

Keywordanalysen im Politikmonitoring

WISSENSCHAFTLICHE BERATER DES PROJEKTS

Christopher Georgi

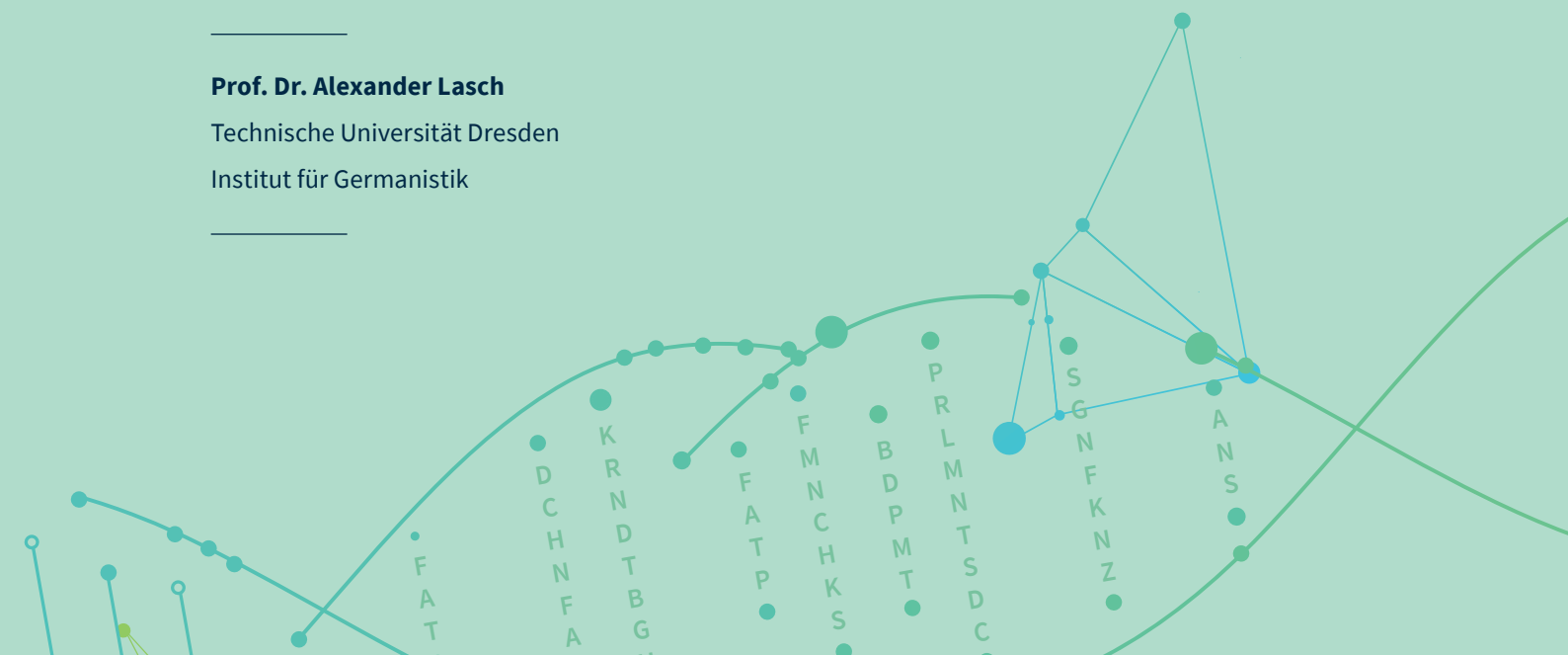
Universität Paderborn

Institut für Germanistik und Vergleichende Literaturwissenschaft

Prof. Dr. Alexander Lasch

Technische Universität Dresden

Institut für Germanistik



Polit-X baut ein dynamisch wachsendes Monitorkorpus (6 Mio. laufende Wortformen pro Tag) auf der Grundlage politischer Dokumente auf Bundes- und Landesebene auf.

