ZfdG

Zeitschrift für digitale Geisteswissenschaften

Projektvorstellung aus:
Zeitschrift für digitale Geisteswissenschaften, Heft 11 (2026)

Titel:
R für Philolog*innen: Ein praktischer Einstieg mit politischen Reden

Autor*in:
Ángeles González-Miguel

Kontakt: angeles.gonzalez.miguel@uva.es
Institution: Universität Valladolid
GND: 1416980601 ORCID: 0000-0002-5989-0680
Contribution (CRediT): Conceptualization | Methodology | Software | Writing – original draft | Writing – review & editing

Autor*in:
Francisco Javier Muñoz-Acebes

Kontakt: fjavier.munoz@uva.es
Institution: Universität Valladolid
GND: 1067645519 ORCID: 0000-0001-6527-203X
Contribution (CRediT): Conceptualization | Methodology | Software | Writing – original draft | Writing – review & editing

DOI des Beitrags:
10.17175/2026_012

Nachweis im OPAC der Herzog August Bibliothek:
1982617446

Erstveröffentlichung:
15.09.2026

Letzte Überprüfung aller Verweise:
28.08.2026

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Computerlinguistik | Textanalyse | R | Bundespräsident (Deutschland)

Empfohlene Zitierweise:
Ángeles González-Miguel / Francisco Javier Muñoz-Acebes: R für Philolog*innen: Ein praktischer Einstieg mit politischen Reden. In: Zeitschrift für digitale Geisteswissenschaften 11 (2026). 15.09.2026. HTML / XML / PDF. DOI: 10.17175/2026_012

Ángeles González-Miguel / Francisco Javier Muñoz-Acebes

R für Philolog*innen: Ein praktischer Einstieg mit politischen Reden

Abstract

Der vorliegende Beitrag stellt ein praxisorientiertes R-Skript für Philologiestudierende ohne Programmiervorkenntnisse vor. Anhand des *German Political Speeches Corpus* – einer XML-Sammlung der Reden deutscher Bundespräsidenten – werden grundlegende Methoden der korpusgestützten Analyse eingeführt: *Tokenisierung*, *Frequenzanalyse*, *Bigramme*, *lexikalische Vielfalt* und *diachrone Schlüsselwortanalyse*. Das Tutorial wurde im Rahmen eines Seminars zur digitalen Philologie an der Universidad de Valladolid erprobt und folgt den Prinzipien des projektbasierten Lernens.

This paper presents a practice-oriented R script for philology students without prior programming experience. Using the *German Political Speeches Corpus* – an XML collection of speeches by German Federal Presidents – it introduces fundamental corpus-based methods: *tokenisation*, *frequency analysis*, *bigrams*, *lexical diversity*, and *diachronic keyword analysis*. Tested in a digital philology seminar at the Universidad de Valladolid, it follows the principles of project-based learning.

1. Einführung

Die vorliegende Publikation ist als praktisches Beispiel für den Einsatz von R in der text- und datenbasierten Analyse konzipiert.[1] Sie richtet sich an Studierende mit geringen oder keinen Vorkenntnissen in R: Alle Schritte werden im Skript kommentiert, so dass der Code Zeile für Zeile nachvollzogen werden kann. Das hier präsentierte Skript wurde im Rahmen von Seminaren zur digitalen Philologie mit Studierenden der Germanistik an der Universidad de Valladolid (Spanien) eingesetzt. Ziel ist es, an einem konkreten Beispiel – den Reden der deutschen Bundespräsidenten – grundlegende Arbeitsweisen der korpusgestützten Analyse (Datenaufbereitung, quantitative Untersuchungen, Visualisierungen, einfache lexikalische Auswertungen) kennenzulernen. [1]

Das Material versteht sich als Vorlage: Die verwendeten Dateien und Einstellungen können leicht angepasst werden (z. B. an andere Korpora, andere Schlüsselwörter, weitere Visualisierungen), so dass das Skript in unterschiedlichen Lehrkontexten wiederverwendet und erweitert werden kann. [2]

1.1 Kontext: R in den Digital Humanities

Der Einsatz von Programmiersprachen wie R in der philologischen Forschung stellt einen bedeutenden methodischen Wandel dar.[2] Traditionell wurde die Textanalyse durch genaues Lesen (*close reading*) durchgeführt; heute ermöglichen digitale Werkzeuge auch die Analyse aus der Distanz (*distant reading*, Konzept von Franco Moretti[3]), d. h. die Untersuchung großer Textmengen durch quantitative Methoden und Datenvisualisierung. [3]

R ist für diese Art von Analyse besonders geeignet, weil es [4]

[1] Generative KI-Werkzeuge kamen in diesem Beitrag in zweierlei Hinsicht zum Einsatz: Zum einen setzten die Studierenden im Rahmen des Seminars den KI-Assistenten Claude (Anthropic) unterstützend beim Erlernen von R ein; zum anderen griffen auch die Verfasser*innen bei der Erstellung des begleitenden R-Skripts auf Claude (Anthropic, Sonnet-Modellreihe) zur Fehlersuche und Kommentierung einzelner Codeabschnitte zurück. Die konzeptionelle und methodische Gestaltung der Analyse erfolgte eigenständig durch die Autor*innen.

[2] Vgl. Arnold et al. 2019.

[3] Vgl. Moretti 2013.

- kostenlos und Open Source ist,
- eine sehr aktive wissenschaftliche Community hat,
- spezialisierte Pakete für Textanalyse bietet (z. B. tidytext, quanteda, stylo),
- Reproduzierbarkeit ermöglicht: andere Forscher*innen können unsere Analysen nachvollziehen;
- in der Entwicklungsumgebung RStudio eine vollständig integrierte Arbeitsumgebung bietet: Skript-Editor, Konsole, Datenobjekte und Visualisierungen sind in einem einzigen Fenster zugänglich, ohne dass externe Konfigurationen oder zusätzliche Werkzeuge erforderlich sind. Dies unterscheidet R von Arbeitsumgebungen wie Python, wo vergleichbare Workflows typischerweise mehrere Komponenten voraussetzen (Interpreter, Paketverwaltung, Jupyter Notebook oder Ähnliches), deren Einrichtung für Studierende ohne Programmiererfahrung eine Einstiegshürde darstellt.

[5] Das vorgestellte Skript steht in der Tradition der Korpuslinguistik, angewandt auf politische Texte, und kombiniert quantitative Methoden mit qualitativer Interpretation – ein Vorgehen, das als methodisch notwendig gilt, wenn computergestützte Verfahren hermeneutisch fundiert und für die Zieldisziplin transparent bleiben sollen.[4] Beispiele für eine solche Verbindung im Kontext deutschsprachiger politischer Korpora finden sich etwa in diachronen Frequenz- und Themenanalysen, die quantitative Befunde konsequent mit inhaltlicher Interpretation verknüpfen.[5]

1.2 Stand der Forschung: R-Tutorials in den Digital Humanities

[6] Die verfügbaren Lehrressourcen für den Einsatz von R in geisteswissenschaftlichen Curricula lassen sich grob in zwei Kategorien einteilen: technisch orientierte Einführungswerke und anwendungsorientierte Tutorials.[6] Was diesen Ressourcen gemeinsam ist, ist ein gewisser Abstraktionsgrad: Die Beispielkorpora sind oft auf das Englische ausgerichtet, und der hochschuldidaktische Kontext mit Philologiestudierenden ohne Programmiervorkenntnisse tritt häufig in den Hintergrund.

[7] Eine wichtige Ausnahme bildet Kristine Hildebrandt,[7] die zeigt, wie digitale Werkzeuge im linguistischen Seminar mit fachlich heterogenen Studierenden ohne Programmierkenntnisse eingesetzt werden können; ein expliziter Fokus auf der Einführung in Programmiersprachen wie R fehlt bei ihr jedoch. Im deutschsprachigen Raum hat sich forTEXT (TU Darmstadt) als zentrale Plattform für DH-Lehrmaterialien etabliert: Lerneinheiten zur Textvisualisierung mit Voyant[8] und zur Stilometrie mit Stylo[9] bieten niedrigschwellige Zugänge zur digitalen Textanalyse für Studierende in den Geisteswissenschaften ohne Programmiererfahrung; sie setzen dabei jedoch auf grafische Benutzeroberflächen statt auf direktes Programmieren und beziehen sich vorwiegend auf literarische Korpora. Für die digitale Philologie bleibt damit weiterhin offen, wie ein methodischer Einstieg aussehen kann, der digitale Verfahren nicht als Selbstzweck, sondern als Erweiterung philologischer Erkenntnisinteressen vermittelt und Studierenden zugleich die direkte Arbeit mit einer Programmiersprache erschließt.[10]

[8] Das vorliegende Tutorial setzt an diesem Desiderat an und orientiert sich an den Grundsätzen des projektbasierten Lernens,[11] das für Studierende ohne Vorkenntnisse in formalen Methoden[12] als besonders lernwirksam beschrieben worden ist.[13] Die Wahl des *German Political Speeches Corpus*[14] ist didaktisch motiviert:

[4] Vgl. Blessing et al. 2015.
[5] Vgl. Adam et al. 2023.
[6] Zu Letzteren zählen: Silge / Robinson 2017; Jockers / Thalken 2020; Fradejas Rueda 2025 (für den hispanophonen Raum besonders relevant).
[7] Vgl. Hildebrandt 2024.
[8] Vgl. Flüh 2024.
[9] Vgl. Stilometrie 2026.
[10] Vgl. Horstmann / Kleymann 2019; Bartsch et al. 2023.
[11] Vgl. Brown et al. 2019; Featherstone 2025.

Politische Reden sind für Germanistikstudierende inhaltlich zugänglich und linguistisch reich genug, um Konzepte wie Tokenisierung, Frequenzanalyse oder lexikalische Vielfalt an authentischem Material zu illustrieren; zugleich eröffnet das XML-Format eine zweite Lernachse im Umgang mit strukturierten Textdaten. Das Skript ist zudem explizit als Vorlage konzipiert, sodass Korpus, Schlüsselwörter und Visualisierungsformen mit geringem Aufwand an andere Lehrkontexte angepasst werden können.

2. Hintergrund: Das Projekt von Adrien Barbaresi

[9] Das in diesem Skript verwendete Korpus stammt aus dem *German Political Speeches Corpus*, das von Adrien Barbaresi aufgebaut wurde.[15] Es handelt sich um ein Korpus politischer Reden auf Deutsch, vor allem von hochrangigen Amtsträgern (z. B. Bundespräsident, Bundeskanzler*in), die nach ihrer politischen Relevanz ausgewählt wurden.

[10] Die Reden wurden aus offiziellen Quellen (u. a. Websites des Bundespräsidenten und der Bundesregierung) gesammelt und im XML-Format[16] mit Metadaten aufbereitet (Titel, Sprecher*in, Datum, Ort, Quelle, Anrede usw.).

[11] Aktuelle Versionen des Korpus werden über Zenodo bereitgestellt, sodass das Material leicht nachnutzbar und zitierbar ist. In diesem R-Skript verwenden wir den Teil des Korpus, der die Reden der Bundespräsidenten umfasst (`Bundespräsidenten.xml`). Die XML-Struktur ermöglicht es uns, die Reden und ihre Metadaten systematisch zu analysieren.

3. Didaktisches Konzept und Seminarkontext

[12] Das vorliegende Tutorial entstand im Rahmen eines Seminars zur digitalen Philologie, das im vierten Studienjahr des Bachelorstudiengangs *Lenguas Modernas* (Moderne Sprachen) an der Universidad de Valladolid durchgeführt wurde. Die Teilnehmenden hatten sich auf Deutsch als Hauptfachsprache spezialisiert und brachten solide philologische Kenntnisse mit – Erfahrungen mit Programmiersprachen oder statistischen Auswertungstools besaßen sie jedoch nicht. Das Seminar verfolgte damit ein klar definiertes Ziel: die Einführung in R und in die Arbeitsumgebung RStudio als philologisches Werkzeug, nicht die informatische Kompetenz um ihrer selbst willen.

3.1 Lernziele und Aufbau

[13] Die Lernziele des Seminars lassen sich in drei Dimensionen gliedern:

1. Technisch: Grundlegende Bedienung von RStudio (Skript-Editor, Konsole, Environment, Plot-Fenster); Installation und Laden von Paketen; Lesen und Verarbeiten von XML-Dateien; Erstellen einfacher Visualisierungen mit ggplot2.

[12] Damit sind technische bzw. MINT-nahe Methoden wie Programmierung oder Statistik gemeint, die im Curriculum geisteswissenschaftlicher Studiengänge in der Regel nicht vermittelt werden.
[13] Vgl. Havenga 2015; Novalia et al. 2025.
[14] Barbaresi 2011–2019.
[15] Vgl. Barbaresi 2011–2019.
[16] Dabei handelt es sich nicht um ein standardisiertes Schema wie TEI, sondern um ein von Barbaresi eigens für dieses Korpus entwickeltes, einfaches XML-Schema mit interner DTD. Es definiert lediglich die Elemente „collection", „text" und „rohtext"; die Metadaten (person, titel, datum, ort, untertitel, url, anrede) sind als Attribute des <text>-Elements abgebildet, der Redetext selbst liegt als Fließtext im Element <rohtext> vor. Ein <teiHeader> oder andere TEI-spezifische Strukturen sind nicht vorhanden.

2. Methodisch: Verständnis korpuslinguistischer Grundbegriffe (*Token*, *Type*, *Lemma*, *Stoppwörter*, *Type-Token-Ratio* (*TTR*), *Konkordanz*); kritische Einordnung quantitativer Befunde im Verhältnis zu qualitativer Textinterpretation.
3. Epistemisch: Reflexion über die Möglichkeiten und Grenzen des *distant reading* als Ergänzung – nicht Ersatz – für das philologische *close reading*.

[14] Das Seminar folgte einem iterativen Aufbau: Jeder Block begann mit der Vorstellung eines Konzepts (z. B. *Tokenisierung*), gefolgt von einem gemeinsam im Plenum ausgeführten Code-Abschnitt, und endete mit einer freien Übungsphase, in der die Studierenden Parameter eigenständig variierten – andere Präsident*innen, andere Schlüsselwörter, andere Visualisierungsformen. Dieses Vorgehen entspricht dem Modell des *scaffolded learning*, bei dem die Lehrperson zunächst stark unterstützt und diese Unterstützung schrittweise zurückzieht, bis die Lernenden selbstständig arbeiten können. Anders als das oben genannte projektbasierte Lernen, das den übergeordneten didaktischen Rahmen des Seminars bildet, bezieht sich das *scaffolded learning* auf die konkrete Unterstützungsstruktur innerhalb der einzelnen Lerneinheiten.[17]

3.2 KI-Unterstützung als didaktisches Element

[15] Eine der methodischen Neuerungen dieses Seminars war die explizite Integration eines KI-Assistenzsystems – konkret Claude Sonnet von Anthropic – in den Lernprozess. Die Studierenden wurden ausdrücklich dazu ermutigt, beim Schreiben und Debuggen ihrer R-Skripte auf dieses Werkzeug zurückzugreifen, allerdings unter einer zentralen Bedingung: Der generierte Code musste verstanden und kommentiert werden, bevor er als Teil des eigenen Skripts akzeptiert wurde.

