Publizistik (2023) 68:291–303
https://doi.org/10.1007/s11616-023-00785-9

PUBLIZISTIK

AUFSATZ

Swiss-AL: Plattform für Sprachdaten zur Analyse öffentlicher Kommunikation in der Schweiz

Philipp Dreesen · Julia Krasselt

Eingegangen: 12. Dezember 2022 / Angenommen: 17. Mai 2023 / Online publiziert: 28. Juni 2023

Zusammenfassung Der Beitrag stellt Swiss-AL (=Swiss-Applied Linguistics) vor, eine Plattform für die Forschung zu öffentlicher Kommunikation in der Schweiz, die an der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) entwickelt und bereitgestellt wird. Swiss-AL enthält rund 4,5 Mrd. Wörter in vier Sprachen (Deutsch, Französisch, Italienisch, Rätoromanisch). Es umfasst online publizierte Texte aus Politik und Verwaltung, Wirtschaft, Wissenschaft und Zivilgesellschaft sowie journalistischen Medien. Swiss-AL ist für diskursanalytische Forschungen entwickelt worden, stellt jedoch ebenso eine wichtige Datenbasis für die Kommunikations- und Medienwissenschaft dar, beispielsweise für quantitativ ausgerichtete/standardisierte Medieninhaltsforschung. Swiss-AL wird gegenwärtig zu einer Open-Research-Data-Ressource für die Angewandten Wissenschaften weiterentwickelt. Dazu gehört die Weiterentwicklung einer browserbasierten Workbench (Zugang: www.swiss-al.linguistik.zhaw.ch), die Zugang zu den Daten ermöglicht und den Anforderungen des Daten- und Urheberrechtsschutzes der Schweiz und der EU entspricht. Im Beitrag werden aggregierende Analysemöglichkeiten präsentiert, die diese Workbench gegenwärtig bereitstellt und die insbesondere für inhaltsanalytische Fragestellungen anwendbar sind. Der Beitrag schließt mit einem Ausblick auf zukünftige Entwicklungen im Bereich Open Research Data und skizziert geplante Implementierungen von FAIR-Prinzipien.

Schlüsselwörter Korpus · Textdaten · Diskursanalyse · Inhaltsanalyse · Open Research Data · Öffentliche Kommunikation

Prof. Dr. Philipp Dreesen · ✉ Dr. Julia Krasselt
Institute of Language Competence, Departement Angewandte Linguistik, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Theaterstrasse 17, 4801 Winterthur, Schweiz
E-Mail: julia.krasselt@zhaw.ch

Prof. Dr. Philipp Dreesen
E-Mail: philipp.dreesen@zhaw.ch

Swiss-AL: language data platform for the analysis of public communication

Abstract The paper presents Swiss-AL (=Swiss-Applied Linguistics), a platform for research on public communication in Switzerland, which is developed and provided at the Zurich University of Applied Sciences (ZHAW). Swiss-AL contains around 4.5 billion words in four languages (German, French, Italian, Romansh). It includes texts published online from the fields of journalistic media, politics & administration, business, science, and civil society. Swiss-AL was developed for discourse studies but is equally an important database for communication and media studies, for example for quantitatively oriented/standardised media content research. Swiss-AL is currently being developed into an open research data resource for the applied sciences. This includes the further development of a browser-based workbench (access: www.swiss-al.linguistik.zhaw.ch) that allows access to the data and meets the requirements of data and copyright protection in Switzerland and the EU. The article presents the aggregating analysis options that this workbench currently provides and that are particularly applicable to content analysis questions. The article concludes with an outlook on future developments in the area of Open Research Data and outlines planned implementations of FAIR principles.

Keywords Corpus · Text data · Discourse analysis · Content analysis · Open research data · Public communication

1 Einleitung

Die aktuelle Etablierung von Open Research Data (ORD) in den Geistes- und Sozialwissenschaften betrifft in besonderer Weise die nachhaltige Nutzung von Sprachdaten. Die wissenschaftliche Nachfrage nach Textdaten ist nicht auf philologische Disziplinen beschränkt; vielmehr besteht in vielen Disziplinen ein genuines Erkenntnisinteresse an empirisch analysierbaren Kommunikationsprozessen (z. B. in Public Health, Rechtswissenschaften, Kommunikationswissenschaft). Vor diesem Hintergrund ist die Bereitstellung von Sprachdaten eine disziplinenübergreifende Angelegenheit. Im Beitrag wird vorgestellt, wie nicht-linguistische Disziplinen einen niedrigschwelligen Zugang zu aufbereiteten Sprachdaten erhalten und bei der Analyse großer Textmengen Unterstützung finden: Swiss-AL (=Swiss Applied Linguistics) ist eine diskursanalytisch ausgerichtete Plattform, die Datenverarbeitung, Datenzugang und nutzer:innenfreundliche Analysetools vereint, wobei der Schwerpunkt auf aggregierenden Methoden zur Identifikation sprachlicher Gebrauchsmuster liegt (sog. *distant reading*). Im Folgenden stellen wir – ausgehend von Nutzungsoptionen für die Kommunikationswissenschaft – die Zusammenstellung der Datensätze, die Diskursmodellierungen sowie die Besonderheiten der Datenaufbereitung und Zugangsoptionen vor (Abschn. 2). Beispiele für kommunikationswissenschaftliche Anwendungsoptionen mit inhaltsanalytischem Bezug sind gesondert aufgeführt (Abschn. 3). Der Beitrag schließt mit einem Ausblick auf die derzeitige Umgestaltung von Swiss-AL zu einer ORD-Plattform (Abschn. 4).

