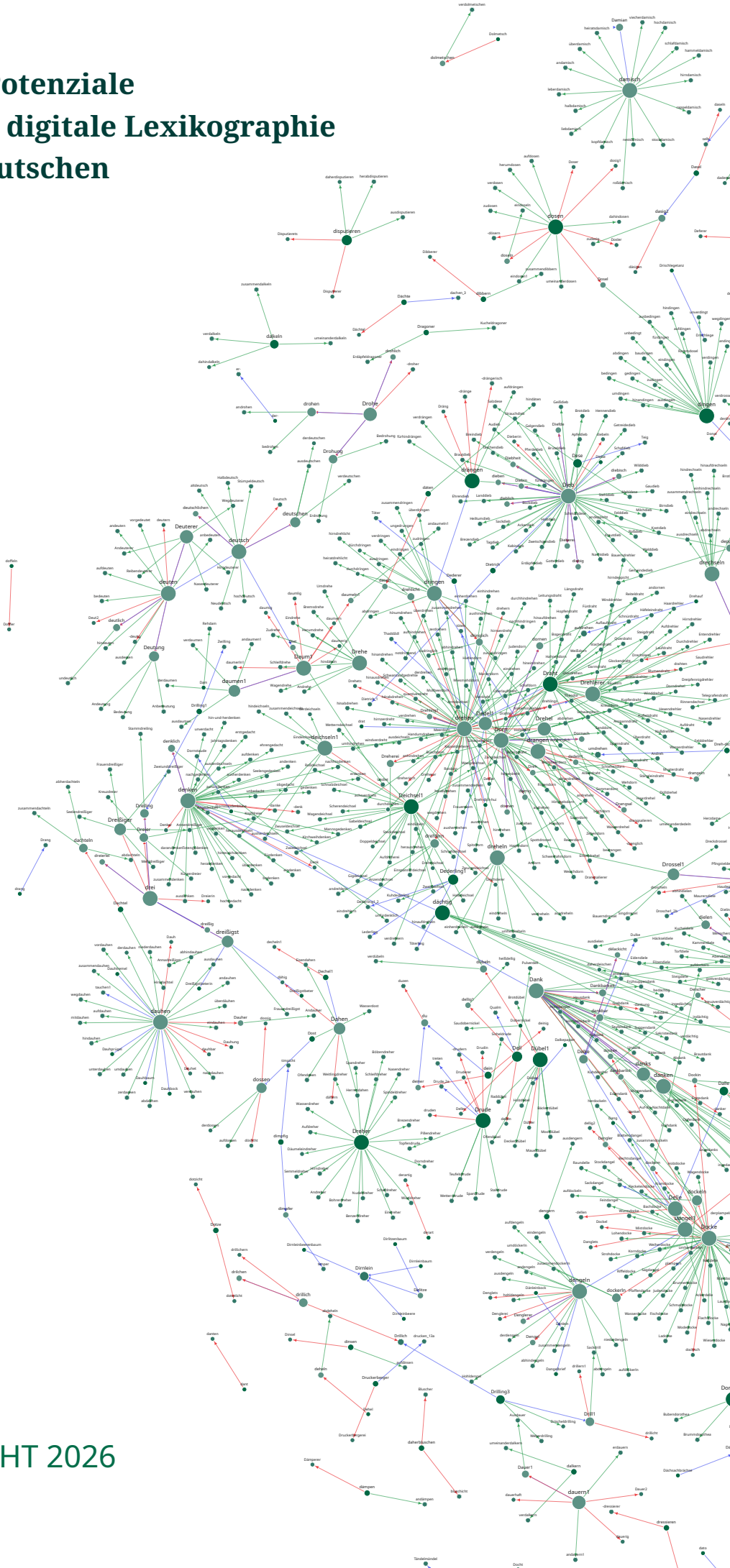
Neue Potenziale für die digitale Lexikographie des Deutschen



FORTSCHRITTSBERICHT 2026

Einleitung und Kurzfassung

Nach genau einem Jahr Laufzeit fasst der vorliegende Fortschrittsbericht die ersten Ergebnisse des Projekts "Neue Potenziale für die digitale Lexikographie des Deutschen" (PDL) zusammen und bietet einen Ausblick auf das zweite Jahr.