[16] Diese Entscheidung gründet auf einer wachsenden empirischen Evidenz zur Rolle von Large Language Models (LLMs) im Programmierunterricht. Neuere Studien zeigen, dass KI-Assistenz den Erwerb von Programmierkenntnissen nicht notwendigerweise untergräbt, sondern – wenn sie reflektiert eingesetzt wird – kognitive Zugangshürden abbaut und das Verstehen vertiefen kann: Studierende, die KI nicht nur zur Codeproduktion, sondern zur aktiven Erklärung und Fehlerdiagnose nutzen, zeigen bessere Lernergebnisse als solche, die Code lediglich übernehmen.[18] Für Studierende der Geisteswissenschaften ohne Programmiervorkenntnisse ist dieser Befund besonders relevant: Die natürlichsprachliche Schnittstelle eines LLM erlaubt es, Fehlermeldungen in eigenen Worten zu beschreiben und konzeptuelle Fragen in fachlicher Sprache zu stellen – eine Form der Interaktion, die dem Lernstil von Philologiestudierenden strukturell entgegenkommt.

[17] Gleichwohl wurden im Seminar auch die Risiken dieses Ansatzes thematisiert: die Gefahr der unkritischen Übernahme von fehlerhaftem oder unverständlichem Code, die Tendenz zur Umgehung eigener Denkprozesse (*cognitive offloading*) sowie Fragen der akademischen Integrität. Diese Reflexion über die epistemischen Grenzen KI-generierter Analysen ist ihrerseits ein philologisch relevantes Thema und konnte so in den fachlichen Diskurs des Seminars integriert werden.

3.3 Block 0 – Vorbereitung und Laden der XML-Datei

[18] Zuerst brauchen wir das Korpus der Reden. Die Datei liegt im XML-Format vor, also in einer strukturierten Textform mit Metadaten (Autor, Datum, Ort usw.).

[17] Vgl. Havenga 2015; Novalia et al. 2025.
[18] Vgl. Shen / Tamkin 2026. Die Studie zeigt, dass kognitiv aktive Interaktionsmuster mit KI (Erklärung, Fehlerdiagnose, hybrider Einsatz) den Lernerfolg erhalten, während passive Delegation ihn beeinträchtigt. Vgl. ferner Wills et al. 2024; Lepp / Kaimre 2025.

Damit die Arbeit so einfach wie möglich ist, wird die Datei beim ersten Ausführen automatisch von Zenodo heruntergeladen. Sie müssen nichts manuell herunterladen. [19]

Wir werden: [20]

1. Die notwendigen Pakete laden.
2. Prüfen, ob die XML-Datei im Arbeitsverzeichnis existiert.
3. Falls nicht, das ZIP-Archiv von Zenodo herunterladen und entpacken.
4. Die Datei
   `Bundespräsidenten.xml`
   mit
   `xml2::read_xml()`
   in R einlesen.
5. Das vollständige R-Skript mit allen Codeblöcken, Kommentaren und Beispieldaten steht im GitHub-Repositorium der Autor*innen zur freien Nachnutzung zur Verfügung.

4. Block 1 – Data Frame mit Reden erstellen

Nun wollen wir das XML in eine besser handhabbare Form bringen: eine Tabelle (*tibble*), in der jede Zeile eine Rede ist und jede Spalte eine Information (Präsident, Datum, Ort, Text usw.). [21]

```
df <- tibble(
person = xml_attr(nodos_text, "person"),
titel = xml_attr(nodos_text, "titel"),
datum_raw = xml_attr(nodos_text, "datum"),
rohtext = xml_text(xml_find_all(xml_data, ".//rohtext"))
) |>
mutate(datum = ymd(datum_raw), jahr = year(datum))
head(df)
person datum ort jahr
Joachim Gauck 2016-05-02 Saarbrücken 2016
Joachim Gauck 2014-02-26 Schloss Bellevue 2014
Horst Köhler 2009-03-13 Madrid 2009
Joachim Gauck 2013-11-09 (leer) 2013
Horst Köhler 2009-12-06 Berlin 2009
Christian Wulff 2011-09-10 Schloss Bellevue 2011
```
[22]

Einige einfache Fragen, die wir jetzt sofort beantworten können: Wie viele Reden gibt es? (2045) Welche Bundespräsidenten kommen im Korpus vor? (sechs: Christian Wulff, Horst Köhler, Joachim Gauck, Johannes Rau, Richard von Weizsäcker, Roman Herzog) Welche Zeitspanne deckt das Korpus ab? (1984–2017) [23]

5. Block 2 – Einfache quantitative Auswertungen und erste Grafiken

Wir beginnen mit sehr einfachen quantitativen Auswertungen: Wie viele Reden hat jeder Präsident, und wie viele gibt es pro Jahr? Abbildung 1 zeigt die Gesamtzahl der Reden pro Bundespräsident als Balkendiagramm; Abbildung 2 stellt die zeitliche Verteilung nach Jahren dar. Als Übungsvarianten stehen [24]

ein farbkodiertes Balkendiagramm zur Verfügung (vgl. Abbildung 3), das dieselben Daten nach Präsidenten farblich differenziert, sowie ein Liniendiagramm derselben Zeitreihe (vgl. Abbildung 4), das die Entwicklung über den gesamten Erfassungszeitraum besonders anschaulich macht und sich als Einstieg in die diachrone Analyse eignet.

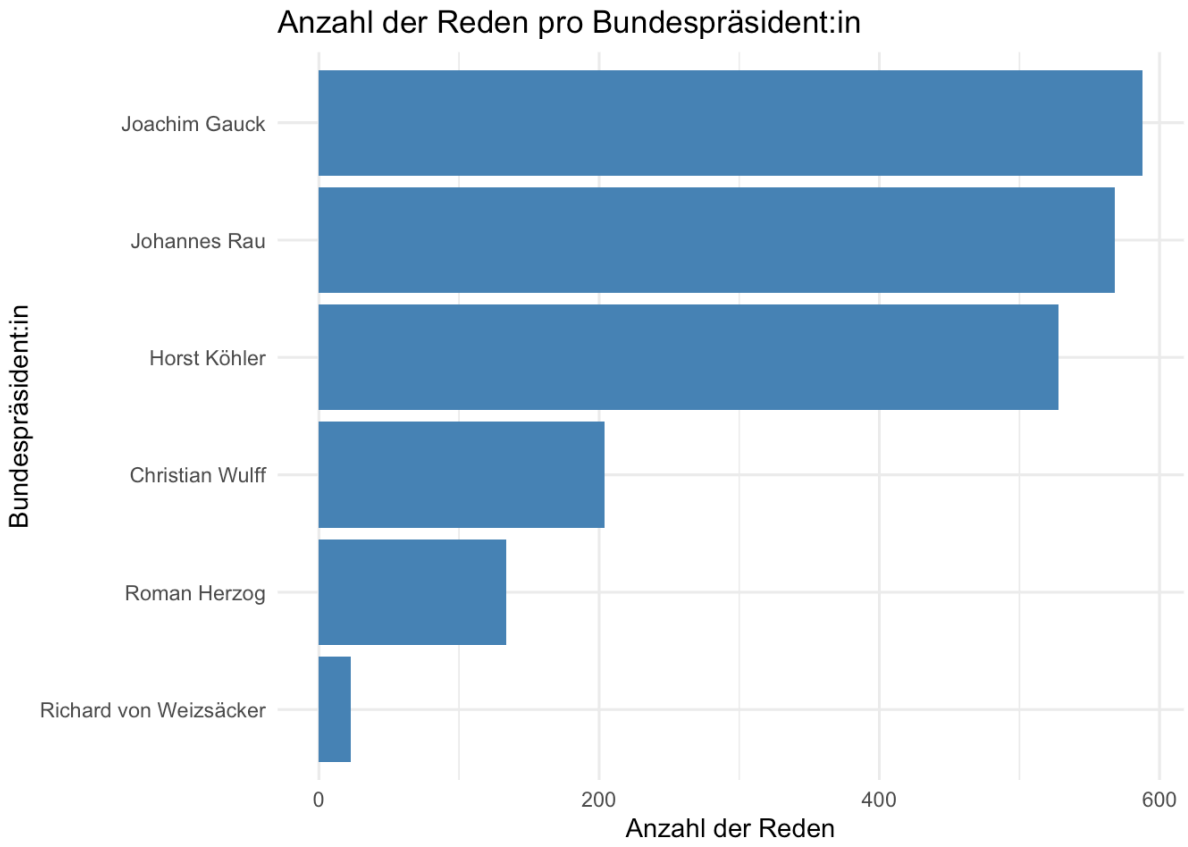

Abb. 1: Anzahl der Reden pro Bundespräsident im *German Political Speeches Corpus*. [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

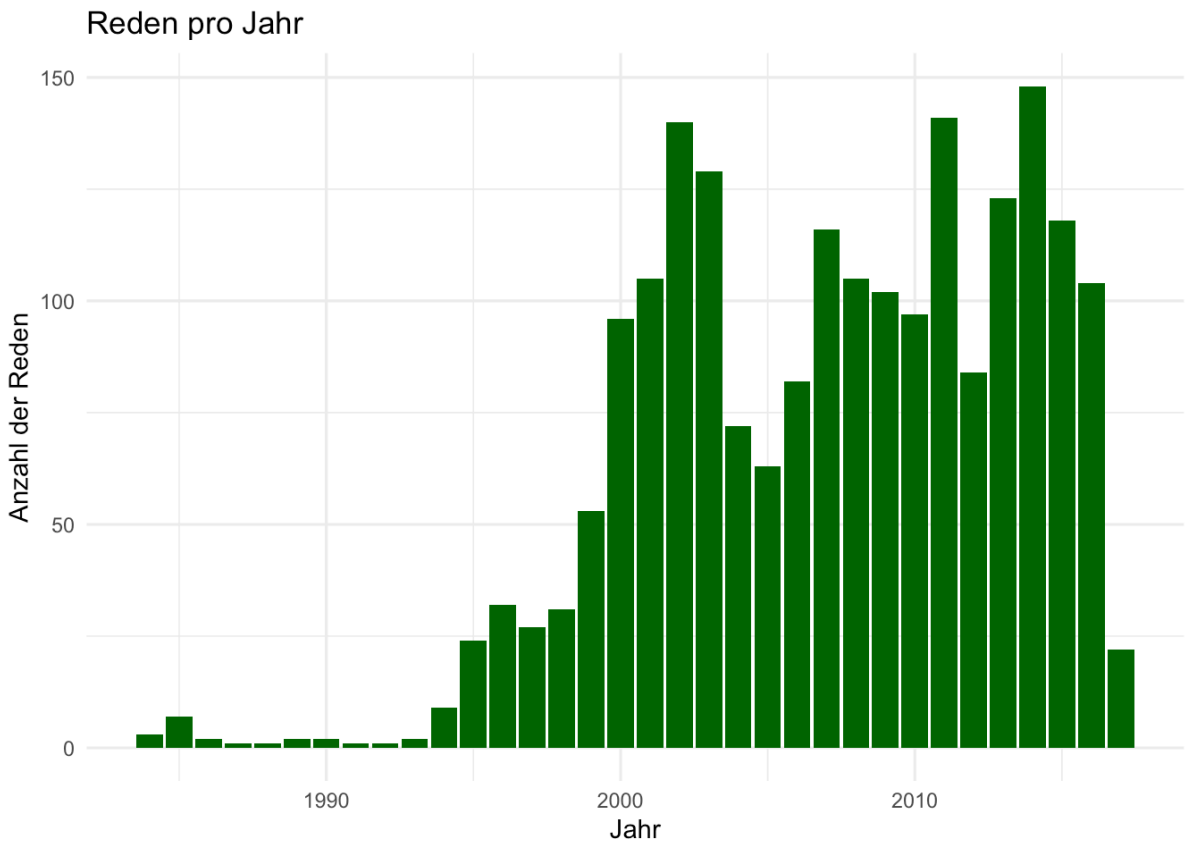

Abb. 2: Reden pro Jahr im *German Political Speeches Corpus* (1984–2017). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

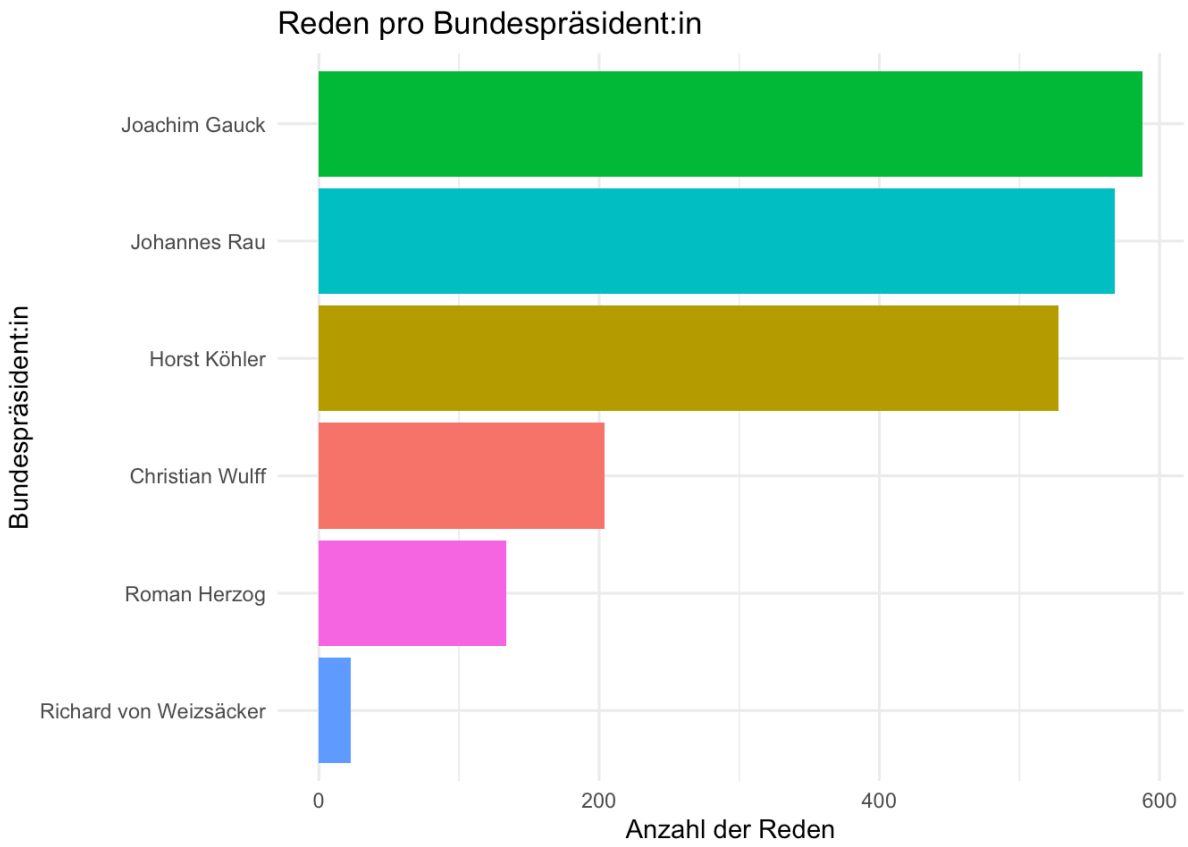

Abb. 3: Reden pro Bundespräsident im *German Political Speeches Corpus* (farbkodiert nach Person). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

