vorhandenseins von lexikographischer Information konzentrieren konnten. Wir wollten diesen Effekt nicht mit anderen Faktoren vermischen wie bspw. dem Suchen der relevanten Information in Nachschlagewerken und der nötigen Verbindung mit problematischen Stellen. Das ist ein ganz entscheidender Schritt: Denn selbstverständlich kann nur jene Information gewinnbringend eingesetzt werden, die auch gefunden wird. Daher sprechen die Ergebnisse unserer Studie u.E. dafür, dass sich nicht nur das Erstellen von lexikographischen Ressourcen lohnt, sondern auch, dass es sich lohnt, Menschen im Umgang mit diesen Ressourcen zu schulen. Denn auch die besten Wörterbücher, Übersetzungsprogramme oder Grammatiken helfen nicht bei der Lösung von sprachlichen Problemen, wenn man die darin enthaltene Information nicht findet und auf das konkrete sprachliche Problem in einer konkreten Situation übertragen kann.

Eine logische Weiterentwicklung der Studie, die wir hier vorgestellt haben, besteht darin, Menschen vor ein konkretes sprachliches Problem zu stellen, ihnen dabei aber nicht die relevante Information praktisch „verzehrbereit" vorzusetzen, sondern zu sehen, ob und wie die Teilnehmer*innen die Information selbst finden und verarbeiten können. Eine solche Studie haben wir in einem anderen Kontext, nämlich mit Deutschlernenden aus dem romanischen Sprachraum durchgeführt (Müller-Spitzer/Nied Curcio/Domínguez Vázquez/Dias/Wolfer, 2018; 2019). Den Lernenden haben wir deutsche Sätze mit Interferenzfehlern[11] aus romanischen Sprachen vorgegeben, die sie korrigieren sollten. Auch die Methode der wissenschaftlichen Herangehensweise haben wir in dieser Studie variiert. Wir setzten dort keine experimentelle Variation ein, wie wir sie hier dargestellt haben, sondern konzipierten eine Beobachtungsstudie mit qualitativen Elementen in der Auswertung.

Sie mögen sich fragen, wie relevant eine Studie noch ist, die sich hauptsächlich mit Wörterbüchern und der Art von Information beschäftigt, die sich darin finden, da Schreibprozesse in der Zukunft stärker automatisch unterstützt werden können. Allerdings darf dabei nicht vergessen werden, dass der Bedarf an sorgfältig erarbeiteten lexikographischen Ressourcen allgemein ungebrochen ist. Viele Systeme zur automatischen Verarbeitung von natürlicher Sprache verlassen sich auf lexikographisch aufbereitete Datenbanken, um diese Information bei der Verarbeitung von Sprache einzubeziehen. Auch computerbasierte Systeme, die Menschen bei der Lösung von sprachlichen Aufgaben helfen sollen, nutzen in großem Stil lexikographische Information. Den Benutzer*innen wird das aber häufig nicht bewusst, weil die Information eben nicht mehr in einem Format aufbereitet ist, das sie von Wörterbüchern kennen. Das offensichtlichste Beispiel ist vielleicht noch, dass Google auf der Ergebnisseite der Suchen in manchen Fällen Auszüge aus Online-Wörterbüchern präsentiert (zumindest zum Zeitpunkt, zu dem wir diesen Beitrag verfassten). Ein Ausgangspunkt der Studie war, wie anfangs skizziert, auch die Idee, eine Art computergestützte Schreibumgebung zu entwickeln, die den Schreiber*innen automatisch relevante Ressourcen zur Verfügung stellt, wenn ein NLP-Algorithmus[12] Probleme im verfassten Text feststellt. So fern dieses Ziel einer automatischen Schreibumgebung auch noch sein mag: Unsere Studie hat gezeigt, dass Schreiber*innen wohl von einem solchem System profitieren würden.