Swiss-AL ist Teil der linguistischen Forschungsinfrastruktur CLARIN-CH und der europäischen CLARIN-Gemeinschaft.[1] Bei der Beschaffung, Aufbereitung und Weiterentwicklung setzt Swiss-AL auf die FAIR-Prinzipien und implementiert diese sukzessive in die bestehende Infrastruktur. Swiss-AL wird vom ZHAW Digital Discourse Lab für angewandte und transdisziplinäre Forschung mit Kompetenzen aus Korpuslinguistik, Diskursanalyse sowie Organisationskommunikation betrieben.[2]

2 Swiss-AL: eine Plattform für die Analyse öffentlicher Kommunikation in der Schweiz

Das Departement Angewandte Linguistik der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) entwickelt und teilt die digitale Plattform Swiss-AL. Sie umfasst die größte mehrsprachige Korpusfamilie der Schweiz mit aktuell rund 4,5 Mrd. Wörtern sowie eine dazugehörige Workbench zur Analyse dieser Daten (www.swiss-al.linguistik.zhaw.ch). Im Unterschied zu anderen Textkorpora dieser Größe (für das Deutsche in der DACH-Region z. B. das DeReKo des Instituts für Deutsche Sprache, das DWDS der Berlin-Brandenburgischen Akademien der Wissenschaften oder das C4-Korpus, vgl. Deppermann et al. 2023; für das Englische vgl. z. B. https://www.english-corpora.org) ist Swiss-AL diskursanalytisch ausgerichtet: Es handelt sich um eine Korpusfamilie, deren einzelne Korpora speziell für die Analyse öffentlicher Diskurse und der in diesen Diskursen sprechenden Akteure und Akteursgruppen konzipiert sind (vgl. Krasselt et al. 2020, 2023; Dreesen und Stücheli-Herlach 2019). Somit ist Swiss-AL *nicht* als Referenzkorpus für *das Deutsche* oder *das Französische* in der Schweiz konzipiert (was zu einer anderen Zusammensetzung führen würde), sondern enthält Texte öffentlich sprechender Kollektivakteure (bspw. Behörden, Parteien, journalistische Medien), die sich als Datengrundlage für die Exploration gesellschaftlicher Diskurse eignen.

Kommunikationswissenschaftlich ist Swiss-AL speziell zu Zwecken der Organisationskommunikation modelliert: Parallel zu journalistischen Perspektiven werden insbesondere organisationale Perspektiven im Diskurs aufgezeigt (vgl. Wehmeier et al. 2013, S. 15), die sich aus der Wissenskonstruktion und -distribution von Akteursrollen und den Relationen von Akteuren zueinander ergeben (vgl. Dreesen und Krasselt 2021, S. 390, 399–404). Verorten lassen sich diese Anwendungsfälle in der Communicative Constitution of Organizations (CCO) (zur Übersicht vgl. Brummans et al. 2014), sodass sich diese Forschung als „metadiscursive“ (Craig 1999, S. 120) verstehen lässt. In dieser konstruktivistischen Vorstellung von Organisationen wird mit Verweis auf die Diskursanalyse der wiederholte, musterhafte Sprachgebrauch zum erklärenden Sachverhalt (vgl. Jablin und Putnam 2001, S. 81).

Der Nachweis musterhaften Sprachgebrauches kann ein geeigneter Zugang von kommunikationswissenschaftlicher Diskursanalyse (vgl. Fraas und Pentzold 2016, S. 231–233) und Inhaltsanalyse sein, insbesondere mit Blick auf zeichenhaft ver-

[1] https://clarin-ch.ch [gesehen am 15. April 2023].

[2] ZHAW Digital Discourse Lab: https://www.zhaw.ch/de/linguistik/dienstleistung/digital-discourse-lab/ [gesehen am 15. April 2023].

fasste soziale Wirklichkeit (vgl. Merten 1995, S. 87; Berger und Luckmann 1977). Die für kommunikationswissenschaftliche Forschung relevanten (Text-)Daten (z. B. Texte aus Print- und Onlinemedien, Medienmitteilungen, Beiträge in sozialen Medien, vgl. Brosius et al. 2012, S. 5) werden in der Inhaltsanalyse (vgl. Mayring und Brunner 2007) auch unter Verwendung (halb-)automatisierter Verfahren durchgeführt (vgl. Scharkow 2013), etwa in der Kommunikatorforschung und Medieninhaltsforschung (vgl. Pürer 2014). Hier ergeben sich Nutzungsperspektiven auf korpuslinguistisch aufbereitete Textdaten.

2.1 Zusammensetzung der Daten

Swiss-AL umfasst Teilkorpora für unterschiedliche diskursanalytisch motivierte Verwendungszwecke (Abb. 1)[3]:

a. *Swiss-AL Base* besteht aus automatisiert heruntergeladenen (gecrawlten) Webdaten von öffentlichen Schweizer Kollektivakteuren aus zentralen gesellschaftlichen Teilsystemen: Politik/Verwaltung (z. B. sämtliche Schweizer Kantone, im Bundesparlament vertretene Parteien), Wirtschaft (z. B. die größten Schweizer Branchenverbände und Gewerkschaften), Wissenschaft (z. B. alle Universitäten der Schweiz) und Zivilgesellschaft (z. B. Umweltorganisationen). Die Daten stammen von den zentralen Webseiten dieser Akteure mit einem Schwerpunkt auf News, Medienmitteilungen und Blogbeiträgen[4]). Zusätzlich enthält *Swiss-AL Base* Artikel der auflagenstärksten journalistischen Medien der Schweiz (bereitgestellt für Forschungszwecke durch eine Kooperation mit der Schweizer Mediendaten- dank SMD sowie der Universität Zürich[5]).
   Die Bezeichnung *Base* ist dabei bewusst gewählt: Das Teilkorpus dient zur projektunabhängigen Exploration tendenziell gesamtgesellschaftlicher Diskurse in der Schweiz. Forschende können es einsetzen, um Fragestellungen zu erarbeiten und erste Forschungshypothesen zu formulieren (z. B. „Seit wann befassen sich die Kantone der Deutschschweiz mit Thema X?“, „Sprechen neben journalistischen Medien auch andere öffentliche Akteure über #MeToo? Wenn ja, welche?“).
b. *Swiss-AL Media* eignet sich für die Analyse massenmedialer Kommunikation: Es besteht ausschließlich aus journalistischen Medien der SMD. Neben den auflagenstärksten Medien der Schweiz sind regionale Medientitel der großen Schweizer Verlagshäuser (z. B. Tamedia, NZZ Group) enthalten. Es wird unter anderem als empirische Datenbasis für die Wahl des Schweizer Wortes des Jahres eingesetzt (vgl. Perrin et al. 2020).