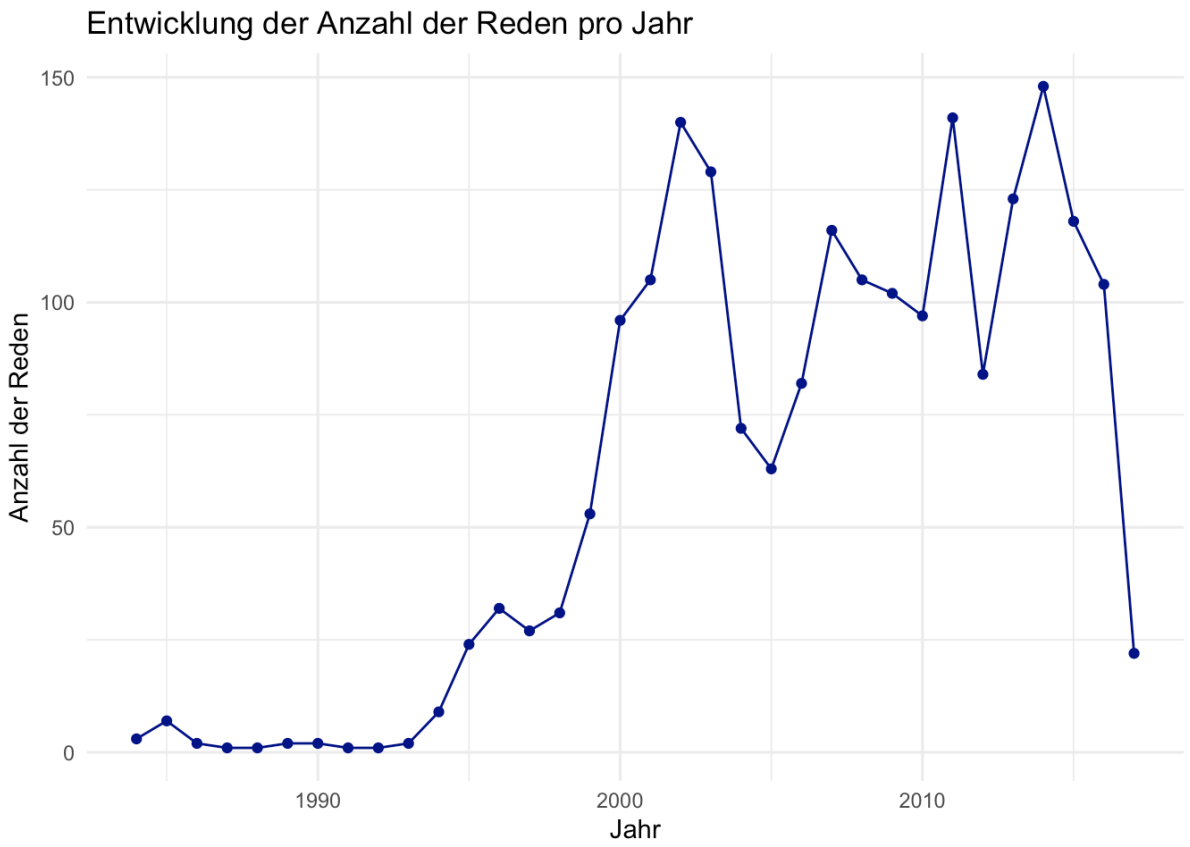

Abb. 4: Entwicklung der Anzahl der Reden pro Jahr im *German Political Speeches Corpus* (1984–2017). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

5.1 Interpretation: Was uns die quantitativen Befunde sagen

Die Ergebnisse von Block 2 sind auf den ersten Blick eindeutig: Joachim Gauck (588 Reden), Johannes Rau (568) und Horst Köhler (528) führen das Ranking an, während Richard von Weizsäcker mit lediglich 23 Reden das Schlusslicht bildet. Diese Zahlen einfach als Maß präsidialer ›Aktivität‹ zu lesen, wäre jedoch ein methodischer Fehler – und genau an dieser Stelle beginnt die philologische Interpretation. [25]

Das Korpus ist nicht gleichmäßig verteilt. Das Balkendiagramm der Reden pro Jahr zeigt deutlich, dass das *German Political Speeches Corpus* erst ab Mitte der 1990er Jahre eine vollständige Abdeckung bietet. Die Amtszeit von Richard von Weizsäcker (1984–1994) ist im Korpus kaum vertreten – nicht, weil er weniger gesprochen hätte, sondern weil die systematische digitale Erfassung von Bundespräsidentenreden online erst später einsetzte. Die niedrige Zahl seiner Reden ist mithin ein Artefakt der Datenerhebung, kein Befund über seine Aktivität als Redner. [26]

Die Amtsdauer erklärt einen Teil der Varianz. Normiert man die absoluten Zahlen auf die Amtsdauer, ergibt sich ein anderes Bild: Christian Wulff etwa war nur gut zwei Jahre im Amt (2010–2012), bevor er infolge einer politischen Affäre zurücktrat – seine 204 Reden entsprechen damit einer bemerkenswert hohen Jahresrate. Joachim Gauck hingegen absolvierte eine volle fünfjährige Amtszeit (2012–2017), was die absolute Spitzenposition relativiert. Eine sinnvolle Ergänzung zu den absoluten Zählungen wäre daher die Berechnung der Reden pro Amtsjahr – eine Übung, die sich als Aufgabe für die Studierenden eignet. [27]

Der Jahresverlauf zeigt strukturelle Muster. Das Balkendiagramm nach Jahren lässt zwei ausgeprägte Aktivitätsphasen erkennen: eine erste um 2001–2003 (Amtszeit Johannes Raus) und eine zweite zwischen 2012 und 2016 (Amtszeit Joachim Gaucks), mit Spitzenwerten von über 140 Reden pro Jahr. Der starke Einbruch am Ende der Zeitreihe – der letzte Balken fällt auf rund 25 Reden – markiert höchstwahrscheinlich ein unvollständiges Jahr in der Datenerhebung und sollte bei Folgeanalysen herausgefiltert werden. Diese Überlegungen illustrieren ein zentrales Prinzip der korpusgestützten Analyse: Quantitative Befunde sind Ausgangspunkte, keine Schlussfolgerungen. Jede Zahl verlangt nach einem historischen und methodischen Kommentar – eine Anforderung, die dem philologischen Ethos des genauen Lesens strukturell verwandt ist. [28]

6. Block 3 – Länge der Reden

Jetzt wollen wir wissen, wie lang die Reden sind (in Zeichen, Wörtern, Sätzen). Wir können auch die mittlere Länge pro Präsident und pro Jahr berechnen (vgl. Abbildung 7). [29]

6.1 Philologische Interpretation: Was uns die Redenlänge verrät

Die quantitativen Grundmaße des Korpus sind aufschlussreich: Die durchschnittliche Rede umfasst 1.401 Wörter bei einem Mittelwert von 80 Sätzen – das entspricht einer vorgetragenen Dauer von etwa zehn bis fünfzehn Minuten und deutet auf ein Format mittlerer Länge hin, dass weder auf eine sehr kurze, noch auf eine außergewöhnlich lange Redeform hindeutet. Diese Zahlen sind jedoch Durchschnittswerte, die eine beträchtliche Varianz verbergen, wie Abbildung 5 unmittelbar zeigt. [30]

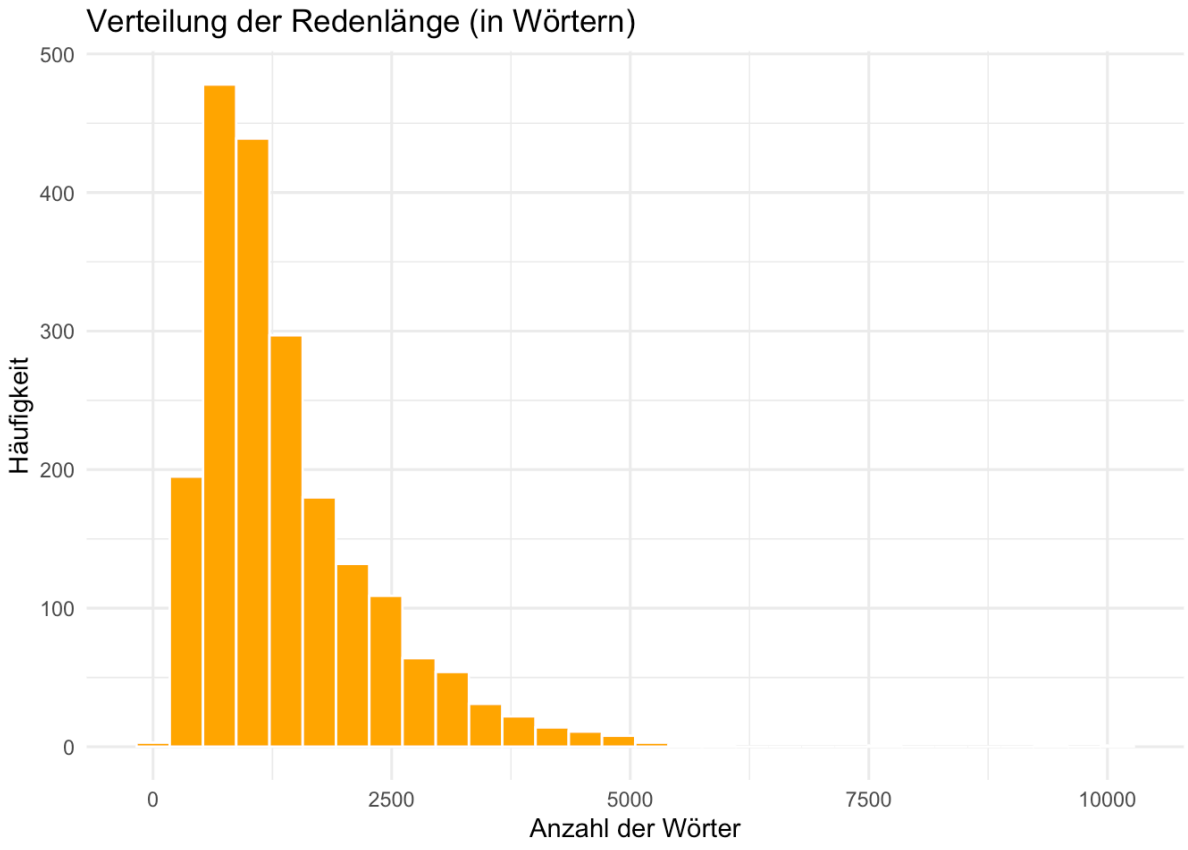

Abb. 5: Verteilung der Redenlänge im *German Political Speeches Corpus* (in Wörtern; rechtsschiefe Verteilung mit Schwerpunkt unter 2.500 Wörtern). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

Die Verteilung ist rechtsschief (vgl. Abbildung 5). Der Großteil der Reden konzentriert sich im Bereich von 500 bis 2.000 Wörtern, mit einem deutlichen Häufigkeitsgipfel um 750–1.000 Wörter. Gleichzeitig existiert ein langer rechter Ausläufer mit einzelnen Reden von über 5.000 Wörtern. Diese Tendenz ist diskursanalytisch interpretierbar: Das Korpus enthält zwei strukturell unterschiedliche Redetypen – kurze, zeremonielle Ansprachen (Glückwünsche, Eröffnungsworte, Danksagungen) und längere politische Grundsatzreden, in denen der Bundespräsident programmatisch zu gesellschaftlichen Fragen Stellung nimmt. Die Länge ist in diesem Sinne kein triviales Maß, sondern ein Indikator für die diskursive Funktion der jeweiligen Rede. [31]

Der nach Präsidenten aufgeschlüsselte Boxplot (vgl. Abbildung 6) offenbart klare stilistische Differenzen. Roman Herzog und Richard von Weizsäcker weisen die höchsten Medianwerte und die breiteste Streuung auf – ihre Reden sind im Durchschnitt deutlich länger als die ihrer Nachfolger, was auf ein elaborierteres, argumentativ dichteres Redeformat hindeutet. Johannes Rau, Joachim Gauck und Horst Köhler zeigen niedrigere Mediane (unter 1.500 Wörter) mit zahlreichen Ausreißern nach oben: Sie hielten viele kurze Ansprachen, aber auch vereinzelte sehr lange Grundsatzreden. Christian Wulff hat den schmalsten Interquartilsbereich – ein Befund, der sich mit seiner kurzen, durch politische Krisen geprägten Amtszeit in Verbindung bringen lässt. Diese Unterschiede reflektieren zugleich Veränderungen im kommunikativen Format: Kürzere Reden können eine Anpassung an die Medienlogik moderner Öffentlichkeit signalisieren, in der prägnante Interventionen sichtbarer sind als ausführliche Argumentationen. Abbildung 7 veranschaulicht die durchschnittliche Redenlänge pro Bundespräsident und bestätigt, dass Roman Herzog und Richard von Weizsäcker mit Mittelwerten von über 2.000 Wörtern pro Rede deutlich über dem Korpusdurchschnitt liegen. [32]