Durch die unlängst erfolgte Bewilligung des „Akademienzentrums Digitale Lexikographie des Deutschen“ (ADL) durch die GWK wird die Zukunftsperspektive konkret, und die Inhalte des ADL rücken weiter in den Vordergrund.

Bezugspunkt des Jahresberichtes ist der zu Beginn aufgestellte Projektplan, der in mehreren Arbeitsbereichen verschiedene Pflichtaufgaben und optionale Aufgaben definierte.

Im Arbeitsbereich „Technische Grundlagen“ wurden verschiedene Formate, Technologien und Strukturen mit Blick auf die Antragsinhalte „Wortschatzportal“ und „Forschungsdatenbank“ exploriert und umgesetzt. Dazu gehören die Datenspeicherung in verschiedenen Datenbanksystemen (mit unterschiedlichen Einsatzgebieten), verschiedene Datenstrukturen und -formate, sowie die automatisierte Transformation zwischen diesen Formaten. Mit der Webapplikation lexoterm.de wurde eine erste Version einer voll funktionalen Webseite inklusive Wörterbuchsuche realisiert. Komplementiert wurden diese Schritte durch nötige technische Begleitmaßnahmen wie den Betrieb virtueller Server am LRZ, die versionierte Code-Verwaltung in Github, das containerbasierte, leistungsfähige Deployment mittels Dokploy, sowie Tests der Leistungsfähigkeit der Infrastruktur durch aufwändige Simulationen (sog. Loadtest). Schließlich wurden auch erste Schnittstellen mit und für Partner im Produktiveinsatz realisiert (LexFCS-Endpunkt für Text+).

Einen weiteren wichtigen Arbeitsbereich stellten Werkzeuge für den lexikographischen Workflow dar. Bei allen Werkzeugen wurde besonders viel Wert auf eine leichte Nutzbarkeit ohne Schulung oder technisches Fachwissen gelegt, weshalb sie durchgehend als Webapplikationen mit grafischem Interface ausgeführt sind. Das Tool „Wörterbuchkonsistenzprüfung“ bietet eine umfangreiche, stets wachsende Sammlung an Prüfmodi zur Konsistenzüberprüfung nach technischen, inhaltlichen und stilistischen Kriterien. Die „Sachgruppenvorhersage“ erlaubt es, verschiedene eigene Machine-Learning-Modelle auf Basis von Wörterbuchdaten zu trainieren, um automatisiert Sachgruppen vorherzusagen. Die beiden neuen Register zu Fischers Schwäbischem Wörterbuch und zu zahlreichen Quellen des Bayerischen Wörterbuchs erleichtern ganz konkret den Redaktionsalltag, ebenso wie ein Tool zur morphologischen Segmentierung. Für das Forschungsprojekt zur „Zeitenwende im Wortfeld“ wurde eine aufwändige Applikation entwickelt, die semantische Vektoren visualisiert – natürlich generalisierbar und nachnutzbar. Diese und weitere Tools sind im entsprechenden Abschnitt weiter unten ausführlich dargestellt.

Manche der Arbeiten legen auch Konferenzeinreichungen bzw. wissenschaftliche Veröffentlichungen nahe. Hier sind insbesondere eine gemeinsame Einreichung mit Kolleg*innen der Abteilung Digital Humanities sowie einem externen Autor zur drei verschiedenen Ansätzen der Sachgruppenvorhersage zu erwähnen. Das bereits erwähnte „Zeitenwende“-Projekt zum semantischen Wandel in Kooperation mit der Deutschen Nationalbibliothek verbindet mehrere neue analytische Ansätze mit einer aufwändigen Korpusauswertung und einer anspruchsvollen Visualisierung.