[3] Übersicht über Swiss-AL-Teilkorpora und darin enthaltene Akteure (verstanden als Textemittenten): https://swiss-al.linguistik.zhaw.ch/docs/ord/korpora/ [gesehen am 15. April 2023].

[4] Beispiel: Die Webseite des Kantons Zürich (www.zh.ch) publiziert alle aktuellen Meldungen (z. B. Medienmitteilungen) gesammelt auf einer zentralen Unterseite (https://www.zh.ch/de/news-uebersicht.html). Alle dort verlinkten Texte werden automatisiert heruntergeladen und sind Bestandteil von *Swiss-AL Base*.

[5] https://www.liri.uzh.ch/en/services/swissdox.html [gesehen am 15. April 2023].

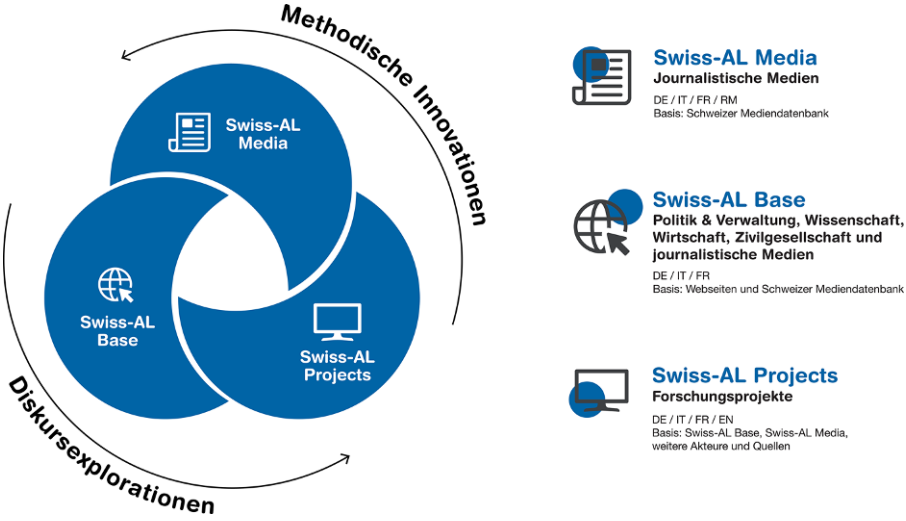

Abb. 1 Swiss-AL Korpusfamilie, bestehend aus Swiss-AL Base, Swiss-AL Media und Swiss-AL Projects

c. Schließlich besteht *Swiss-AL Projects* aus thematisch bestimmten Teilkorpora, die für interne und drittmittelbasierte Forschungsprojekte zusammengestellt und mit der Forschungsgemeinschaft geteilt werden. Hierunter fallen etwa der mehrsprachige Diskurs über COVID-19 oder die Energiestrategie 2050 (vgl. Stücheli-Herlach et al. 2018). Korpora in *Swiss-AL Projects* bestehen v. a. aus Quellen, die bereits in *Swiss-AL Base* oder *Media* enthalten sind, werden aber in Hinblick auf Forschungsfragen und zu lösende Praxisprobleme um weitere Quellen ergänzt (im Falle des COVID-19-Diskurses bspw. um Twitter-Daten). *Swiss-AL Projects* ist nicht auf die Schweiz beschränkt, sondern umfasst Daten aus weiteren europäischen Ländern, z. B. dem deutschen rechtsextremen und islamophoben Weblog PI-News (vgl. Dreesen und Krasselt 2022). Methodische Innovationen innerhalb eines Forschungsprojektes werden an *Swiss-AL Media* und *Base* zurückgespielt (z. B. Verbesserung und Erweiterung von Annotationen).

Die Teilkorpora aus *Swiss-AL Projects* bleiben nach Projektabschluss mehrheitlich statisch. *Swiss-AL Base* und *Media* werden hingegen einmal jährlich aktualisiert. Für Nutzer:innen steht eine Dokumentation zur Verfügung, die für Informationen über die Zusammensetzung der Korpora herangezogen werden kann (Link siehe Fußnote 4).