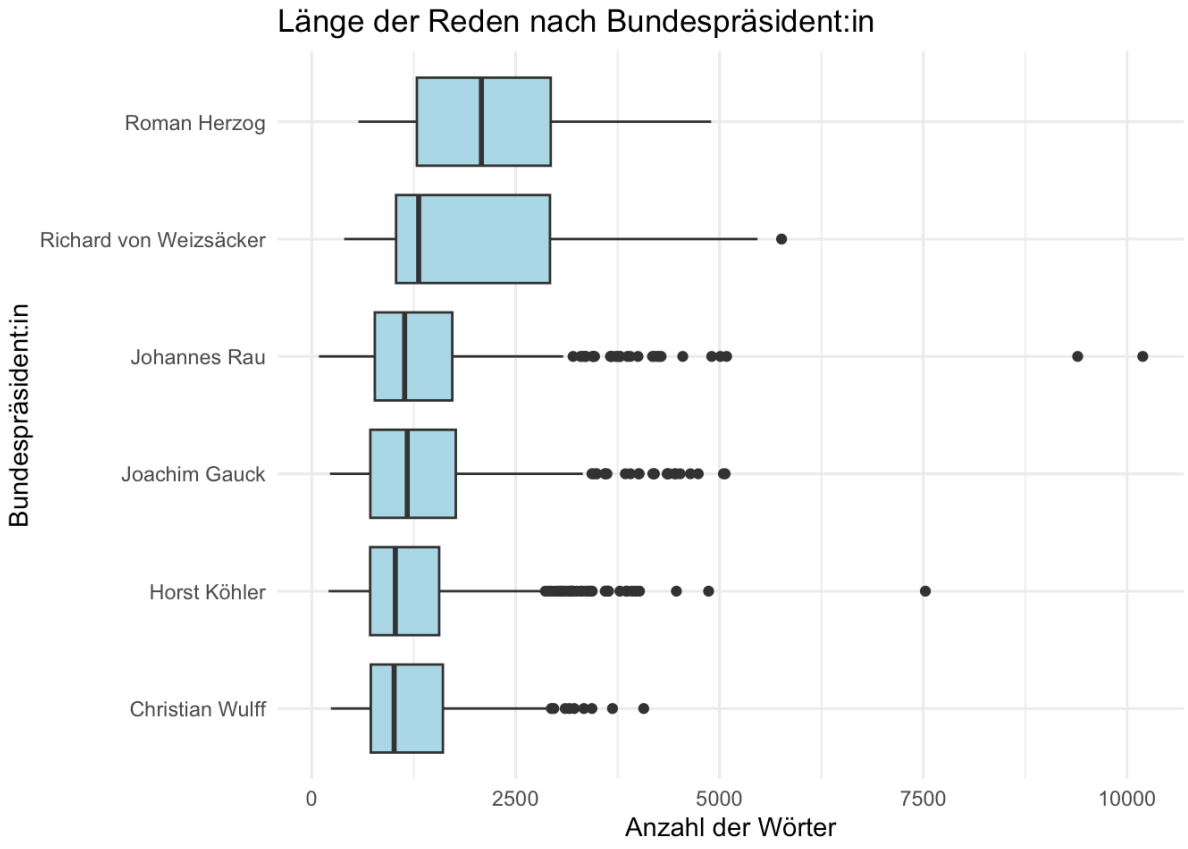

Abb. 6: Länge der Reden nach Bundespräsident im *German Political Speeches Corpus* (in Wörtern; Boxplot mit Ausreißern). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

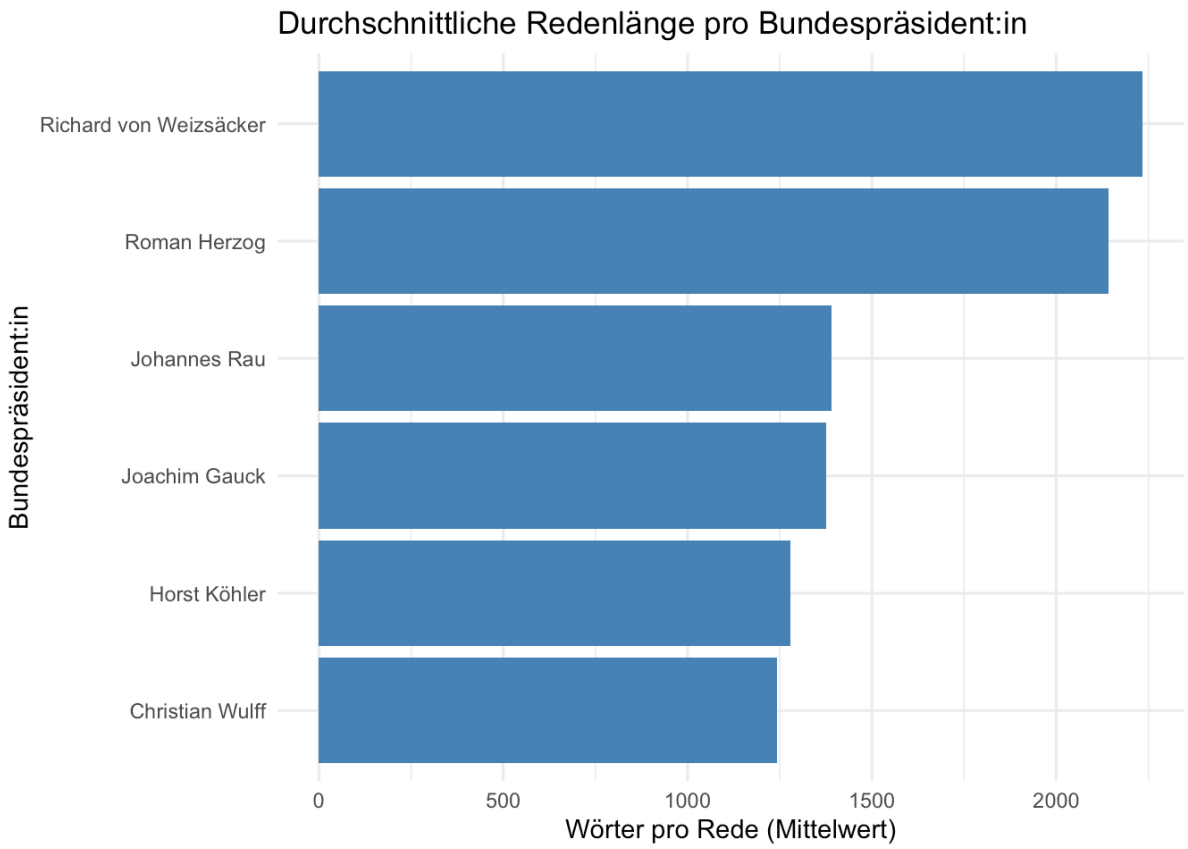

Abb. 7: Durchschnittliche Redenlänge pro Bundespräsident im *German Political Speeches Corpus* (Mittelwert in Wörtern pro Rede). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

7. Block 4 – Tokenisierung und Wortfrequenzen

7.1 Grundbegriffe: Tokens, Types und Lemmata

In der Korpuslinguistik unterscheiden wir: Token (jedes einzelne Vorkommen eines Wortes im Text), Type (jede einzigartige Form) und Lemma (die Grundform eines Wortes: ›spricht‹, ›sprach‹, ›gesprochen‹ sind verschiedene Tokens, gehören aber zum selben Lemma ›sprechen‹). Die Tokenisierung ist der Prozess, einen Text in Einheiten (normalerweise Wörter) zu zerlegen. Im Deutschen stellt dies eine besondere Herausforderung dar: Komposita wie ›Bundespräsident‹ werden als ein einziger Token gezählt, obwohl sie semantisch zwei Konzepte enthalten. [33]

Stoppwörter sind sehr häufige grammatische Funktionswörter (Artikel, Präpositionen, Pronomen), die wenig semantische Information für Inhaltsanalysen liefern. Ihre Entfernung aus dem Analysekorpus hilft dabei, das distinktive thematische Vokabular zu identifizieren. Wir zerlegen die Reden zunächst in einzelne Wörter (d. h., die Datenstruktur wechselt von ›eine Rede pro Zeile‹ zu ›ein Wort pro Zeile‹) und entfernen anschließend die Stoppwörter. Die Häufigkeitsverteilung der verbleibenden Wörter ist in Abbildung 8 als Balkendiagramm dargestellt. [34]

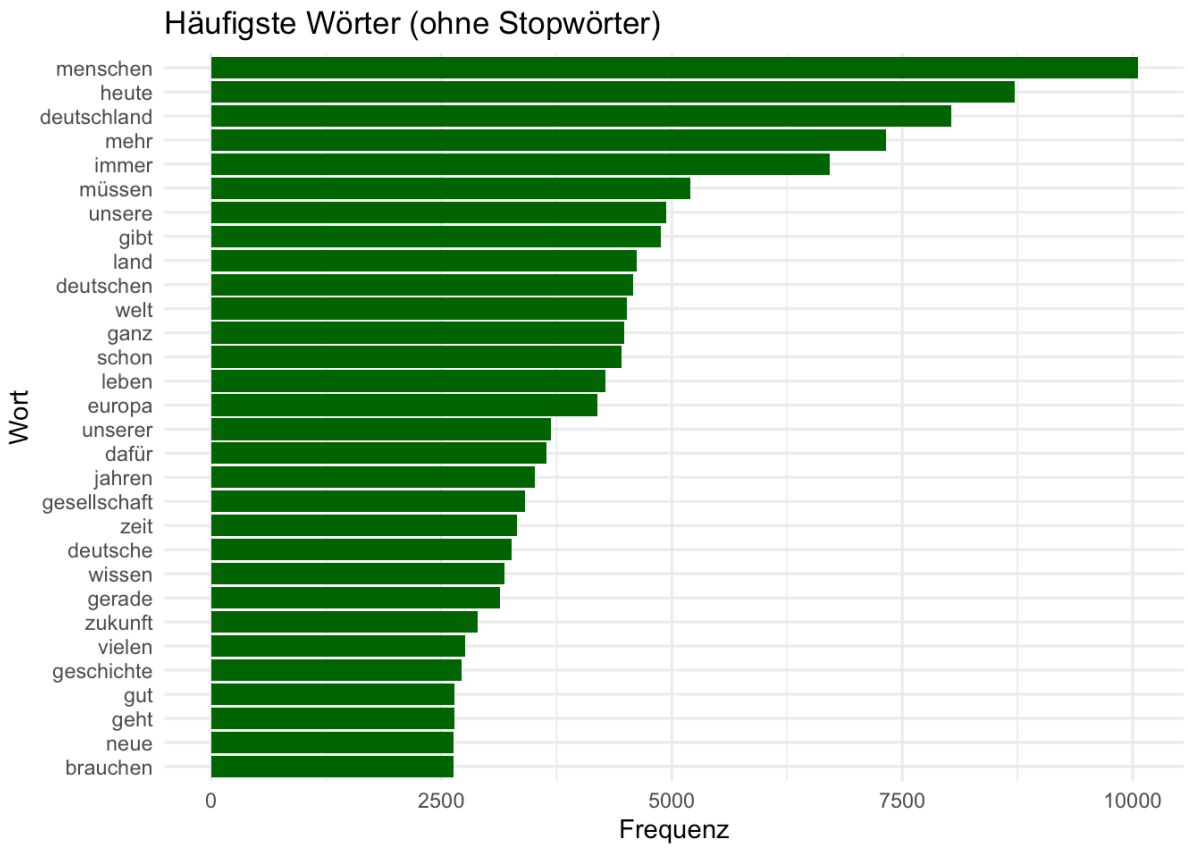

Abb. 8: Die 30 häufigsten Wörter im *German Political Speeches Corpus* nach Entfernung der Stoppwörter (Gesamtkorpus, ohne Lemmatisierung). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

7.2 Philologische Interpretation: Was uns die Wortfrequenzen verraten

Bereits die Rohfrequenzliste zeigt einen methodisch bedeutsamen Befund: Das häufigste Wort vor der Stoppwortbereinigung ist ›dass‹ mit über 20.000 Vorkommen – ein Konjunktionalwort, das für die syntaktische Komplexität der deutschen Präsidentenreden charakteristisch ist. Abhängige Sätze mit ›dass‹ markieren argumentative Strukturen, Behauptungen und normative Aussagen (›Es ist wichtig, dass…‹, ›Ich glaube, dass…‹), die für die politische Rede typisch sind. Der hohe Wert ist damit nicht lexikalisch, aber grammatisch interpretierbar und rechtfertigt die Aufnahme des Wortes in die manuelle Stoppliste. [35]

Das bereinigte Vokabular ist stark wertebezogen (vgl. Abbildung 8). Nach der Entfernung der Stoppwörter dominieren ›Menschen‹ (10.054), ›heute‹ (8.722), ›Deutschland‹ (8.034), ›mehr‹ (7.327) und ›immer‹ (6.717). Dieses Kernvokabular ist für präsidiale Reden hochgradig charakteristisch:[19] ›Menschen‹ als zentrales Subjekt des politischen Diskurses, ›heute‹ als deiktischer Zeitmarker, der die Gegenwartsorientierung der Reden betont, und ›Deutschland‹ als konstante nationale Referenz. Die häufige Verwendung des Modalverbs ›müssen‹ (5.197) ist rhetorisch bedeutsam: Es signalisiert normative Dringlichkeit und appellative Funktion – der Bundespräsident spricht nicht nur beschreibend, sondern auffordernd. [36]

[19] Vgl. Burkhardt 2003, S. 353–357.

Das Lexem ›Europa‹ erscheint bereits unter den 30 häufigsten Wörtern (ca. 3.800 Belege) – ein Befund, der die Zentralität des europäischen Themas in den Präsidentenreden bestätigt und in Kapitel 12 einer detaillierteren Analyse unterzogen wird. Gleiches gilt für ›Gesellschaft‹, ›Geschichte‹, ›Zukunft‹ und ›Welt‹: Diese Lexeme beschreiben das typische thematische Repertoire des Amtes – ein Präsident, der historisch einbettet, gesellschaftlich reflektiert und zukunftsorientiert appelliert. [37]

Eine methodische Einschränkung bleibt festzuhalten: Da die Texte nicht lemmatisiert wurden, werden flektierte Formen getrennt gezählt. So erscheinen ›deutschen‹ und ›deutsche‹ als separate Einträge, obwohl sie zum selben Lemma gehören. Eine Lemmatisierung würde die tatsächliche Reichweite dieser Konzepte im Korpus noch deutlicher sichtbar machen – und stellt, wie in den Schlussbemerkungen (Kapitel 14) skizziert, eine natürliche Erweiterung des hier vorgestellten Workflows dar. Hinzu kommt, dass bei der Tokenisierung sämtliche Wortformen klein geschrieben werden, wodurch auch Groß- / Kleinschreibungsunterschiede aufgehoben werden: So lassen sich etwa das Substantiv ›Deutsche‹ (im Sinne von ›Person deutscher Staatsangehörigkeit‹) und das Adjektiv ›deutsche‹ nach der Tokenisierung nicht mehr unterscheiden. Auch dieser Informationsverlust ließe sich durch ein vorgeschaltetes POS-Tagging vermeiden. [38]

8. Block 5 – Wortwolke

Eine Wortwolke ist eine sehr visuelle Art darzustellen, welche Wörter am häufigsten in einem Text oder Textkorpus vorkommen. Abbildung 9 zeigt die 100 häufigsten Wörter in den Reden Roman Herzogs; Abbildung 10 stellt dieselbe Visualisierung für die Reden Joachim Gaucks dar und eignet sich als Ausgangspunkt für den Vergleich zwischen einzelnen Präsidenten. [39]

Abb. 9: Wortwolke der 100 häufigsten Wörter in den Reden Roman Herzogs im *German Political Speeches Corpus* nach Entfernung der Stoppwörter (ohne Lemmatisierung). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

Abb. 10: Wortwolke der 100 häufigsten Wörter in den Reden Joachim Gaucks nach Entfernung der Stoppwörter (ohne Lemmatisierung). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