Die anstehende Gründung des ADL legt einen neuen Fokus auf die Zusammenarbeit mit den weiteren Akademien im Verbund. Konnte bisher schon mit drei Akademien die Bereitstellung von Testdatensätzen aus dort verantworteten Wörterbüchern erreicht werden, so gilt es nun, diesen inhaltlichen, technischen und materiellen Austausch auszuweiten. Auch einige technische Details werden in der Folge dieses Austausch an Kontur gewinnen, wodurch auch der Zusammenarbeit mit dem LRZ zusätzliches Gewicht zukommt. Entsprechende Absprachen vor allem hinsichtlich Inferenz-Rechenleistung

sind bereits erfolgt. Die Ausschreibung und Besetzung der gemäß Antrag vorgesehenen Stellen stellt ebenfalls eine wichtige Weichenstellung für den weiteren Erfolg dar.

Neben diesen bereits auf die Zukunft gerichtete Aufgaben tritt der weitere Ausbau der Funktionalität für lexoterm.de, die Fortschreibung eines Konzepts zum fairen Datenaustausch zwischen den Akademien sowie die Kooperation mit der BBAW im Bereich Linked Open Data (vgl. ELEXAI). Weiterhin ist das Training eines Instruct-Modells auf Basis vorhandener Dialekt-Wörterbuchdaten vorgesehen.

Projektplan und Schwerpunkte

Der zu Beginn des Projektes PDL erstellte Projektplan, der mit allen Projektbeteiligten abgestimmt wurde, gibt die Leitplanken und Ziele des Projektes vor. Er unterscheidet zwischen organisatorischen, technischen, wissenschaftlichen und kommunikativen Arbeiten und trennt zwischen Kernmodulen, die in jedem Fall erreicht werden müssen, und optionalen Modulen, aus denen je nach Projektfortgang und Schwerpunktbildung gewählt werden kann.

Der Projektplan legt auch fest, dass einer vielseitigen Exploration des Aufgabenfeldes im Rahmen des Pilotprojektes Vorrang gegeben wird; im Gegenzug wird nicht jede Aufgabe in der maximal möglichen akademischen Tiefe ausgearbeitet. Diese nötigen und wünschenswerten Vertiefungen sind gerade Aufgabe der Leads im Rahmen des ADL, und eine Vorwegnahme ist einerseits nicht leistbar, und würde andererseits den Ablauf des Hauptprojektes unterminieren. Gleichwohl sollen alle Arbeiten handwerklich und wissenschaftlich fehlerfrei ausgeführt werden und als Basis für darauf aufbauende Arbeiten dienen können.

Dieser Ansatz hat zudem den Vorteil, sowohl die Kompetenzprofile für die Stellenbesetzungen nachschärfen, wie auch Bereiche, die weiteren Kompetenzaufbau erfordern, frühzeitig identifizieren zu können. Weiterhin erklärt sich damit das vielseitige Werkzeugportfolio, das im weiteren Verlauf dargestellt wird.

Zudem hat sich gezeigt, dass manche Themenbereiche, insbesondere die Detailspezifikation der Forschungsdatenbank, einer intensiveren Befassung durch ein größeres Plenum der beteligten Partnerakademien bedürfen – deren Vorstellungen, Wünsche und auch Vorbehalte sind maßgeblich für das Design und den Erfolg einer solchen Anwendung.

Der untenstehende Zeitstrahl sowie die Berichte zu technischen Arbeit, Werkzeugen, Forschungsarbeiten und Organisation analysieren den Stand im Detail. Zusammenfassend lässt sich festhalten, dass die Kernmodule alle umgesetzt sind, und sich bereits einige optionale Module in einem fortgeschrittenen Stadium befinden. Der Zeitplan wurde weitestgehend eingehalten und eine erfolgreiche Erfüllung der angestrebten Aufgaben sollte keine besondere Schwierigkeit darstellen.