Die täglich aktualisierte Datenbasis übersteigt nicht nur die Datenmenge, die man lesend zur Kenntnis nehmen kann (sie entspricht ca. 50 Romanen), bei weitem, sondern ist in dieser Aktualität, Qualität der Aufbereitung und Quantität auch aus korpuslinguistischer Forschungsperspektive bemerkenswert. Berücksichtigt werden nämlich dabei breitgefächerte Dokumentarten wie Plenarprotokolle, Kleine Anfragen, Pressemitteilungen, Ausschreibungen, Beschlussempfehlungen u.a.m..

Das besondere gegenüber anderen Projekten, die Text-Korpora maschinell erfassen, ist die Tagesaktualität der Datenbasis und die Möglichkeit zur dynamischen Auswertung: **Die Daten sind sofort verfügbar für weitere Analyseschritte und Visualisierungen, die als Basis für statistisch abgesicherte Interpretationen dienen können.**

[illegible]

Automatische und statistisch abgesicherte Themenfindung und Keywordanalyse

WICHTIGSTE WÖRTER IDENTIFIZIE- REN

Um die wichtigsten Wörter in Texten zu identifizieren, werden die grammatisch sortierten Keywords (exemplarisch aus dem Dokument *Armutsgefährdung, Einbuße, Armutsgefährdungsgrenze*) im betreffenden Dokument mit einem frei wählbaren Ausschnitt aus dem Monitorkorpus verglichen.

Es wird danach gefragt, ob die Häufigkeit von *Armutsgefährdung* z.B. in der Pressemitteilung *3,1 Millionen Erwerbstätige waren 2019 hierzulande von Armut bedroht* **besonders hoch oder niedrig (statistisch signifikant) im Vergleich zu anderen Texten ist.**

Dazu werden gängige statistische Signifikanzmessungen verwendet und Signifikanzprofile nach zeitlich und quantitativ frei wählbaren Korpuschichten und Teilkorpora angelegt. Letztere können zusammengesetzt sein aus verschiedenen Quellen (z.B. Bundestag und/oder Landtag), bestimmten Dokumentarten (z.B. Kleine Anfrage und/oder Plenarprotokoll) oder themenspezifisch (z.B. ausgehend von der vorliegenden Pressemitteilung zum Thema Armut).

Signifikanzmessung

SIGNIFIKANZ- MESSUNGEN GENERIEREN KEYWORD- LISTEN

Die Signifikanzmessungen ergeben Keyword-Listen, also Listen mit Schlagworten, für jedes einzelne Dokument im Vergleich zum Monitorkorpus. **Dabei handelt es sich um jene Wörter, die besonders charakteristisch für das jeweilige Dokument sind.** In unserem Beispiel sind fünf besonders wichtige Keywords beispielsweise: *Armutsgefährdung, Kernerwerbstätige, Armutsgefährdungsgrenze, Realschulabschluss und Hauptschulabschluss.*