8.1 Methodische Anmerkung zur Wortwolke

Die Wortwolke bestätigt visuell, was die Frequenztabelle aus Block 4 (Kapitel 7) bereits gezeigt hat: ›Menschen‹, ›heute‹ und ›Deutschland‹ dominieren das Korpus, gefolgt von einem dichten Cluster aus Wörtern wie ›Gesellschaft‹, ›Europa‹, ›Zukunft‹, ›Demokratie‹ und ›Freiheit‹, die das thematische Repertoire der Präsidentenreden sichtbar machen. [40]

Als Visualisierungswerkzeug hat die Wortwolke klare didaktische Vorzüge: Sie ist intuitiv lesbar, visuell ansprechend und vermittelt auf einen Blick ein erstes Bild des dominanten Vokabulars. Dies macht sie besonders geeignet, um Studierende ohne Vorkenntnisse für die Möglichkeiten der Textanalyse zu motivieren. Für analytische Zwecke ist sie jedoch mit Vorsicht zu verwenden: Die Schriftgröße kodiert Frequenz, aber ohne Achsenbeschriftung sind die genauen Werte nicht ablesbar, und die zufällige Anordnung der Wörter kann semantische Zusammenhänge suggerieren, die im Text nicht existieren. Das Balkendiagramm aus Block 4 (vgl. Abbildung 8) bleibt das präzisere und wissenschaftlich nachvollziehbarere Instrument. [41]

Ein besonders produktiver Einsatz der Wortwolke im Seminar war der Vergleich nach Präsidenten: Wenn Studierende die Wolke für Joachim Gauck (vgl. Abbildung 10) neben die von Roman Herzog (vgl. Abbildung 9) oder Johannes Rau stellen, werden stilistische und thematische Unterschiede unmittelbar sichtbar – auch ohne statistische Vorkenntnisse. Diese Art der explorativen Visualisierung dient als Einstieg in gezieltere Analysen, nicht als deren Ersatz. [42]

9. Block 6 – Bigramme (Wortpaare)

Bigramme sind Paare von aufeinanderfolgenden Wörtern und helfen, einfache Kollokationen zu erkennen (z. B. ›soziale Marktwirtschaft‹). Die 20 häufigsten Bigramme nach Entfernung der Stoppwörter sind in Abbildung 11 dargestellt. [43]

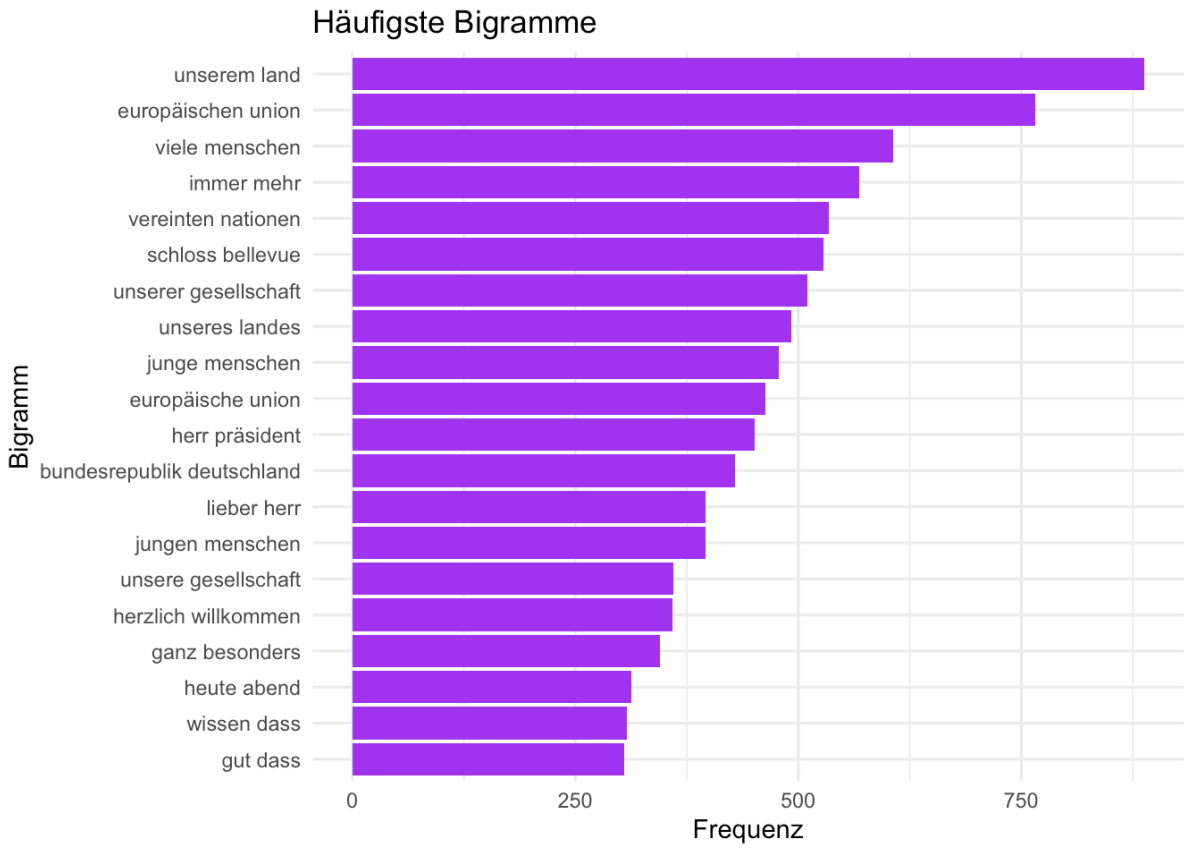

Abb. 11: Die 20 häufigsten Bigramme im *German Political Speeches Corpus* nach Entfernung der Stoppwörter (Gesamtkorpus, ohne Lemmatisierung). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

9.1 Interpretation: Was uns die Bigramme verraten

Die Bigrammanalyse (vgl. Abbildung 11) liefert qualitativ andere Erkenntnisse als die Einzelwortfrequenzen: Während Tokens isolierte Lexeme zählen, zeigen Bigramme syntagmatische Muster – also Wörter, die der Präsident gemeinsam und wiederholt verwendet. Das Ergebnis ist diskursanalytisch aufschlussreich. [44]

Das häufigste Bigramm ist ›unserem Land‹ (ca. 850 Belege), gefolgt von ›Europäischen Union‹ (ca. 760) und ›viele Menschen‹ (ca. 660). Diese drei Kollokationen beschreiben treffend die Grundkoordinaten präsidialer Rhetorik: eine nationale Rahmung (›unserem Land‹), eine europäische Einbettung (›Europäischen Union‹) und ein volksbezogener Appell (›viele Menschen‹). Interessant ist, dass ›Europäischen Union‹ und ›europäische Union‹ als zwei separate Einträge erscheinen – ein erneutes Indiz für die Grenzen einer nicht-lemmatisierten Analyse, bei der unterschiedliche Flexionsformen separat gezählt werden. [45]

Einige Bigramme sind funktional-zeremonielle Formeln. ›Schloss Bellevue‹ (ca. 530), der offizielle Amtssitz des Bundespräsidenten, erscheint als häufiger Redeort und markiert Reden, die explizit dort gehalten wurden. ›Herr Präsident‹, ›Lieber Herr‹ und ›Herzlich willkommen‹ sind typische Anredeformeln, die auf die zeremonielle Dimension vieler Reden hinweisen. Ihr hohes Vorkommen erklärt teilweise den großen Anteil kurzer Reden im Korpus: Viele Texte sind Begrüßungsansprachen, keine politischen Grundsatzreden. [46]

Diskursiv besonders bedeutsam sind die Bigramme ›immer mehr‹, ›junge Menschen‹ / ›jungen Menschen‹ und ›unserer Gesellschaft‹: Sie verweisen auf typische Argumentationsmuster des Amtes: demographische Herausforderungen, gesellschaftlicher Zusammenhalt und eine Tendenz zum sprachlichen Ausdruck kontinuierlicher Zunahme (›immer mehr‹), die auf normative Dringlichkeit schließen lässt. Das Bigramm ›Vereinten Nationen‹ (ca. 540) belegt zudem die außenpolitische Dimension der Präsidentenreden, die in der Einzelwortanalyse weniger sichtbar war. [47]

Für Studierende ist der Vergleich zwischen Unigrammen (Kapitel 8) und Bigrammen besonders lehrreich: Erst in der Kollokation zeigt sich, ob ›Menschen‹ in einem sozialen (›viele Menschen‹, ›junge Menschen‹), einem normativen (›für die Menschen‹, hier als Trigramm) oder einem abstrakten Kontext verwendet wird. Allgemein spricht man hierbei von n-Grammen – zusammenhängenden Sequenzen aus n Wörtern, wobei Bigramme (n = 2) und Trigramme (n = 3) die gebräuchlichsten Fälle sind. Dies illustriert den Mehrwert der n-Gramm-Analyse gegenüber der einfachen Frequenzauszählung. [48]

10. Block 7 – Lexikalische Vielfalt und Komplexität

Die lexikalische Vielfalt, gemessen anhand der TTR (Type-Token-Ratio), also dem Verhältnis von Tokens und Types (unterschiedliche Wortformen), gibt an, welcher Anteil der Wörter in einer Rede unterschiedlichen Wortformen zuzuordnen ist. Abbildung 12 zeigt die Verteilung der TTR-Werte über das Gesamtkorpus. Außerdem lässt sich die sprachliche Komplexität eines Textes näherungsweise über die Anzahl der Wörter pro Satz bestimmen. In methodischer Hinsicht ist zu beachten, dass die TTR hier auf Basis der bereinigten Token-Liste (ohne Stoppwörter) berechnet wird, nicht auf dem Rohtext. Dies hebt die absoluten TTR-Werte gegenüber der klassischen Berechnung an und macht sie nicht direkt mit Studien vergleichbar, die die TTR auf den unbereinigten Text anwenden. Der intern konsistente Vergleich zwischen den Amtsinhabern bleibt jedoch valide. [49]

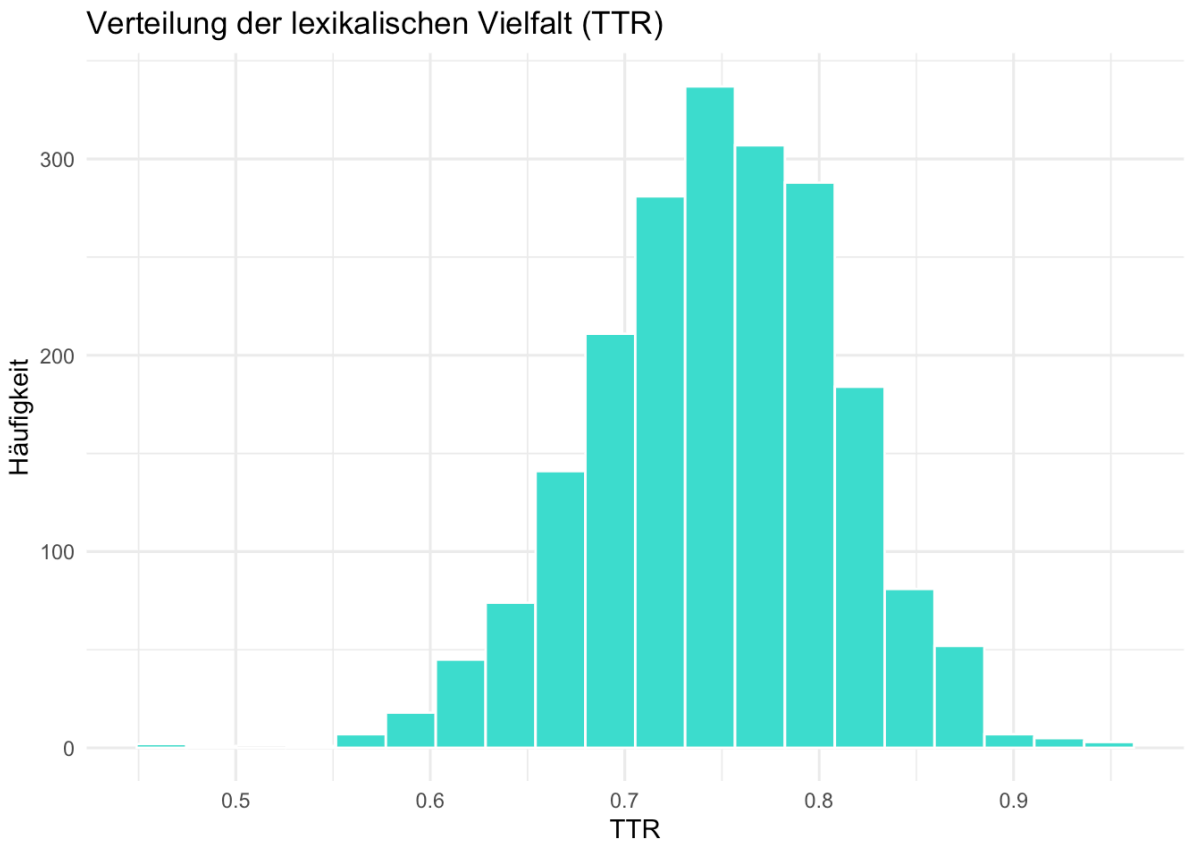

Abb. 12: Verteilung der Type-Token-Ratio (TTR) im *German Political Speeches Corpus* (Gesamtkorpus; Schwerpunkt zwischen 0,65 und 0,80). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

10.1 Interpretation: Lexikalische Vielfalt in den Präsidentenreden

Die Verteilung der TTR-Werte (vgl. Abbildung 12) zeigt ein annähernd normalverteiltes Bild, konzentriert zwischen 0,65 und 0,85, mit einem Gipfel um 0,72–0,75. Das bedeutet, dass im Durchschnitt etwa 72–75 % der Wörter einer Rede unterschiedliche Formen haben – ein Wert, der für die hier vorliegenden Reden mittlerer Länge auf ein abwechslungsreiches, aber nicht übermäßig technisches Vokabular hinweist. [50]

Der Boxplot nach Präsidenten (vgl. Abbildung 13) zeigt, dass die Unterschiede zwischen den Amtsinhabern in absoluten TTR-Werten gering sind: Alle Mediane liegen zwischen 0,70 und 0,76. Roman Herzog und Richard von Weizsäcker weisen die höchsten Medianwerte auf, was konsistent mit ihren in Kapitel 7 beobachteten längeren Reden ist – ein scheinbarer Widerspruch, der sich auflöst, wenn man bedenkt, dass beide zahlreiche kurze Reden gehalten haben, die tendenziell höhere TTR-Werte aufweisen; da diese kurzen Reden den Großteil ihres Redekorpus ausmachen, liegt ihr Median entsprechend im oberen Bereich. Joachim Gauck, Horst Köhler und Christian Wulff clustern eng zusammen, was auf einen ähnlichen diskursiven Stil in Bezug auf lexikalische Variation hindeutet. Die größte Streuung zeigt Johannes Rau – er umfasst sowohl sehr knappe Begrüßungsansprachen als auch die längsten Grundsatzreden des gesamten Korpus. [51]