Aktueller Zeitplan und Meilensteine

2025
Oktober
01

Projektstart

Projektabstimmung mit Prof. Habermann & Prof. Evert
kontinuierlich während der Projektlaufzeit

Diskussion Technischer Grundlagen mit LRZ
kontinuierlich während der Projektlaufzeit

Vernetzung der Kooperationspartner des PDL
Schwerpunkt zu Projektbeginn

Basis-Daten-Pipeline
Quell-XML → Transformation Zielformat → Speicherung DB → Ausspielung Web

November
02

Projektwebsite I
Informations-Webseite als Teil der BAdW-Präsenz; ohne Funktionalität

Definition Datenformat
Festlegung eines (vorläufigen) Datenformats für die zentrale Speicherung

Definition Techstack
Festlegung eines Techstacks; Anpassungen während der Projektlaufzeit

Kommunikationsstrategie
Ausarbeitung eines Kommunikationskonzeptes; Umsetzung fortwährend

Dezember
03

Werkzeuge:
Wörterbuchgraph, Visueller Wortausschluss, Sachgruppenvorhersage
Erste funktionale Versionen online

2026
März
06

Projektwebsite II: Portal
Erste Version von Lexoterm.de online, mit Wörterbuchfunktion (BDO-Bestand)

Datentransformer
Framework für erweiterbare Format-Transformer: Quell-Wörterbücher zu Zielformat

April
07

Anbindung BDO an LexFCS/Text+
In Abstimmung mit SAW und BBAW

Werkzeuge:
Wörterbuchkonsistenzprüfung, Morphologische Segmentierung
Erste funktionale Versione online

Integration Testdaten DWDS
Automatisierte Transformation von DWDS-Format in Lexoterm-Format, Bereitstellung

Mai 08

Verbesserungen LexoTerm-Backend
Dokploy für einfaches Deployment, Durchführung von Lasttests

Werkzeuge: Bedeutungsnummerierung
Prüfung von Nummerierungskonsistenz in Wörterbüchern

Juni 09

Konzeption Erweitertes Datenformat
Erprobungen von hybriden SQL-Formaten

Werkzeuge: Fischer-Register, BWB-Meta-Register
Neue Tool-Klasse: Register auf Basis von Quelltexten, inkl. OCR-Modell-Training

Juli 10

Detailkonzept und Projektplan für "Zeitenwende im Wortfeld"
Literatursichtung, Arbeitsplan, Abstimmung DNB

August 11

Werkzeuge:
Interaktives ADL-Organigramm, Visualisierung Semantischer Vektoren
VSV-Tool als wesentliche Komponente des DNB-Projektes

September 12

Werkzeuge:
Modellspielplatz (+ neue Funktionen für Wörterbuchkonsistenzprüfung)
Testumgebung für offene ML-Modelle, z.B. OCR/HTR für Sütterlin

Aktueller Zeitpunkt

Dezember 15

Projektwebsite II: Portal
funktionale öffentliche Version mit Zugangskontrolle (academic credentials)

Abschlusss Datenauswertung für Projekt "Zeitenwende im Wortfeld"
Datenerhebung an der DNB im November, Auswertung Dezember, Publikation 2027

Integration Testdaten DGS und AWB
Testdaten zum Gebärdensprachen-Korpus und zum Althochdeutschen Wörterbuch wurden von AWHH und SAW bereits bereitgestellt, Transformation und Integration ggw. ausstehend

2027 Januar 16

Die Arbeitsschwerpunkte für die Restlaufzeit mit parallelem Anlauf des ADL werden in einem gemeinsamen Treffen der Projektbeteiligten im Herbst 2026 festgelegt. Offene Punkte gemäß ursprünglichem Projektplan inkludieren:

Projektwebsite II: Portal *(finale Version)*

OPTION **Annotations-Copilot**

OPTION **LLM-gestützte Lexikographie-Workflows**

OPTION **Vektor-Datenbank / Embeddings**

OPTION **Lexikographischer Wissensgraph**

KOOPERATIONSPROJEKT **Transfer Learning für Dialekte** *(BAVA-MInD)*

Für weitere Projektideen siehe Kapitel "Ausblick".

Technische Arbeiten: Forschungsdatenbank

Das ADL wird einen stetig wachsenden Bestand lexikographischer Daten hervorbringen. Um diese Daten für quantitative Methoden wie statistische Analysen, maschinelles Lernen und LLMs vorzubereiten, wurden in PDL die konzeptuellen und technischen Grundlagen für den Aufbau einer Forschungsdatenbank und einer zugehörigen Such-Oberfläche gelegt. Die Forschungsdatenbank soll dafür ausgewählte Elemente der Quelldaten über geeignete Schnittstellen in einem performanten, hoch-skalierbaren Format bereitstellen.