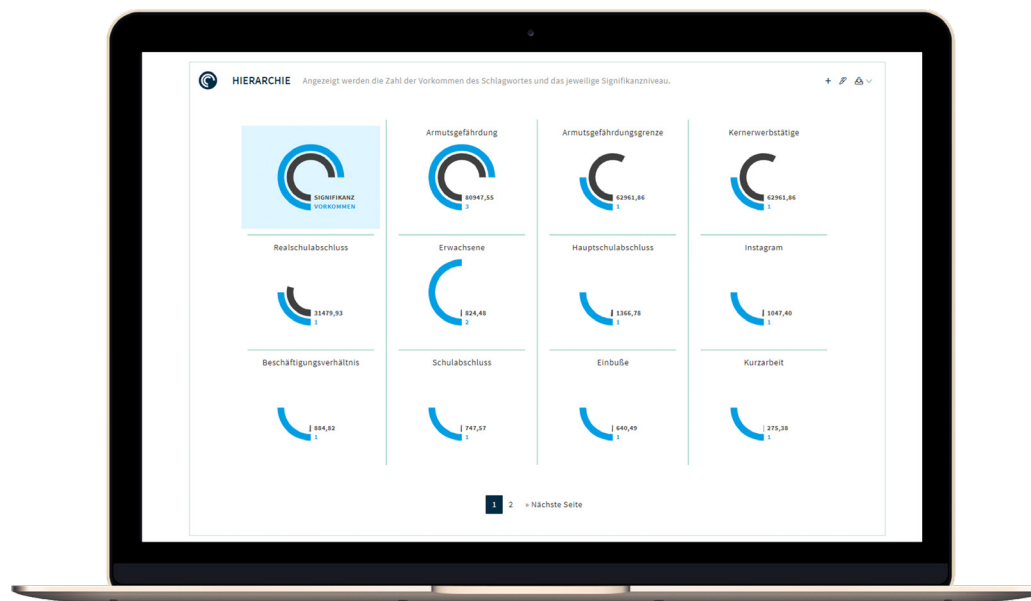


Abbildung 1: Die Keywords der Pressemitteilung 3,1 Millionen Erwerbstätige waren 2019 hierzulande von Armut bedroht werden mit dem Polit-X Feature visualisiert.

Nennungen im zeitlichen Verlauf

NENNUNGEN ALS BASIS FÜR TRENDS

Neben der Signifikanz lassen sich auch Nennungen im zeitlichen Verlauf im Monitorkorpus gut nachvollziehbar darstellen. Nennungen bilden die Basis um Trends über einen Zeitraum zu entdecken und sehr leicht **Zusammenhänge zwischen den Keywords des Dokuments zu identifizieren**, die im zeitlichen Verlauf durch je unterschiedliche Frequenzen auffallen. So ist es möglich, auf unterschiedliche Themen einzugehen und festzustellen, wann welche Themen öffentlich diskutiert worden sind. Im Unterschied zu anderen Trend-Berechnungen bildet hier die Zahl der korpuslinguistisch errechneten Keywords die Datengrundlage und stellt das **Trend-Ergebnis somit auf einen methodisch breiteren Sockel**.

Betrachtet man bspw. *Armutsgefährdung* als ein wichtiges Keyword der vorliegenden Pressemitteilung und interessiert sich für die Häufigkeit des Keywords im zeitlichen Verlauf im Monitorkorpus, dann wird man feststellen, dass *Armutsgefährdung* kontinuierlich häufiger genannt und damit thematisch im Monitorkorpus relevant wird.