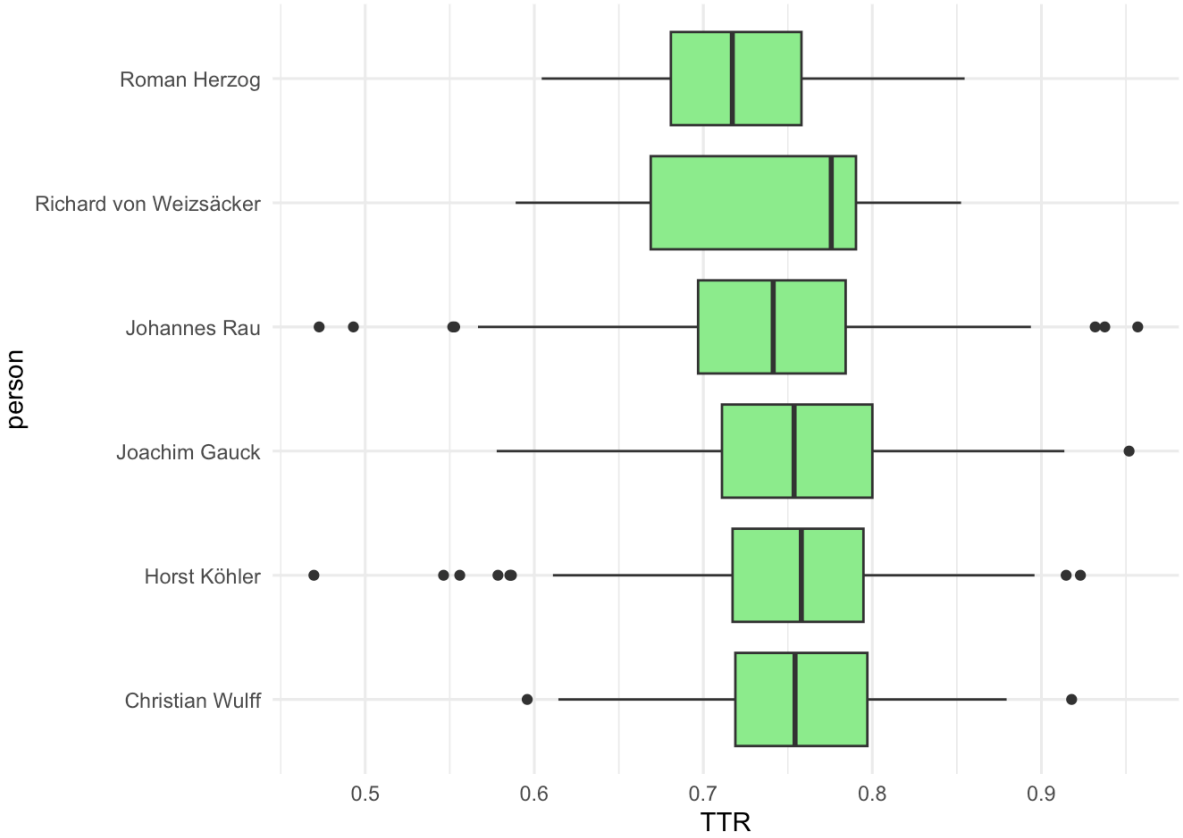

Abb. 13: Lexikalische Vielfalt (TTR) nach Bundespräsident im *German Political Speeches Corpus* (Boxplot mit Ausreißern). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

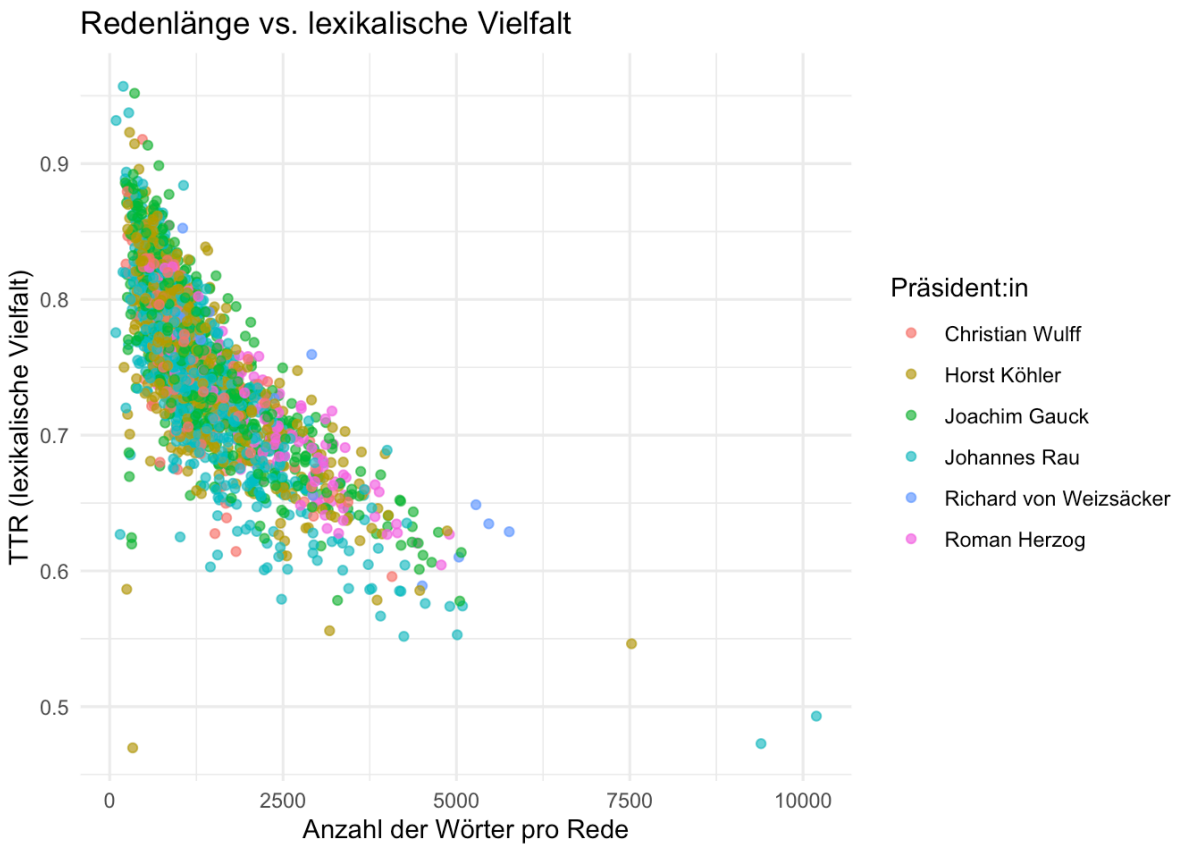

Abb. 14: Zusammenhang zwischen Redenlänge (in Wörtern) und lexikalischer Vielfalt (TTR) im *German Political Speeches Corpus*, farbkodiert nach Bundespräsident (negative Korrelation: längere Reden weisen tendenziell niedrigere TTR-Werte auf). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

Das Streudiagramm (vgl. Abbildung 14) offenbart eine strukturelle Gesetzmäßigkeit, die für die methodische Reflexion zentral ist: Je länger eine Rede, desto niedriger die TTR. Dieser negative Zusammenhang ist kein inhaltlicher Befund, sondern ein mathematisches Artefakt: In längeren Texten werden Wörter zwangsläufig häufiger wiederholt, was den TTR-Wert senkt, unabhängig vom tatsächlichen Reichtum des Vokabulars. Die sehr langen Reden (über 7.500 Wörter, überwiegend von Johannes Rau und Horst Köhler) weisen TTR-Werte unter 0,55 auf, während kurze Ansprachen unter 500 Wörtern TTR-Werte von über 0,90 erreichen können. Dieses Phänomen ist in der Stilometrie gut bekannt und hat zur Entwicklung längennormierter Maße wie *MATTR* (*Moving Average Type-Token Ratio*)[20] oder *MTLD* (*Measure of Textual Lexical Diversity*)[21] geführt, die als Erweiterung dieses Blocks eingeführt werden könnten. [52]

Für Studierende ist dieser Block besonders lehrreich, weil er die Grenzen quantitativer Maße direkt erfahrbar macht: Eine hohe TTR bedeutet nicht unbedingt einen reicheren Stil, sondern kann schlicht Kürze signalisieren. Die Interpretation verlangt, den Zahlenwert immer im Kontext von Textlänge und Gattung zu lesen – eine Übung im kritischen Umgang mit Daten, die über die Philologie hinaus Gültigkeit hat. [53]

[20] Vgl. Covington / McFall 2010 (MATTR).
[21] Vgl. McCarthy / Jarvis 2010 (MTLD).

11. Block 8 – Schlüsselwörter und Konkordanzen (KWIC)

11.1 Relevanz der Schlüsselbegriffe im deutschen Präsidialdiskurs

Die Begriffe, die wir im Folgenden analysieren, sind nicht willkürlich gewählt, sondern spiegeln zentrale Achsen der zeitgenössischen deutschen politischen Debatte wider: ›Europa‹ (die europäische Integration als konstitutives Thema der deutschen Nachkriegsidentität), ›Freiheit‹ (Grundbegriff der liberalen Demokratie und wiederkehrendes Thema nach der Wiedervereinigung), ›Krise‹ (Indikator für die Wahrnehmung politischer Dringlichkeit in verschiedenen Perioden) und ›Flüchtlinge‹ (Zentralität des Migrationsthemas insbesondere seit 2015). Die diachrone Analyse dieser Begriffe erlaubt es, die politische Agenda und die dominierenden Deutungsrahmen (*frames*) in verschiedenen Perioden nachzuzeichnen – ein Verfahren, das auch in vergleichbaren korpusgestützten Untersuchungen deutschsprachiger politischer und medialer Diskurse eingesetzt wurde.[22] [54]

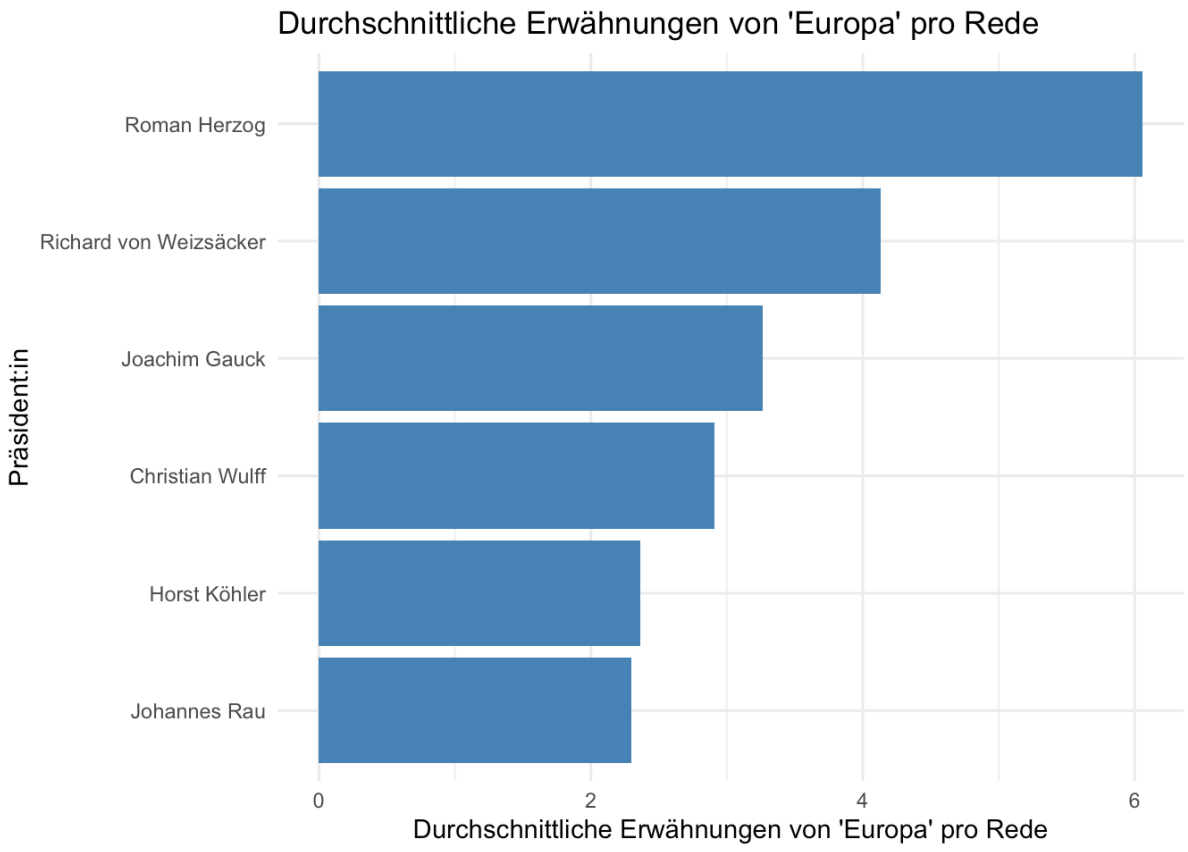

Abb. 15: Durchschnittliche Erwähnungen des Begriffs ›Europa‹ pro Rede nach Bundespräsident im *German Political Speeches Corpus* (Roman Herzog mit dem höchsten Wert). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

Abbildung 15 zeigt, wie oft der Begriff ›Europa‹ durchschnittlich pro Rede bei den einzelnen Bundespräsidenten vorkommt. Darüber hinaus erlaubt die *KWIC*-Funktion (*Key Word In Context*) einen qualitativen Blick auf die Verwendungskontexte der Schlüsselbegriffe: Sie macht sichtbar, ob ein Begriff in einem affirmativen, kritischen oder appellativen Rahmen erscheint und illustriert damit exemplarisch das Zusammenspiel von quantitativer Frequenzanalyse und qualitativer Interpretation, das den methodischen Kern dieses Tutorials bildet. [55]

[22] Vgl. Adam et al. 2023.

11.2 Interpretation: Schlüsselbegriffe in den Präsidentenreden

[56] Das Gesamtvorkommen der vier Schlüsselbegriffe spiegelt unmittelbar die thematischen Prioritäten der Präsidentenreden wider: ›Europa‹ (5.973) und ›Freiheit‹ (3.630) dominieren mit großem Abstand vor ›Krise‹ (1.229) und ›Flüchtlinge‹ (368). Diese Rangfolge ist sowohl inhaltlich als auch historisch interpretierbar.

[57] ›Europa‹ ist das stabilste Grundthema des Amtes. Mit fast 6.000 Belegen über alle Amtsinhaber ist es das weitaus häufigste der analysierten Schlüsselwörter. Abbildung 15 zeigt jedoch erhebliche Unterschiede in der Intensität: Roman Herzog nennt Europa im Durchschnitt über 6-mal pro Rede – der höchste Wert im gesamten Korpus –, was mit seiner Amtszeit (1994–1999) zusammenhängt, in der zentrale europäische Integrationsschritte stattfanden (Maastricht-Vertrag 1993, Vorbereitung des Euro). Richard von Weizsäcker (ca. 4,2 pro Rede) und Joachim Gauck (ca. 3,2) folgen, während Johannes Rau und Horst Köhler trotz vieler Reden die niedrigsten Durchschnittswerte aufweisen – ein Befund, der zur Hypothese einlädt, dass europäische Themen in ihrer Amtszeit stärker auf andere Institutionen (z. B. Bundesregierung, Bundestag) delegiert wurden.