---

11 Interferenzfehler bezeichnen sprachliche Fehler, die bei der unzulässigen Übertragung von sprachlichen Eigenschaften (Semantik/Morphologie/Syntax) aus einer Sprache in eine andere entstehen. Ein Beispiel ist der Satz „Obwohl er sich beeilt hat, hat er die U-Bahn *verloren*", wo eine unzulässige Übertragung von bspw. ital. „perdere" auf dt. „verpassen" stattfindet.

12 NLP steht für *natural language processing*.

## Zum Weiterlesen

Wer sich über die Visualisierung von linguistischen Daten informieren möchte, kann dies im Sammelband (Open Access) von Bubenhofer & Kupietz (2018) tun. Wolfer & Hansen-Morath (2017) geben in einem Online-Tutorial einen Überblick über einige Visualisierungsmöglichkeiten innerhalb der Statistikumgebung *R*.

Die inferenzstatistischen Modelle, die wir in diesem Beitrag verwendeten, werden von Baayen (2008, insb. Kapitel 7) und Winter (2020, insb. Kapitel 14) vorgestellt. Beide Bücher enthalten auch umfassende und verständliche Einführungen in R sowie deskriptive Verfahren der Statistik.

## Literatur


Abel, Andrea / Aivars Glaznieks / Lionel Nicolas / Egon Stemle (2014): KoKo: An L1 Learner Corpus for German, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, S. 2414–2421.

Baayen, Harald R. (2008): *Analyzing linguistic data. A practical introduction to statistics using R*, Cambridge: Cambridge University Press.

Bates, Douglas / Martin Maechler / Ben Bolker / Steve Walker (2015): Fitting Linear Mixed-Effects Models Using lme4, in: *Journal of Statistical Software*, Jg. 67(1), S. 1–48.

Bubenhofer, Noah / Marc Kupietz (Hrsg.) (2018): *Visualisierung sprachlicher Daten: Visual Linguistics – Praxis – Tools*, Heidelberg: Heidelberg University Publishing, [online] https://heiup.uni-heidelberg.de/heiup/catalog/book/345.

Müller-Spitzer, Carolin / Martina Nied Curcio / María José Domínguez Vázquez / Idalete Maria Silva Dias / Sascha Wolfer (2018): Correct hypotheses and careful reading are essential: results of an observational study on learners using online language resources, in: *Lexikos*, Bd. 28, S. 287–315.

Müller-Spitzer, Carolin / Martina Nied Curcio / María José Domínguez Vázquez / Idalete Maria Silva Dias / Sascha Wolfer (2019): Rechercheprаxis bei der Verbesserung von Interferenzfehlern aus dem Italienischen, Portugiesischen und Spanischen: Eine explorative Beobachtungsstudie mit DaF-Lernenden, in: *Lexicographica*, Bd. 34, Berlin / Boston: de Gruyter, S. 157–182.

R Core Team (2019): R: A language and environment for statistical computing, Vienna: R Foundation for Statistical Computing, [online] https://www.R-project.org/.

Tufte, Edward (2001): *The visual display of quantitative information*, 2. Aufl., Cheshire: Graphics Press.

Winter, Bodo (2020): *Statistics for linguists: An introduction using R*, New York / London: Routledge.

Wolfer, Sascha / Sandra Hansen-Morath (2017): Visualisierung linguistischer Daten mit der freien Grafik- und Statistikumgebung R, [online] http://kograno.ids-mannheim.de/VisR-OnlinePub.

Wolfer, Sascha / Thomas Bartz / Tassja Weber / Andrea Abel / Christian M. Meyer / Carolin Müller-Spitzer / Angelika Storrer (2018): The effectiveness of lexicographic tools for optimising written L1-texts, in: *International Journal of Lexicography*, Jg. 31 H. 1, S. 1–28.


Die Adressen aller Webseiten und Online-Ressourcen in diesem Beitrag wurden zuletzt auf Aktualität überprüft am 6. April 2021.