2.2 Modellierungskriterien

Die Zusammensetzung der Swiss-AL-Teilkorpora ist theoretisch begründet in Umfang und Auswahl, indem sie für forschungsspezifische Zwecke modelliert werden. Swiss-AL folgt hier dem Forschungsdesign *Diskurslinguistik in Anwendung* (DIA), das auf die transdisziplinäre, datenbasierte Analyse von Diskursen abzielt (vgl. Dree-

sen und Stücheli-Herlach 2019). Im Kontext einer transdisziplinären Forschung bedeutet Modellierung zunächst: Die Datenerhebung erfolgt reflektiert mit dem Ziel, dass entstehende Korpora typische Ausprägungen des zu untersuchenden Diskurses enthalten und für die Lösung konkreter kommunikativer Praxisprobleme nützlich sind. Als Modelle sollen die Swiss-AL-Teilkorpora Diskurse abbilden, indem sie insbesondere Parameter auf den Ebenen Akteur (wer spricht), Medialität (in welchem Medium), Semantik (mit welchen Ausdrücken/Konzepten), Zeitraum (wann) und Mehrsprachigkeit (in welchen Sprachen) abbilden (vgl. Stachowiak 1973). Im Fall von *Swiss-AL Base* umfasst dieses Modell die Webseiten von Akteuren zentraler gesellschaftlicher Teilsysteme sowie journalistische Medien. Akteurskategorien sind auf Bundes- und kantonaler Ebene angewandt, um der föderalen, mehrsprachigen Schweiz gerecht zu werden. Für explorative Zwecke ist dieses Modell ausreichend, weil es erlaubt, durch die Äußerungen der untersuchten gesellschaftlichen Akteure der Schweiz Schlüsse über geteiltes Wissen und Sagbarkeit zu ziehen. Bei der Modellierung diskursspezifischer Teilkorpora (*Swiss-AL Projects*) spielen zusätzlich semantische Kriterien eine Rolle. Gemeint ist damit die Entscheidung, welche Wörter obligatorisch in einem Text vorkommen müssen, damit er Teil eines zu untersuchenden thematischen Diskurses ist. Hierfür sind wiederum Explorationen in *Swiss-AL Base* dienlich, um bereits die Modellierung datengeleitet umzusetzen.

2.3 Datenverarbeitung und Zugang

Die Verarbeitung der gesammelten Textdaten erfolgt durch eine modularisierte computerlinguistische Pipeline (vgl. Krasselt et al. 2020). Die Rohdaten können in unterschiedlichem Format vorliegen (z. B. HTML und PDF) und werden durch Webcrawling, APIs und Zugriffe auf existierende Datenbanken automatisiert gesammelt und in einer Datenbank abgespeichert. Relevante, textbezogene Metadaten (z.B. Publikationsdatum, Textemittent, Sprache, URL) werden automatisiert erkannt. Ein spezielles Modul entfernt nicht relevante Elemente aus den Rohdaten. Bei Webseitendaten sind dies z. B. Impressa, Werbung oder Adresszeilen. Nach einer automatisierten Spracherkennung erfolgen Annotationen auf unterschiedlichen linguistischen Beschreibungsebenen, die automatisierte Entfernung von Duplikaten sowie der Import in geeignete Analysetools.

Ein solches Analysetool stellt die Swiss-AL Workbench dar (vgl. Krasselt et al. 2021, vgl. Abb. 2). Sie ermöglicht Forschenden einen Zugriff auf *Swiss-AL Base, Swiss-AL Media* und eine Auswahl aus *Swiss-AL Projects*. Die Workbench ermöglicht wortbasierte Abfragemethoden (z. B. Abfrage von Wort-Distributionen im Zeitverlauf) wie auch Analysen mit Machine- und Deep-Learning-Verfahren (Topic Modeling, Word Embeddings). Der Schwerpunkt der Swiss-AL Workbench liegt auf aggregierenden Methoden der Datenanalyse (*distant reading*), bei denen Erscheinungen auf der Sprachoberfläche quantitativ zusammengefasst werden. Durch die Abfrageoption von aggregierten Daten wird ein Zugang ermöglicht, der den Bestimmungen des schweizerischen Urheberrechtsschutzes entspricht (d. h. kein Volltextzugriff; vgl. Kap. 4 für den Umgang mit dieser Herausforderung).

Die in Abschn. 2.2 und 2.3 vorgestellten Modellierungskriterien und Verarbeitungsschritte verdeutlichen den Wert von Swiss-AL gegenüber Angeboten wie Facti-

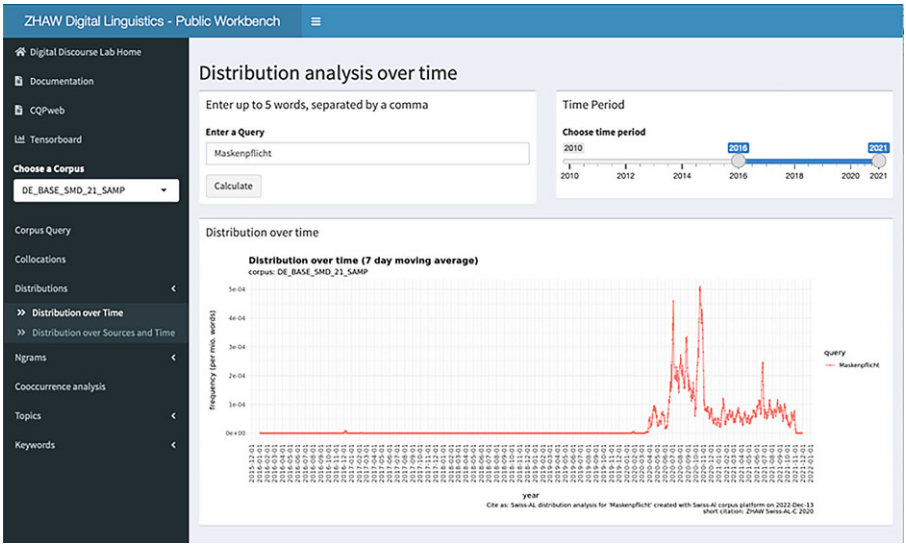

Abb. 2 Swiss-AL Workbench. *Links*: Auswahl eines Teilkorpus und einer gewünschten Analysemethode, Zugriff auf die Dokumentation. *Rechts*: Darstellung der Ergebnisse, hier bspw. der Methode *distribution over time.* Dargestellt ist die Frequenz eines Wortes (*Maskenpflicht)* im Korpus Swiss-AL-Base im Zeitverlauf (2016–2021)

va und Swissdox (beides zentrale Datenbanken für die Kommunikationsforschung): Mit Swiss-AL ist ein Blick auf öffentliche Kommunikation möglich, der nicht nur auf journalistischen Sprachdaten beruht, sondern auf Sprachdaten eines breiten Spektrums von öffentlich relevanten Akteuren. Darüber hinaus werden in Swiss-AL Daten in systematisch durchsuchbarer Form zur Verfügung gestellt. Das ermöglicht insbesondere denjenigen Forschenden Zugang zum datengetriebenen, quantitativen Arbeiten, die nicht über Kenntnisse der automatisierten Datenverarbeitung verfügen.