[58] ›Freiheit‹ ist ein Kernbegriff der deutschen politischen Identität, besonders nach der Wiedervereinigung 1990. Seine Häufigkeit (3.630) deutet auf eine kontinuierliche normative Rahmung hin, die sich durch alle Amtszeiten zieht. Die deutlich geringere Frequenz von ›Krise‹ (1.229) mag überraschen, ist aber funktional erklärbar: Bundespräsidenten meiden als staatstragende Institution tendenziell Krisenvokabular, das als alarmistisch wahrgenommen werden könnte. Diese These ist mit der KWIC-Funktion unmittelbar überprüfbar.

[59] ›Flüchtlinge‹ (368) ist mit Abstand das seltenste Wort der Liste, was angesichts der Prominenz des Themas in der deutschen Politik seit 2015 zunächst überrascht. Hier sei jedoch auf den Zeitschnitt des Korpus hingewiesen: Der Großteil der Reden stammt aus den Jahren vor 2016. Eine diachrone Analyse (Block 9, Kapitel 12) wird zeigen, ob die Frequenz dieses Begriffs in den letzten Jahren des Korpus deutlich ansteigt.

[60] Die KWIC-Funktion erlaubt es, über das reine Zählen hinauszugehen. Die Beispielkontexte aus den Reden Horst Köhlers vom März und November 2009, dem Jahr nach der globalen Finanzkrise, ermöglichen eine direkte Überprüfung, ob ›Europa‹ in jenem Jahr in einem Krisenrahmen verwendet wird oder eher als positive Referenz. Diese Art der qualitativen Validierung quantitativer Befunde ist methodisch zentral: Frequenzen zeigen, wo zu lesen ist, die KWIC-Konkordanz zeigt, was dort steht.

12. Block 9 – Zeitliche Dimension von Schlüsselbegriffen

[61] Zum Schluss betrachten wir die diachrone Entwicklung von Begriffen wie ›Europa‹ und ›Freiheit‹. Abbildung 16 zeigt die relative Häufigkeit beider Begriffe pro Jahr über den gesamten Erfassungszeitraum des Korpus. Im Unterschied zu den bisherigen absoluten Zahlen gleicht die relative Häufigkeit die ungleichmäßige Verteilung der Reden über die Jahre aus (vgl. Block 2, Kapitel 5): Sie ergibt sich aus der Anzahl der Vorkommen eines Begriffs in einem bestimmten Jahr, geteilt durch die Gesamtzahl aller Tokens in diesem Jahr. Ein Wert von 0,009 bedeutet, dass der Begriff in besagtem Jahr etwa 9-mal pro 1000 Tokens vorkommt (bzw. 0,9 % aller Wörter ausmacht).

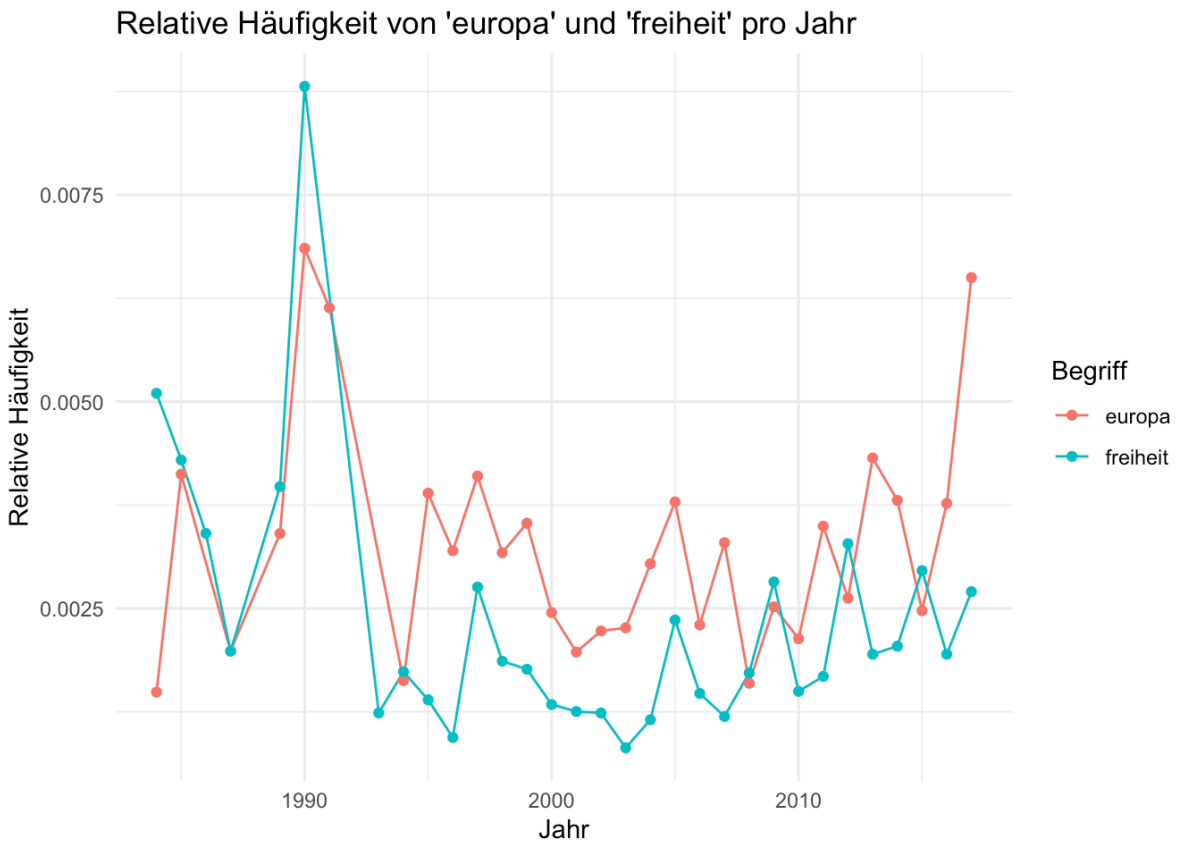

Abb. 16: Relative Häufigkeit der Begriffe ›europa‹ und ›freiheit‹ pro Jahr im *German Political Speeches Corpus* (1984–2017); markanter Doppelgipfel um 1990 im Kontext der deutschen Wiedervereinigung und des Mauerfalls von 1989. [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

12.1 Interpretation: Diachrone Muster in den Präsidentenreden

[62] Das Liniendiagramm (vgl. Abbildung 16) ist die philologisch ergiebigste Visualisierung des gesamten Tutorials: Es macht sichtbar, dass quantitative Muster in politischen Texten nicht zufällig verteilt sind, sondern mit historischen Ereignissen korrespondieren, und damit interpretierbar werden. Dass diachrone Frequenzverläufe gesellschaftlich relevante Diskursverschiebungen abbilden können, hat sich auch in vergleichbaren Korpusanalysen deutschsprachiger politischer und medialer Texte gezeigt.[23] Das vorliegende Tutorial macht dieses Prinzip für Studierende an konkretem historischen Material erfahrbar.

[63] Der markanteste Befund ist der Doppelgipfel um 1990. Sowohl ›Europa‹ als auch ›Freiheit‹ erreichen in diesem Jahr ihre absoluten Höchstwerte im gesamten Korpus, mit ›Freiheit‹ klar an der Spitze (relative Häufigkeit über 0,009). Das Jahr 1990 markiert im unmittelbaren Nachklang des Mauerfalls von 1989 die deutsche Wiedervereinigung und den Beginn einer neuen europäischen Ordnung. Dies findet in den Präsidentenreden unmittelbar und eindeutig Niederschlag. Der Begriff ›Freiheit‹ wird in diesem Kontext nicht abstrakt verwendet, sondern als direkte Referenz auf ein historisches Ereignis, das die politische Identität der Bundesrepublik nachhaltig geprägt hat. Dieser Peak illustriert für Studierende exemplarisch, wie eine diachrone Frequenzanalyse als Seismograph politischer Geschichte funktioniert.

[23] Vgl. Adam et al. 2023.

Nach 1990 folgt ein gemeinsamer Rückgang beider Begriffe, der bis etwa 2002 anhält. Die relative Häufigkeit stabilisiert sich danach auf einem niedrigeren Niveau (zwischen 0,002 und 0,004, d. h. etwa 2 bis 4 Vorkommen pro 1.000 Tokens), mit moderaten Schwankungen. Dies lässt sich als Normalisierung des Diskurses interpretieren: Die außerordentliche historische Situation der Wendezeit wich einem routinierteren Redeformat der konsolidierten Demokratie. [64]

Ab ca. 2014 zeigt ›Europa‹ einen deutlichen Wiederanstieg, der bis zum Ende des Korpus anhält und kurz vor 2016 seinen zweiten Gipfelwert (ca. 0,006) erreicht. Dieser Anstieg fällt zeitlich zusammen mit der Eurokrise (2010–2015), der Ukraine-Krise (2014) und dem Brexit-Referendum (2016), also mit Ereignissen, die den europäischen Zusammenhalt in Frage stellten und das Thema wieder auf die präsidiale Agenda brachten. ›Freiheit‹ hingegen zeigt in diesem Zeitraum keinen vergleichbaren Anstieg und bleibt auf niedrigem Niveau, was darauf hindeutet, dass die Bedrohungswahrnehmung der 2010er Jahre eher geopolitisch als freiheitsbezogen gerahmt wurde. [65]

13. Methodische Limitationen und kritische Überlegungen

Die in den vorangehenden Kapiteln eingeführten Methoden sind mächtige Explorationswerkzeuge, aber keine neutralen Messinstrumente. An mehreren Stellen dieses Tutorials sind wir auf spezifische Grenzen gestoßen, die hier zusammenfassend reflektiert werden. [66]

13.1 Technische Limitationen

Die wichtigste technische Limitation ist das Fehlen einer Lemmatisierung. Wir haben in den Blöcken 4 und 6 (Kapitel 8 und 10) beobachtet, dass flektierte Formen desselben Lexems (z. B. ›deutschen‹, ›deutsche‹, ›deutschem‹) als separate Types gezählt werden und im Ranking separat erscheinen. Dadurch wird systematisch die tatsächliche diskursive Reichweite zentraler Konzepte unterschätzt. Ähnliches gilt für die Satzsegmentierung (Block 3, Kapitel 6), die über einfache Zeichenregeln approximiert wurde, und für den TTR-Wert (Block 7, Kapitel 10), dessen mathematische Abhängigkeit von der Textlänge eine direkte Vergleichbarkeit zwischen Reden unterschiedlicher Länge verhindert. Diese Einschränkungen sind durch den Einsatz von POS-Taggern wie udpipe, längennormierten Diversitätsmaßen wie MATTR oder syntaktischen Parsern lösbar. So zeigt Abbildung 17 den Unterschied zwischen lemmatisierter und nicht-lemmatisierter Analyse. Eine nähere Diskussion dieser Analyse übersteigt jedoch den Rahmen einer Einführungsveranstaltung und wird als Erweiterungsoption in den Schlussbemerkungen aufgeführt.[24] [67]

[24] Exkurs: Eine Lemmatisierung lässt sich in R mit dem Paket udpipe und einem vortrainierten deutschen Modell (german-gsd-ud-2.5) durchführen. Der direkte Vergleich von Frequenzlisten mit und ohne Lemmatisierung zeigt (vgl. Abbildung 17), dass flektierte Formen wie ›Kind‹ und ›Kindern‹ oder ›europäisch‹, ›europäischen‹ und ›Europäer‹ ohne Lemmatisierung als verschiedene Types gezählt werden, was die diskursive Reichweite zentraler Konzepte systematisch unterschätzt. Da die vollständige Lemmatisierung des Gesamtkorpus rechenintensiv ist, steht der vollständige Code (inkl. Vergleichsvisualisierung) im GitHub-Repositorium der Autor*innen zur Verfügung (vgl. González-Miguel / Muñoz-Acebes 2026).

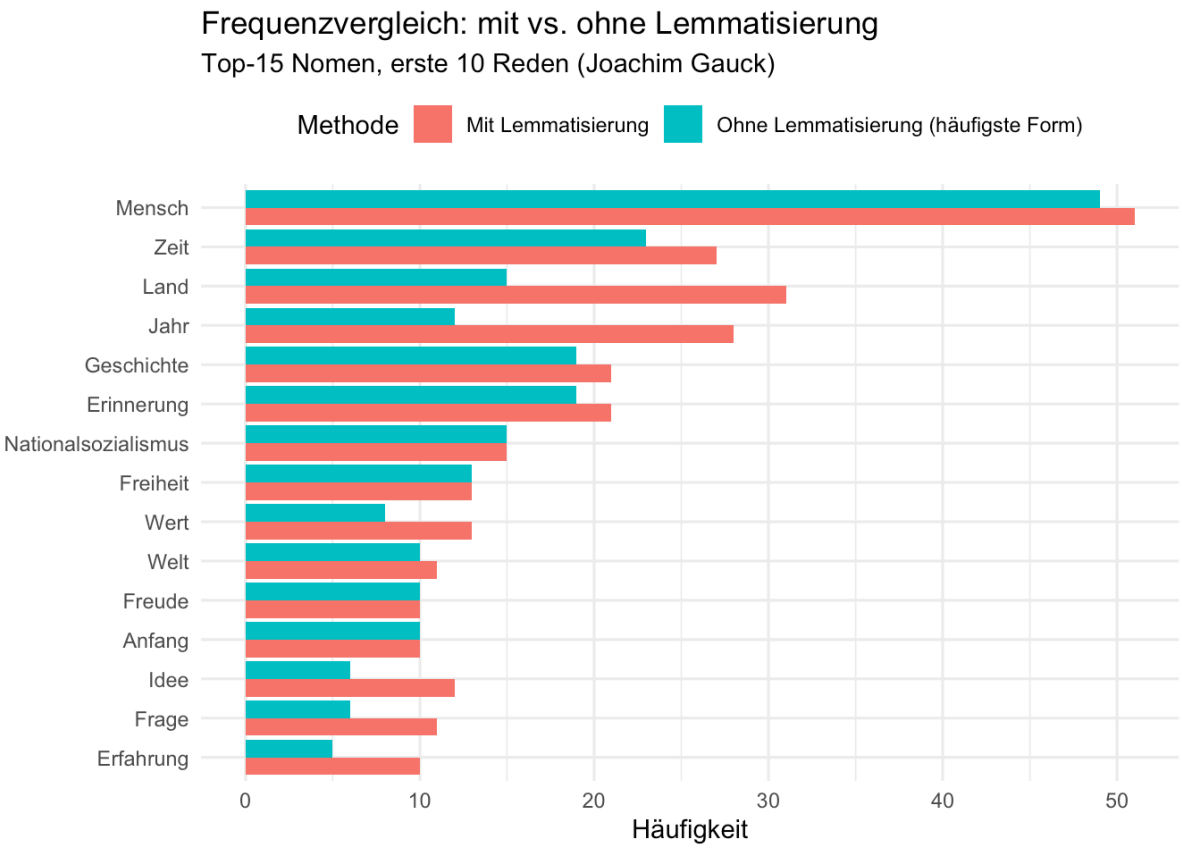

Abb. 17: Frequenzvergleich der Top-15-Nomen mit und ohne Lemmatisierung in den ersten zehn Reden Joachim Gaucks (udpipe, Modell german-gsd-ud-2.5; gruppiertes Balkendiagramm). [Grafik: Ángeles González-Miguel / Francisco Javier Muñoz-Acebes 2026]