Die Forschungsdatenbank komplementiert somit die föderierte Suche, die im Lead der SAW den gesamten Wörterbuchbestand aller teilnehmenden Institutionen jeweils in Echtzeit und in allen Details abbilden soll.

Datenspeicherung

Da die Forschungsdatenbank primär zur Verarbeitung großer Datenmengen dient und nur Teile der originalen Wörterbuchstrukturen abbildet, müssen bei der Auswahl des Datenformats und der Datenbank-Technologie andere Aspekte berücksichtigt werden als bspw. bei der redaktionellen Bearbeitung von Wörterbuch-Artikeln. De facto erfordert die Forschungsdatenbank nutzerseitig einen hochperformanten, ausschließlich lesenden Zugriff auf strukturierte Daten mit komplexen Suchabfragen. Hierfür legen unsere Versuche mit verschiedenen Datenbanktechnologien nahe, dass ein klassisches, relationales Datenbank-Paradigma wie PostgreSQL langfristig erste Wahl ist.

DATENBANK-TECHNOLOGIE

Für die Pilotphase wurde zunächst MongoDB als Interims-Technologie eingesetzt. MongoDB ist ein Dokumenten-orientiertes (NoSQL) Datenbank-System, das sich aufgrund seiner Flexibilität sehr gut eignet, um in kurzer Zeit verschiedene Datenformate zu erproben, ohne sich auf ein verbindliches Schema festzulegen. Ergebnis dieser Versuche ist eine Struktur, welche die Kern-Elemente der Quelldaten (wie Lemma, Grammatik und Bedeutungen) sehr gut abbildet. In der nächsten Projektphase wird dieses Schema in PostgreSQL als Dauerlösung überführt. Es ist wichtig, zu erwähnen, dass die PostgreSQL-Datenbank nicht die einzige Technologie bleiben muss: So kann es sich z.B. anbieten, die Daten für spezialisierte Anwendungsszenarien über den Einsatzzweck der Forschungsdatenbank hinaus zusätzlich etwa als XML (z.B. in TEI Lex0) oder in Form einer Graph-Datenbank vorzuhalten.

DATENFORMATE

In der Forschungsdatenbank werden u.a. atomare Felder für bspw. Lemmata und Wortarten, aber auch verschachtelte Informationen (Bedeutungen und Unterbedeutungen) abgedeckt. Letztere werden sowohl in XML wie auch Dokumenten-orientieren Datenbanken mittels verschachtelter Strukturen abgebildet. In relationalen Datenbanken werden alle Bedeutungen hingegen in einer flachen Tabelle abgelegt und ihre hierarchischen Beziehungen mithilfe von Zeigern (Kindknoten → Mutterknoten) gespeichert. Diese Struktur hat technisch große Vorteile und bedeutet keinerlei Informationsverlust.

Ein ähnliches Vorgehen ermöglicht es ebenfalls, verschachtelte XML-Auszeichnungen in Fließtext (Mixed Content) als flache Struktur abzulegen. Dafür werden die Auszeichnungen extrahiert und für jedes Element ein Start- und End-Index gespeichert, die den Skopus der Auszeichnung auf dem Quelltext wiedergeben. Dieses Vorgehen (Standoff-Annotation) ist in der Computerlinguistik weit verbreitet und ermöglicht die Abbildung arbiträr verschachtelter und sogar überkreuzender Auszeichnungen sowie eine einfache Ergänzung um zusätzliche Informationen. Dies ermöglicht es z.B., Suchtreffer als weitere Auszeichnungsschicht zu definieren und im Interface genau zu markieren, welche Textseg-

mente von der Suche erfasst wurden, was einen enormen Effizienzgewinn bei der manuellen Sichtung von Ergebnislisten mit sich bringt und mit hierarchischen Formaten (XML) nur sehr eingeschränkt möglich ist.