3 Anwendungsoptionen für die Kommunikationswissenschaft

Die bisherigen Anwendungen von Swiss-AL sind primär geprägt durch sprachbezogene Diskursanalysen in disziplinärer, inter- und transdisziplinärer Nutzung (vgl. Dreesen und Stücheli-Herlach 2019; Stücheli-Herlach et al. (2022). Im Folgenden zeigen wir Anwendungsoptionen für die Kommunikationswissenschaft.

Im Rahmen einer quantitativen, standardisierten Inhaltsanalyse kann Swiss-AL genutzt werden, um durch die Analyse von Mustern an der Sprachoberfläche auf „formale und inhaltliche Merkmale einer größeren Menge an medialen Inhalten“ (Rössler und Geise 2013, S. 271) zu schließen. Zentral ist also der Begriff des *Sprachgebrauchsmusters* (im Sinne rekurrenter, statistisch signifikanter und im jeweiligen Kontext typischer sprachlicher Einheiten, vgl. Bubenhofer 2009): Mit Swiss-AL ist die kommunikationswissenschaftlich zentrale Kategorie *Inhalt* über die Analyse von Sprachoberfläche zugänglich, d. h. über den Gebrauch von Einheiten wie Wörtern oder Wortgruppen und das musterhafte Ko-Vorkommen dieser

Einheiten. Auf der Swiss-AL-Workbench verfügbare Methoden können für die Erfassung zentraler inhaltlicher Kategorien wie Thema, Akteur, Ort oder Bewertung herangezogen werden und eignen sich bspw. für die Analyse von Agenda-Setting, Themendarstellungen und Trends (vgl. Brosius et al. 2012, S. 163). Bereits bei Merten (1995, S. 121) findet sich eine linguistisch orientierte Ausrichtung der Inhaltsanalyse, indem Zugänge wie Wort-, Stil-, Themen- und Assoziationsanalysen den in der Linguistik etablierten Ebenen Syntax, Semantik und Pragmatik zugeordnet werden (vgl. auch Putnam und Fairhurst 2001).

Nutzer:innen der Swiss-AL Workbench entscheiden sich zunächst für eines der verfügbaren Korpora, in dem sie anschließend aggregierende Analysen vornehmen. Aus Perspektive der Kommunikationswissenschaft mag es dabei ungewohnt sein, keinen Zugang zum Einzeltext zu haben. Im Folgenden wird skizziert, dass auch inhaltsanalytisch ausgerichtete Fragen mit der Swiss-AL Workbench bearbeitet werden können (*distant reading*) und sich durch die Kombination mit zusätzlichen Volltextanalysen (*close reading*) ein sinnvoller Methodenmix ergeben kann.

1. Mit den Funktionen *Corpus Query* und *Distribution* kann die Frequenz von Ausdrücken im gesamten Korpus, im Zeitverlauf und in einzelnen Quellen nachgeschlagen werden. Dadurch ergeben sich quantitativ basierte Hinweise auf das Vorhandensein gewisser Themen und Trends in einem Korpus bzw. in einzelnen Quellen des Korpus (vgl. Abb. 2).
2. Ein großes Potenzial für standardisierte Inhaltsanalysen bietet die Funktion *Tensorboard* mit distributionellen semantischen Modellen für einzelne Korpora. Solche *Word-Embedding-Modelle* bieten einen gebrauchsbasierten Zugang zu Semantik, in dem es das Vokabular eines Korpus so in einem dreidimensionalen Raum anordnet, dass Wörter, die in ähnlichen Kontexten gebraucht werden, in räumlicher Nähe im Modell stehen (vgl. Lenci 2018). Ein *Word-Embedding-Modell* bietet einen onomasiologischen Zugriff auf das Korpus und damit die Möglichkeit, nach semantisch äquivalenten Versprachlichungen für ein Konzept zu suchen. Ein Beispiel ist die Suche nach semantischen Äquivalenten zum Wort *Digitalisierung* (vgl. Abb. 3). Diese können dann beispielsweise wiederum mit den Funktionen *Corpus Query* und *Distribution* in ihrer Frequenz und Verteilung untersucht werden (z. B. *digitale Transformation, Automatisierung, technologischer Wandel*). Der Vorteil für standardisierte Inhaltsanalysen besteht darin, Themen und Trends umfassender und datengetrieben untersuchen zu können, ohne Gefahr zu laufen, zentrale Versprachlichungen im Codebuch nicht zu berücksichtigen. Auch für eine Auswahl später zu untersuchender Volltexte ist ein umfangreiches Wissen über verwendetes Vokabular zielführend, da ein Sampling zu untersuchender Texte in der Kommunikationswissenschaft auch auf diesem Parameter beruht. Hier zeigt sich noch einmal der intendierte Nutzen eines Korpus wie *Swiss-AL Base* zur Exploration von Diskursen.
3. Eine weitere inhaltsanalytisch nutzbare Analysemethode stellen Kollokationen dar. Hier finden sich in der kommunikationswissenschaftlichen Literatur äquivalente Methoden. So verwendet Merten (1995, S. 108) den Begriff der *Assoziationsanalyse* bei der Untersuchung von Wortabfolgen. In der Korpuslinguistik werden Kollokationen für die Analyse diskursspezifischer Semantik von Ausdrücken