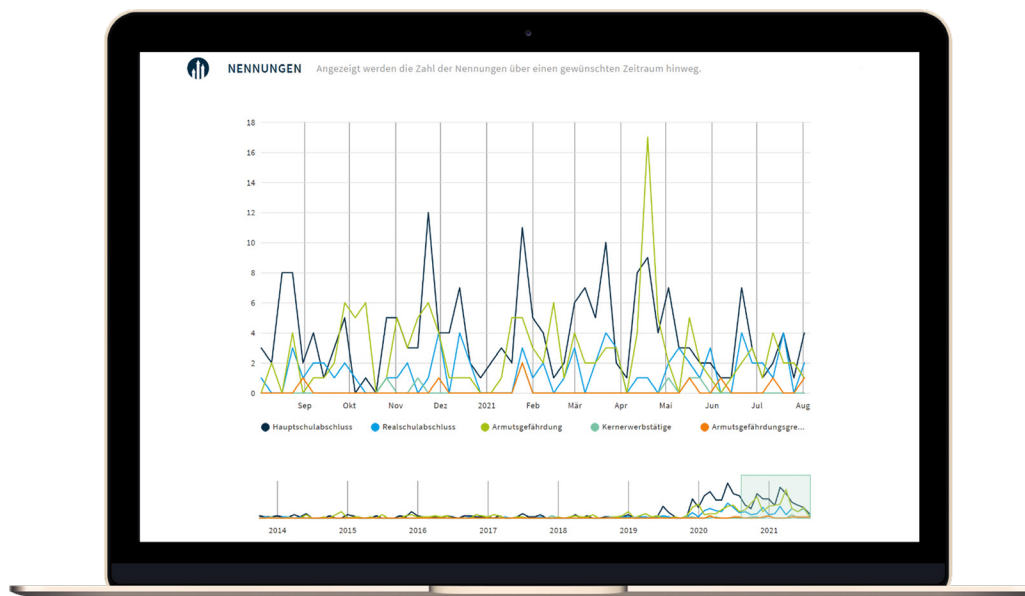


Abbildung 2: Nennungen der Keywords im Verlauf bei Polit-X

Mit computer- linguistischer Intelligenz in die Zukunft der Politikbetrachtung

Debatten und Schlagwörter

KEYWORDS IM ÖFFENTLICH- POLITISCHEN DISKURS

Die Beschäftigung mit Schlagwörtern bzw. Keywords im öffentlich-politischen Diskurs weist in der Linguistik eine lange Tradition auf. Dies ist im Wesentlichen auf die besondere Rolle der Schlagwörter zurückzuführen, die nach Schröter (2015: 396) „oft ein ganzes politisches Programm kondensiert erfass[en]“ und als zentrale Charakteristiken an bestimmte Diskurse zu politisch-gesellschaftlich relevanten Sachverhalten gebunden sind. Darüber hinaus sind mit ihnen bestimmte Sprachstrategien verbunden.

Für diese hat sich speziell in der Politolinguistik der Ausdruck *Begriffe besetzen* etabliert (vgl. Niehr 2014: 64). **Eine dieser Strategien stellt die Begriffsumdeutung dar**, die dazu dient, einem Begriff eine weitere oder andere Bedeutung zuzuschreiben. Gegenwärtig existieren beispielsweise gegenüber den Begriffen *Klimakrise* und *Umweltschutz* sehr unterschiedliche Bedeutungszuschreibungen.

Darüber hinaus ist die Begriffsumwertung eine beliebte Strategie, die wiederum darauf abzielt, **die positive oder negative Wertung eines Begriffs umzukehren**. Diese wird insbesondere bei kontroversen Begriffen wie *Energiewende* deutlich. **Politische Debatten sind demzufolge geprägt von zentralen Schlagwörtern**, die wie der *Klimaschutz* zunehmend populär werden können, im Laufe der Auseinandersetzung Umdeutungen erfahren und schließlich, etwa wie der Begriff *Gastarbeiter*, teilweise nahezu in Vergessenheit geraten.

**KEYWORDS
ALS THE-
MATISCHE
MARKER**

Im Zuge des Data-driven Turn in den Sozial- und Kulturwissenschaften erlangten mit den Keyword-Analysen auch die Schlagwörter bzw. Keywords selbst ungeahnte Popularität, da sie auf der Grundlage von umfassenden Korpora und statistischen Signifikanzmessungen eine breite empirische Basis erhielten (vgl. Bubenhofer/ Scharloth 2015: 7f.).

Die Analysen haben zum Ziel, **Wörter zu finden, die für ein Dokument oder Korpus besonders charakteristisch und demnach als thematische Marker zu verstehen sind**. Die zentrale Schwierigkeit besteht Stefanowitsch (2020: 358) zufolge jedoch darin, bei den Keywords zwischen thematischen Markern und Markern für das Genre zu unterscheiden. In anderen Worten: Es ergeben sich bei den Berechnungen immer auch Keywords, die nicht für das Thema, sondern für die spezifische Textstruktur oder generell die sprachliche Varietät stehen. Für eine sichere Unterscheidung ist fundiertes linguistisches Hintergrundwissen nötig, etwa zu Besonderheiten der Textstruktur. Ferner unterstützt eine genaue grammatikalische Auszeichnung der Korpusdaten die Unterscheidung.