TRANSFORMER

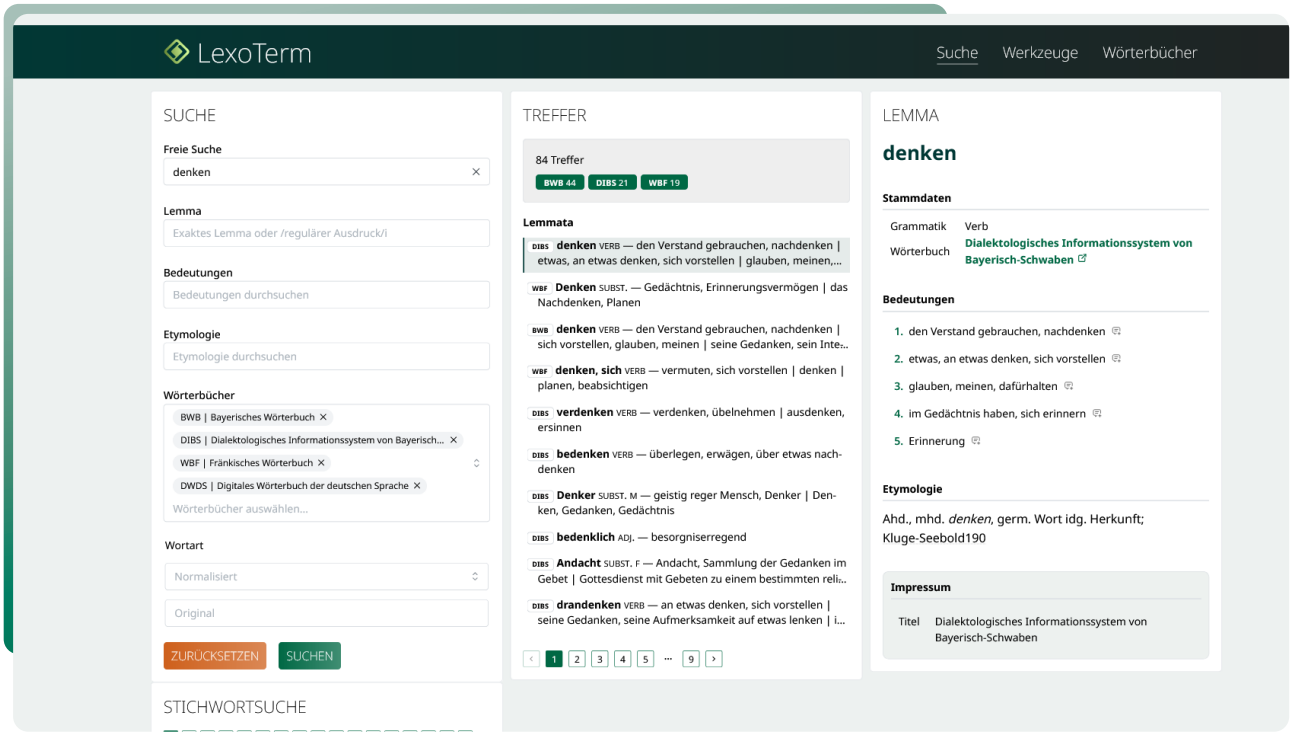
Um in XML vorliegende Inputdaten in das o.g. flache Format zu überführen, wurden dedizierte Konvertierungsmodule in Python implementiert, die die gewünschten Informationen aus den Quelldaten extrahieren und in das Zielformat überführen. Bei der technischen Umsetzung wurde darauf geachtet, ein gut wartbares, deklaratives Programmierparadigma zu verwenden, das sicherstellt, dass der Quellcode auch bei vielen verschiedenen Eingabeformaten und komplexen Datenumwandlungen gut nachvollziehbar und wartbar bleibt.

Diese Transformer existieren sowohl für die Konvertierung von spezifischen Quelldatenformaten (BDO, DWDS, DGS, AWB) in das LexoTerm-Zielformat, wie auch für die Konvertierung ausgesuchter Quellformate in TEI Lex-0. TEI Lex-0 eignet sich in seiner unveränderten Form nicht als effizientes Datenbankformat, sondern dient vielmehr als *Austauschformat*. Es wird angestrebt, alle Wörterbuchartikel jederzeit in TEI Lex-0 exportieren zu können.