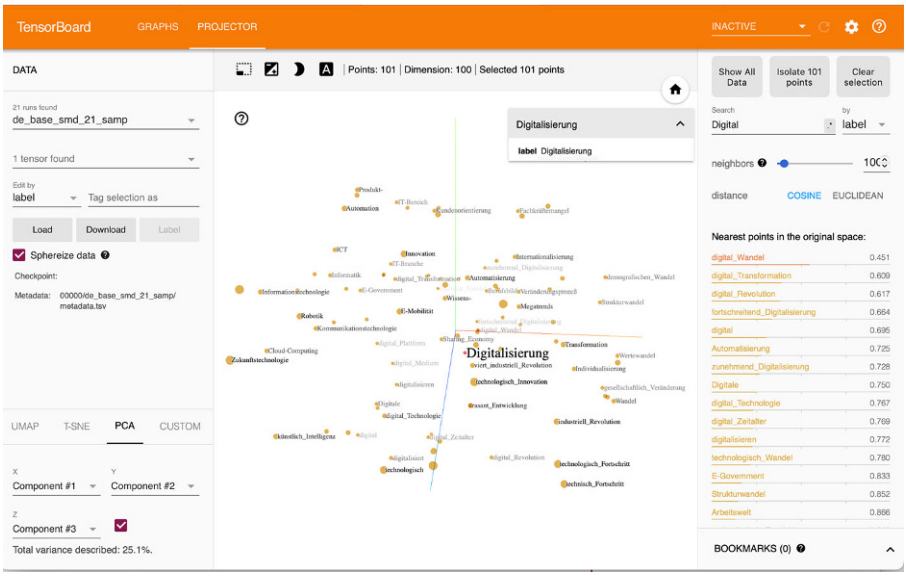

Abb. 3 Die Wortform *Digitalisierung* im Word Embedding Modell für den journalistischen Teil von Swiss-AL Base. Dargestellt werden die sog. 100 *Next Neighbors*, d. h. Wortformen/Bi-Gramme, die in ähnlichen Kontexten gebraucht werden und dadurch über eine ähnliche Semantik verfügen

genutzt (vgl. Baker et al. 2012). Kollokationen sind Ausdrücke, die signifikant häufig miteinander auftreten und stellen eine Operationalisierung der distributionellen Analyse dar: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache." (Wittgenstein 2019, §43) Für Inhaltsanalysen können Kollokationen genutzt werden, um typische thematische Kontexte eines Wortes sowie typische Konnotationen auf Ebene eines Korpus zu untersuchen. Eine Kollokationsanalyse für das Wort *Integration* im journalistischen Teil von *Swiss-AL Base* zeigt beispielsweise eine starke thematische Fokussierung auf die Integration von Personen, die aufgrund von Flucht und Migration in die Schweiz kommen (vgl. Abb. 4). Untergeordnet ist hingegen das Thema Integration von Menschen mit körperlichen oder geistigen Behinderungen. Quantitative Befunde wie dieser können zur Formulierung von Hypothesen genutzt werden (bspw. hinsichtlich implizit kommunizierter Stereotypisierungsmuster, vgl. Rössler und Geise 2013, S. 270), die im Anschluss Inhaltsanalysen zu Grunde liegen können.

4. Schließlich steht mit der Funktion Topic Modeling eine Methode zur Verfügung, um die thematische Struktur von Korpora datengeleitet erfassen zu können. Es handelt sich um eine Methode des maschinellen Lernens, bei der co-vorkommende Wörter in Cluster gruppiert werden, die eine hohe thematische Kohärenz aufweisen (Blei 2012; Maier et al. 2018). Die Swiss-AL Workbench bietet vorberechnete Topic Modelle für einzelne Korpora an, die für eine thematische Exploration dieser Korpora genutzt werden können. Ein Vorteil von Topic Modeling besteht insbesondere darin, dass es mit unterschiedlichen Gebrauchskontexten eines Wortes umgehen kann. Für inhaltsanalytische Fragestellungen ist das nützlich, um zu einer thematischen Binnendifferenzierung zu gelangen, bspw. wenn es um die un-

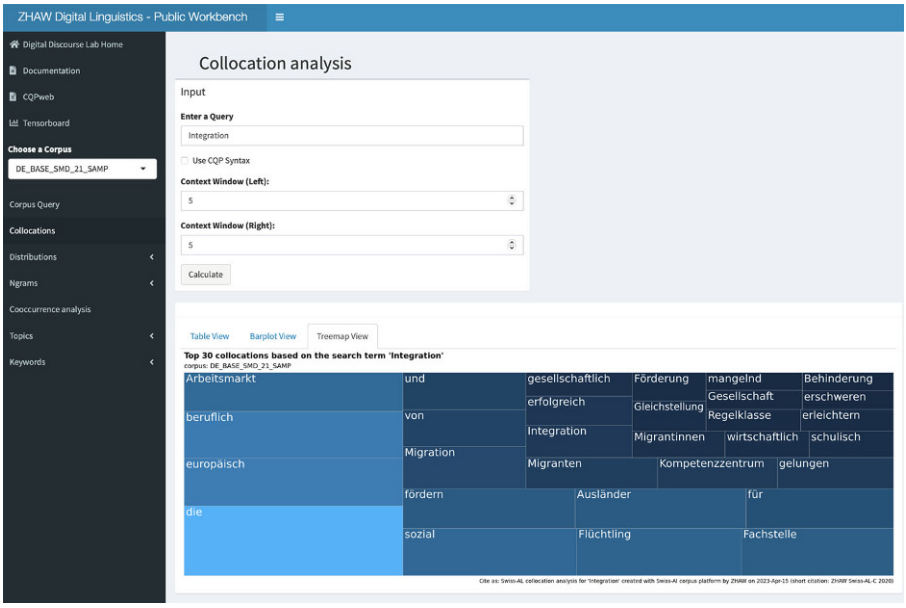

Abb. 4 Kollokationsanalyse auf der Swiss-AL Workbench. Grafisch dargestellt werden die 30 signifikantesten Kollokationen für die Wortform *Integration* in Form einer Treemap, basierend auf einem Kontextfenster („Context Window") von jeweils fünf Wörtern links und fünf Wörtern rechts von *Integration*