Betrachtet man Schlagwörter über einen längeren zeitlichen Verlauf hinweg, so können **neben der Themenbestimmung auch Schlussfolgerungen und Prognosen zu Trends getroffen werden**. Dabei ist wiederum Hintergrundwissen zu wichtigen Quellen und der allgemeinen Vernetzung verschiedener Informationskanäle notwendig.

**BESTIMMUNG
THEMATISCHER
CLUSTER**

Schließlich vollzog sich mit der Anwendung von **Deep Learning-Algorithmen auf Sprachdaten** ein weiterer großer Entwicklungssprung in der Schlagwortanalyse. Mit Hilfe der Berechnung von Word Embeddings (Mikolov et al. 2013) können Wörter in einen n-dimensionalen Vektorraum übertragen werden. Liegt dieser Berechnung ein ausreichend großes Korpus zugrunde – dabei gilt: je mehr Daten, umso besser – können **präzise Aussagen zu Themen und Trends sowie ferner zur Ähnlichkeit von Dokumenten getroffen werden**. Dabei ist die Bestimmung thematischer Cluster auf der Basis von Word Embeddings absoluter State of the Art.

LITERATUR

Bubenhöfer, Noah/ Scharloth, Joachim (2015): Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. In: ZGL 43 (1), S. 1–26.

Mikolov, Tomas/ Chen, Kai/ Corrado, Greg/ Dean, Jeffrey (2013): Efficient Estimation of Word Representations in Vector Space. In: Computer Science [cs], arXiv:1301.3781.

Niehr, Thomas (2014): Einführung in die Politolinguistik. Gegenstände und Methoden. Göttingen: Vandenhoeck & Ruprecht.

Schröter, Melani (2015): Besondere Wörter III: Schlagwörter in der öffentlich-politischen Auseinandersetzung. In: Ulrike Haß, Petra Storjohann: Handbuch Wort und Wortschatz. Berlin/ Boston: de Gruyter, S. 394–412.

Statistisches Bundesamt (2021): 3,1 Millionen Erwerbstätige waren 2019 hierzulande von Armut bedroht. Pressemitteilung Nr. N 008 vom 28. Januar 2021. https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/01/PD21_N008_634.html (Zugriff vom 02.09.2021).

Stefanowitsch, Anatol (2020): Corpus linguistics. A guide to the methodology. Berlin: Language Science Press.

AUTOREN

Christopher Georgi ist Mitarbeiter am Lehrstuhl für Germanistische und Allgemeine Sprachwissenschaft an der Universität Paderborn und Doktorand an der Professur für Angewandte Linguistik der Technischen Universität Dresden. Seine Forschungsschwerpunkte liegen neben dem Themengebiet Sprache und Emotion im Bereich der Korpus- und Computerlinguistik, der forensischen Linguistik sowie der Kulturlinguistik mit dem Schwerpunkt auf Data-driven Studies.

Prof. Dr. Alexander Lasch ist Inhaber des Lehrstuhls für germanistische Linguistik und Sprachgeschichte in an der Technischen Universität Dresden. Er studierte Deutsch und Geschichte und arbeitete an der Ludwig-Maximilians-Universität München und der Christian-Albrechts-Universität zu Kiel. Seine Forschungsschwerpunkte liegen im Bereich konstruktionsgrammatischer Zugänge zur Gegenwartssprache, Erschließung und maschinelle Analyse im Kontext der Digital Humanities, diskurs- und domänenspezifische Kommunikation in Vergangenheit und Gegenwart, funktionale und regionale Varietäten sowie Themen der angewandten Linguistik. Er ist Gründungsmitglied des Arbeitskreises „Konstruktionsgrammatik des Deutschen“ und Vorsitzender der Gesellschaft für germanistische Sprachgeschichte (GGSG).

**Sprechen Sie uns gerne an.
Wir freuen uns auf das Gespräch mit Ihnen!**

**030 / 280 408 46
info@polit-x.de**

POLIT-X IST EIN DIENST DER

HAUPTSTÄDTISCHE Verlags- & Beratungsgesellschaft HBVG mbH

Novalisstraße 8a
D-10115 Berlin