13.2 Die Balance zwischen *distant* und *close reading*

Quantitative Methoden sind ein Explorationswerkzeug, kein Ersatz für genaues Lesen. Numerische Muster müssen philologisch interpretiert werden: Ein Peak in der Frequenz von ›Krise‹ zeigt uns, wo wir hinschauen sollten, aber die präzise Bedeutung erfordert das Lesen der Reden in ihrem historischen und rhetorischen Kontext. Die in diesem Tutorial eingeübte Kombination aus quantitativer Exploration und qualitativer Validierung, die exemplarisch in der KWIC-Funktion realisiert wurde, ist methodisch der Kern der zeitgenössischen Digital Humanities: nicht *distant reading* statt *close reading*, sondern *distant* als Orientierung für *close reading*. [68]

14. Erfahrungsbericht: Erkenntnisse aus dem Seminar

Die praktische Erprobung dieses Tutorials im Seminarkontext hat eine Reihe von Beobachtungen erbracht, die über die technische Dokumentation hinaus für Lehrende von Interesse sein dürften. [69]

14.1 Was gut funktioniert hat

Der projektbasierte Ansatz mit einem realen, inhaltlich relevanten Korpus erwies sich als entscheidender Motivationsfaktor. Studierende, die zu Beginn des Seminars skeptisch gegenüber Programmierung waren, zeigten hohes Engagement, sobald die ersten Visualisierungen eigene Hypothesen über das sprachliche [70]

Verhalten der Bundespräsidenten bestätigten oder widerlegten. Besonders wirksam war der Moment, in dem die Wortwolke (vgl. Abbildung 9) und das Frequenzdiagramm (vgl. Abbildung 8) zum ersten Mal erschienen: Das unmittelbare Erkennen bekannter Begriffe (›Menschen‹, ›Europa‹, ›Freiheit‹) schuf eine Brücke zwischen dem philologischen Vorwissen der Studierenden und der neuen Methode.

[71] Die Blöcke mit dem stärksten Lerneffekt waren Block 4 (Kapitel 7: Tokenisierung und Stoppwörter) und Block 9 (Kapitel 12: diachrone Entwicklung). Im ersten Fall löste die Frage, welche Wörter aus der Analyse ausgeschlossen werden sollen und warum, lebhafte Diskussionen über die Grenzen von Automatisierung aus: Ist ›müssen‹ ein Stoppwort? Sollte ›heute‹ herausgefiltert werden? Diese Debatten zeigten, dass die Studierenden begonnen hatten, methodisch zu denken. Im zweiten Fall produzierte die Visualisierung des Höhepunkts von ›Freiheit‹ im Jahr 1990 (vgl. Abbildung 16) spontane historische Kontextualisierungen, ohne dass die Lehrperson eingreifen musste.

14.2 Wo Schwierigkeiten auftraten

[72] Die größten technischen Hürden lagen erwartungsgemäß in der Installation der Pakete und im Umgang mit Fehlermeldungen. Viele Studierende hatten R und RStudio noch nie gestartet und reagierten auf rote Ausgaben in der Konsole zunächst mit Verunsicherung. Hier bewährte sich der Einsatz des KI-Assistenten (Claude Sonnet 4.5): Studierende beschrieben ihre Fehlermeldungen in eigenen Worten und erhielten unmittelbare Erklärungen, was den Unterrichtsfluss erheblich verbesserte. Kritisch bleibt jedoch, dass einige Studierende dazu neigten, KI-generierten Code zu übernehmen, ohne ihn zu verstehen. Diese Tendenz konnte durch die Anforderung, jeden Code-Block zu kommentieren und im Plenum zu erklären, teilweise abgefedert werden.

[73] Ein konzeptuell schwieriger Moment war die Auseinandersetzung mit dem TTR-Artefakt in Block 7 (Kapitel 10): Der Befund, dass kürzere Reden automatisch höhere TTR-Werte aufweisen, unabhängig von ihrem tatsächlichen Wortreichtum, erforderte eine explizite Thematisierung der Grenzen quantitativer Maße und war für viele Studierende die erste bewusste Erfahrung damit, dass Zahlen interpretierungsbedürftig sind.

14.3 Studentische Eigenproduktionen

[74] Im Anschluss an das gemeinsame Tutorial erarbeiteten die Studierenden eigene kleine Analysen, in denen sie das Skript auf abgewandelte Fragestellungen anwendeten, z. B. den Vergleich zweier Präsidenten nach Lexik, die Analyse eines anderen Schlüsselbegriffs (›Demokratie‹, ›Familie‹, ›Wirtschaft‹) oder die Erstellung präsidentenspezifischer Wortwolken. Diese Arbeiten zeigten, dass das Skript als Vorlage funktioniert: Der Code war für alle Studierenden modifizierbar, ohne dass tiefgehende Programmierkenntnisse erforderlich waren. Die Qualität der Interpretationen variierte stark und reichte von rein beschreibenden Kommentaren bis hin zu kontextualisierenden Analysen mit historischen Referenzen. Das weist auf die Notwendigkeit hin, die philologische Interpretationskompetenz parallel zur technischen Einführung explizit zu fördern.

14.4 Pädagogischer Wert und Übertragbarkeit

[75] Das Tutorial vermittelt nicht nur R-Kenntnisse, sondern eine Haltung gegenüber Texten: die Bereitschaft, sie als strukturierte, messbare Daten zu betrachten, ohne ihre hermeneutische Dimension preiszugeben. Die Vorlage ist auf andere Korpora, Sprachen und philologische Fragestellungen übertragbar; der Einsatz im Fremdsprachenkontext (Germanistik an einer spanischsprachigen Universität) hat gezeigt, dass quantitative Methoden den *close-reading*-basierten Zugang produktiv ergänzen.

R-Pakete und Ressourcen

Kenneth Benoit / David Muhr / Kohei Watanabe: stopwords. Multilingual Stopword Lists. CRAN. 12.11.2017. Version 2.3 vom 28.10.2021. DOI: 10.32614/CRAN.package.stopwords

Ian Fellows: wordcloud. Word Clouds. CRAN. 23.07.2011. Version 2.6 vom 24.08.2018. DOI: 10.32614/CRAN.package.wordcloud

Ángeles González-Miguel / Francisco Javier Muñoz-Acebes: R für Philolog:innen – Begleitendes R-Skript [Code]. GitHub. 07.07.2026. [online]

Garrett Grolemund / Hadley Wickham: Dates and Times Made Easy with lubridate. In: Journal of Statistical Software 40 (2011), H. 3, S. 1–25. DOI: 10.18637/jss.v040.i03

Kirill Müller / Hadley Wickham: tibble. Simple Data Frames. CRAN. 23.03.2016. Version 3.3.1 vom 11.01.2026. DOI: 10.32614/CRAN.package.tibble

Erich Neuwirth: RColorBrewer. ColorBrewer Palettes. CRAN. 31.10.2002. Version 1.1-3 vom 03.04.2022. DOI: 10.32614/CRAN.package.RColorBrewer

R Core Team: R. A Language and Environment for Statistical Computing. Version 4.3.3. Wien 1993–2026. [online]

RStudio Team: RStudio. Integrated Development Environment for R. Version 2023.12.1. Boston 2023. [online]

Julia Silge / David Robinson: tidytext. Text Mining and Analysis Using Tidy Data Principles in R. In: Journal of Open Source Software 1 (2016), H. 3, Art. 37. DOI: 10.21105/joss.00037

Hadley Wickham: ggplot2. Elegant Graphics for Data Analysis. 2. Auflage. New York 2016. [Nachweis im GVK]

Hadley Wickham: stringr. Simple, Consistent Wrappers for Common String Operations. CRAN. 09.11.2009. Version 1.6.0 vom 04.11.2025. DOI: 10.32614/CRAN.package.stringr

Hadley Wickham / Davis Vaughan / Maximilian Girlich: tidyr. Tidy Messy Data. CRAN. 21.07.2014. Version 1.3.2 vom 19.12.2025. DOI: 10.32614/CRAN.package.tidyr

Hadley Wickham / Jim Hester / Jeroen Ooms: xml2. Parse XML. CRAN. 20.04.2015. Version 1.5.2. vom 17.01.2026. DOI: 10.32614/CRAN.package.xml2

Hadley Wickham / Romain François / Lionel Henry / Kirill Müller / Davis Vaughan: dplyr. A Grammar of Data Manipulation. CRAN. 29.01.2014. Version 1.2.1 vom 03.04.2026. DOI: 10.32614/CRAN.package.dplyr

Yihui Xie: knitr. A General-Purpose Package for Dynamic Report Generation in R. CRAN. 17.01.2012. Version 1.51 vom 20.12.2025. DOI: 10.32614/CRAN.package.knitr

Literatur

Raven Adam / Marie Lisa Kogler / Martina Scholger: Aufwind in der Berichterstattung zum Klimaschutz. Langfristige Entwicklung von Themen und Stimmungsbildern in österreichischen Zeitungen. In: Zeitschrift für digitale Geisteswissenschaften 8 (2023). Version 2 vom 27.06.2024. HTML. DOI: 10.17175/2023_006_v2

Taylor Arnold / Nicolas Ballier / Paula Lissón / Lauren Tilton: Beyond Lexical Frequencies: Using R for Text Analysis in the Digital Humanities. In: Language Resources and Evaluation 53 (2019), H. 4, S. 707–733. PDF. DOI: 10.1007/s10579-019-09456-6

Adrien Barbaresi: German Political Speeches Corpus. 2011–2019. Datenset. DOI: 10.5281/zenodo.3611246

Sabine Bartsch / Evelyn Gius / Marcus Müller / Andrea Rapp / Thomas Weitin: Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft. In: Zeitschrift für digitale Geisteswissenschaften 8 (2023). 01.06.2023. HTML. DOI: 10.17175/2023_003

André Blessing / Fritz Kliche / Ulrich Heid / Cathleen Kantner / Jonas Kuhn: Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorie. In: Constanze Baum / Thomas Stäcker (Hg.): Grenzen und Möglichkeiten der Digital Humanities (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 1). Wolfenbüttel 2015. 19.02.2015. HTML. DOI: 10.17175/sb001_013

Benjamin D. Brown: Evolving Project Based Learning Methodology at the University Level. A Need for More Guidance and Accountability. In: Alabama Journal of Educational Leadership 6 (2019), S. 10–19. PDF. [online]

Armin Burkhardt: Das Parlament und seine Sprache. Studien zu Theorie und Geschichte parlamentarischer Kommunikation. Tübingen 2003. [Nachweis im GVK]

Michael A. Covington / Joe D. McFall: Cutting the Gordian Knot. The Moving-Average Type–Token Ratio (MATTR). In: Journal of Quantitative Linguistics 17 (2010), H. 2, S. 94–100. PDF. DOI: 10.1080/09296171003643098

Michael Featherstone: Teaching Programming in Higher Education Using a Hybrid Project-Based Learning Approach. In: IEEE Transactions on Education 68 (2025), H. 4, S. 377–386. PDF. DOI: 10.1109/TE.2025.3577406

Marie Flüh: Lerneinheit: Textvisualisierung mit Voyant. In: forTEXT 1 (2024), H. 5. 07.08.2024. HTML. DOI: 10.48694/fortext.3773

José Manuel Fradejas Rueda: Cuentapalabras. Estilometría y análisis de textos con R para filólogos. Valladolid 2025. HTML. [online]

Marietjie Havenga: Project-Based Learning in Higher Education. Exploring Programming Students' Development towards Self-Directedness. In: South African Journal of Higher Education 29 (2015), H. 4, S. 135–157. 01.01.2015. PDF. [online]

Kristine Hildebrandt: Narrating a Path. Digital Humanities Tools in the Linguistics Classroom. In: Proceedings of the Linguistic Society of America 9 (2024), H. 3. 30.10.2024. PDF. DOI: 10.3765/plsa.v9i3.5848

Jan Horstmann / Rabea Kleymann: Alte Fragen, neue Methoden – Philologische und digitale Verfahren im Dialog. Ein Beitrag zum Forschungsdiskurs um Entsagung und Ironie bei Goethe. In: Zeitschrift für digitale Geisteswissenschaften 4 (2019). 12.12.2019. HTML. DOI: 10.17175/2019_007

Matthew L. Jockers / Rosamond Thalken: Text Analysis with R. For Students of Literature. 2. Auflage. New York 2020. [Nachweis im GVK]

Marina Lepp / Joosep Kaimre: Does Generative AI Help in Learning Programming: Students' Perceptions, Reported Use and Relation to Performance. In: Computers in Human Behavior Reports 18 (2025). HTML. DOI: 10.1016/j.chbr.2025.100642

Philip M. McCarthy / Scott Jarvis: MTLD, vocd-D, and HD-D. A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. In: Behavior Research Methods 42 (2010), H. 2, S. 381–392. DOI: 10.3758/BRM.42.2.381

Franco Moretti: Distant Reading. London 2013. [Nachweis im GVK]

Riska Novalia / Arita Marini / Totok Bintoro / Uyu Muawanah: Project-Based Learning. For Higher Education Students' Learning Independence. In: Social Sciences & Humanities Open 11 (2025). HTML. DOI: 10.1016/j.ssaho.2025.101530

Judy Hanwen Shen / Alex Tamkin: How AI Impacts Skill Formation. arXiv. 28.01.2026. Version 2 vom 01.02.2026. DOI: 10.48550/arXiv.2601.20245

Julia Silge / David Robinson: Text Mining with R. A Tidy Approach. Sebastopol, US-CA 2017. [Nachweis im GVK]

Stilometrie und Literaturanalyse. Drei Schritt-für-Schritt-Tutorials In: forTEXT. Literatur digital erforschen. Letzter Zugriff: 29.06.2026. HTML. [online]

Simon Wills / Sanson T. S. Poon / Arianna Salili-James / Ben Scott: The Use of Generative AI for Coding in Academia. In: Methods in Ecology and Evolution 15 (2024), H. 12, S. 2189–2191. 12.11.2024. PDF. DOI: 10.1111/2041-210X.14454

Abbildungsverzeichnis