terschiedlichen Teilthemen von „Nachhaltigkeit" in der Gesellschaft geht, die sich in je distinkten Topics zeigen können (und so z. B. eine Differenzierung zwischen wirtschaftlicher, sozialer und ökologischer Nachhaltigkeit erlauben).

4 Ausblick

Im Rahmen einer Drittmittelfinanzierung wird Swiss-AL 2023–2024 umfassend überarbeitet, um den Anforderungen einer Open-Research-Data-Plattform zu entsprechen.[6] Legt man die FAIR-Prinzipien zugrunde, ergeben sich für die Bereitstellung von Sprachdaten eine Reihe an Herausforderungen, z. B. ethische, datenschutzrechtliche und urheberrechtliche Beschränkungen (vgl. Maireder et al. 2015), die u. a. Volltextzugriffe verhindern. Zu diesen Herausforderungen zählen auch der Ausbau von Anleitungen zur Verwendung der angebotenen Analysen für nichtlinguistische Disziplinen sowie eine wesentlich detailliertere Dokumentation der ausgewählten Akteure, der Metadaten (z. B. URLs einzelner Texte, Annotationsebenen) sowie von Filterprozessen in der Datenaufbereitung (z. B. Duplikaterkennung, Mindesttextlänge).

Anders als reine *Data Repositories* und Analyseapplikationen, wie sie etwa Text+ und SWISSUbase anbieten, steht für Swiss-AL als Angebot der Ange-

[6] https://www.zhaw.ch/en/linguistics/research/swiss-al-linguistic-open-research-data-practices-for-applied-sciences/ [gesehen am 15. April 2023].

wandten Wissenschaft das Prinzip des Data Life Cycle im Vordergrund, d. h. die projektspezifische Wiederverwendung und projektübergreifende Wiedereinspeisung von Daten.[7] Das Alleinstellungsmerkmal von Swiss-AL besteht darin, Forschenden nicht-linguistischer Disziplinen einen niedrigschwelligen Zugang zu kuratierten, mehrsprachigen, linguistisch aufbereiteten Sprachdaten der öffentlichen Kommunikation zu ermöglichen und webbasierte *Distant-reading-Analysen* anzubieten. Im Zentrum steht neben der Bereitstellung weiterer, insbesondere projektspezifischer Korpora, der Ausbau aggregierender Analysemethoden sowie die Entwicklung von Schnittstellen für Forschende, um eigene Daten in Swiss-AL zu integrieren und mit der Workbench zu analysieren.

- ***Name der Ressource:*** *Swiss-AL (Swiss Applied Linguistics): Eine Plattform für Sprachdaten in den Angewandten Wissenschaften*
- ***Namen der Autor:innen:*** *Julia Krasselt, Philipp Dreesen, Matthias Fluor, Klaus Rothenhäusler*
- ***Link:*** *https://www.zhaw.ch/de/linguistik/forschung/swiss-al/ (weiterführende Informationen) und https://swiss-al.linguistik.zhaw.ch/ (Zugang)*
- ***Kurze Beschreibung:*** *Als mehrsprachige Textsammlung der Schweizer öffentlichen Kommunikation ermöglicht Swiss-AL die datenbasierte Analyse und Simulation gesellschaftlicher Diskurse. Es steht für Forschende und die Öffentlichkeit als Open-Research-Data-Ressource zur Verfügung*

Funding Swiss-AL is currently funded by the Zurich University of Applied Sciences (ZHAW) and swissuniversities. Swiss-AL has also received external funding from the Swiss federal government and cantons, for example from the Swiss Federal Office for Energy (SFOE), the Swiss Federal Office of Public Health (FOPH) and the Swiss National Science Foundation (SNSF).

Funding Open access funding provided by ZHAW Zurich University of Applied Sciences

[7] https://www.text-plus.org/ und https://www.swissubase.ch [gesehen am 15. April 2023].

Literatur

Baker, P., Gabrielatos, C., & McEnery, T. (2012). Sketching muslims: a corpus driven analysis of representations around the word 'muslim' in the british press 1998–2009. *Applied Linguistics*, *34*(3), 255–278.

Berger, P. L., & Luckmann, T. (1977). *Die gesellschaftliche Konstruktion der Wirklichkeit: Eine Theorie der Wissenssoziologie* (5. Aufl.). Frankfurt a.M.: S. Fischer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Brosius, H.-B., Haas, A., & Koschel, F. (2012). *Methoden der empirischen Kommunikationsforschung. Eine Einführung* (6., erweiterte und aktualisierte Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94214-8_10.

Brummans, B. H. J. M., Cooren, F., Robichaud, D., & Taylor, J. R. (2014). Approaches to the communicative constitution of organizations. In L. L. Putnam & D. K. Mumby (Hrsg.), *The SAGE handbook of organizational communication: advances in theory* (3. Aufl.) (S. 173–194). Los Angeles: SAGE.

Bubenhofer, N. (2009). *Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin: De Gruyter.

Craig, R. T. (1999). Communication theory as a field. *Communication theory*, *9*(2), 119–161.

Deppermann, A., Fandrych, C., Kupietz, M., & Schmidt, T. (Hrsg.). (2023). *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial. Berlin*. De Gruyter.

Dreesen, P., & Krasselt, J. (2021). Exploring and analyzing linguistic environments. In F. Cooren & P. Stücheli-Herlach (Hrsg.), *Handbook of management communication* (S. 389–408). De Gruyter.

Dreesen, P., & Krasselt, J. (2022). Medienporträt: PI-NEWS.net. In U. Backes, A. Gallus, E. Jesse & T. Thieme (Hrsg.), *Jahrbuch Extremismus & Demokratie* (S. 237–254). Baden-Baden: Nomos.

Dreesen, P., & Stücheli-Herlach, P. (2019). Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation. *Zeitschrift für Diskursforschung*, *7*(2), 123–162.

Fraas, C., & Pentzold, C. (2016). Diskursanalyse in der Kommunikationswissenschaft. In S. Averbeck-Lietz & M. Meyen (Hrsg.), *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft* (S. 227–240). Wiesbaden: Springer.

Jablin, F. M., & Putnam, L. (2001). Discourse analysis in organizations. Issues and concerns. In L. L. Putnam & G. T. Fairhurst (Hrsg.), *The new handbook of organizational communication* (S. 78–136). Los Angeles: SAGE.

Krasselt, J., Dreesen, P., Fluor, M., Mahlow, C., Rothenhäusler, K., & Runte, M. (2020). Swiss-AL: a multilingual swiss web corpus for applied linguistics. In *Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France* (S. 4138–4144).

Krasselt, J., Dreesen, P., Rothenhäusler, K., & Fluor, M. (2021). A workbench for corpus linguistic discourse analysis. In D. Gromann (Hrsg.), *3rd conference on language, data and knowledge (LDK 2021)* (S. 26:1–26:9). Dagstuhl: Leibniz-Zentrum für Informatik.

Krasselt, J., Dreesen, P., Fluor, M., & Rothenhäusler, K. (2023). Swiss-AL. Korpus und Workbench für mehrsprachige digitale Diskurse. In M. Kupietz & T. Schmidt (Hrsg.), *Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022* (S. 127–142). Tübingen: Narr.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures, 12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754.

Maireder, A., Ausserhofer, J., Schumann, C., & Taddicken, M. (Hrsg.). (2015). *Digitale Methoden in der Kommunikationswissenschaft*. Berlin: Digital Communications Research.

Mayring, P., & Brunner, E. (2007). Qualitative Inhaltsanalyse. In R. Buber & H. Holzmüller (Hrsg.), *Qualitative Marktforschung* (S. 669–680). Wiesbaden: Gabler.

Merten, K. (1995). *Inhaltsanalyse: Einführung in Theorie, Methode und Praxis* (2. Aufl.). Wiesbaden: Springer.

Perrin, D., Whitehouse, M., Liste Lamas, E., & Kriele, C. (2020). Diskursforschung im Schaufenster. Ein transdisziplinärer Ansatz zur Ermittlung und Vermittlung von Wörtern des Jahres. *Zeitschrift für Diskursforschung*, *2*, 164–189.

Pürer, H. (2014). *Publizistik- und Kommunikationswissenschaft* (2. Aufl.). Stuttgart: utb.

Putnam, L., & Fairhurst, G. (2001). Discourse Analysis in Organizations. Issues and Concerns. In F. Jablin & L. Putnam (Hrsg.), *The New Handbook of Organizational Communication* (S. 78–136). Thousand Oaks, CA: SAGE.

Rössler, P., & Geise, S. (2013). Standardisierte Inhaltsanalyse: Grundprinzipien, Einsatz und Anwendung. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 269–287). Wiesbaden: Springer.

Scharkow, M. (2013). Automatische Inhaltsanalyse. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 289–306). Wiesbaden: Springer.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien: Springer.

Stücheli-Herlach, P., Ehrensberger-Dow, M., & Dreesen, P. (2018). *Energiediskurse in der Schweiz: Anwendungsorientierte Erforschung eines mehrsprachigen Kommunikationsfelds mittels digitaler Daten*. Winterthur: ZHAW.

Stücheli-Herlach, P., Dreesen, P., & Krasselt, J. (2022). Öffentliche Diskurse modellieren und simulieren. Wege der transdisziplinären Diskurslinguistik. *Zeitschrift für Diskursforschung. Sonderausgabe zum zehnjährigen Jubiläum*, *10*(2), 245–256.

Wehmeier, S., Rademacher, L., & Zerfaß, A. (2013). Organisationskommunikation und Public Relations: Unterschiede und Gemeinsamkeiten. Eine Einleitung. In A. Zerfaß, L. Rademacher & S. Wehmeier (Hrsg.), *Organisationskommunikation und Public Relations* (S. 7–24). Wiesbaden: Springer.

Wittgenstein, L. (2019). *Philosophische Untersuchungen* (9. Aufl.). Frankfurt a.M.: Suhrkamp.

Dr. Philipp Dreesen ist Professor für Digitale Linguistik und Diskursanalyse an der Zürcher Hochschule für Angewandte Wissenschaften und Mitglied des Kernteam ZHAW Digital Discourse Lab.

Dr. Julia Krasselt ist wissenschaftliche Mitarbeiterin an der Professur Digitale Linguistik und Diskursanalyse an der Zürcher Hochschule für Angewandte Wissenschaften und leitet das Projekt *Swiss-AL*. Sie ist ebenfalls Mitglied des Kernteam ZHAW Digital Discourse Lab.