Datendarstellung (Front-End)

ENTWICKLUNG LEXOTERM.DE

Die Forschungsdatenbank wird nicht nur zum Abruf der Rohdaten über eine API, sondern ebenfalls über das Web-Interface *LexoTerm* (www.lexoterm.de) verfügbar gemacht. Der Name LexoTerm ist ein Kunstwort aus den Begriffen Lex(ikon) und Term(ini) und spielt auf die Bezeichnung exotherm an, die physikalische und chemische Prozesse beschreibt, bei denen gebundene Energie in Form von Wärme freigegeben wird. Diesem Bild folgend ist die Hoffnung, dass mithilfe von LexoTerm linguistische und lexikographische Erkenntnisse, die in den Quelldaten derzeit noch verborgen liegen, freigesetzt werden können.



Lexoterm.de: Wörterbuch-Suchmaske

Das Interface ist als eine auf React und TypeScript basierende Web-Anwendung umgesetzt. Es bietet neben dem Zugriff auf die Forschungsdatenbank Informationen über das Projekt und vernetzt eigene, externe und Partner-Werkzeuge für verschiedene lexikographische Anwendungen. Die Kernfunktion, d.h. die Suchansicht, bietet verschiedene Filter, um die im oben beschriebenen Verfahren aufbereiteten Wörterbuchdaten abzufragen. Derzeit ist die Suche nach Lemmata, Bedeutungen, Etymologie und Wortart über den BDO-Datenbestand möglich.

Weiterhin ist eine kleine Zahl an Testdaten aus dem DWDS integriert, und es liegen weitere Testdaten des Deutschen Gebärdensprachenkorpus und des Althochdeutschen Wörterbuchs bereit, die noch den Transformationsschritt durchlaufen müssen. Ob zu einem späteren Zeitpunkt alle Wörterbücher im ADL-Verbund in ihrer Vollversion oder gekürzter Form als Verzeichnis über das FDB-Portal abrufbar sein sollen, ist eine Entscheidung, welche die ADL-Partner gemeinsam treffen müssen. Die technischen Grundlagen sind gelegt.

Ebenfalls vorbereitet ist die Integration von Korpora, um neben lexikographischen auch korpuslinguistische Fragestellungen bearbeiten zu können. Viele attraktive Korpora (etwa aktuelle Zeitungen oder Medien) unterliegen aber Nutzungsbeschränkungen der Rechteinhaber, so dass ein Ausbau dieser Funktionalität vermutlich auf absehbare Zeit allenfalls in geringem Umfang wird erfolgen können.

Die Verbindung zwischen *Wörterbüchern* und *Korpora* stellen schließlich die *Werkzeuge* dar, die zu einem späteren Zeitpunkt integraler Bestandteil von LexoTerm werden sollen, und dann das Fundament des Aufgabenbereichs "Algorithmisches Rechnen" bilden. Gegenwärtig sind die Werkzeuge noch als unabhängige, nicht voll integrierte Web-Applikationen realisiert (Details siehe unten).

Die Forschungsdatenbank bzw. LexoTerm sind nicht als Angebote für beliebige Nutzer konzipiert, sondern rein für wissenschaftlich Arbeitende. Sie sollen explizit kein konkurrierendes Angebot zu erfolgreichen existierenden Angeboten wie z.B. dwds.de darstellen. Gleichzeitig soll das Angebot der sprachwissenschaftlichen Community leicht zugänglich sein. Infolgedessen werden Zugangsmechanismen integriert, die die Nutzung nach Nachweis der Zugehörigkeit zu einer akademischen Einrichtung ermöglichen.

Backend & Deployment

VIRTUELLE MASCHINEN

LexoTerm sowie alle weiteren im Rahmen von PDL entwickelten Werkzeuge werden auf einem virtuellen Server am Leibniz-Rechenzentrum gehostet. Dort profitiert das System von hoher Verfügbarkeit, betreuter Wartung des Host-Systems sowie umfangreichen Backup-Angeboten. Der Hosting-Server wird ggw. ergänzt durch eine Inferenz-VM zur Durchführung von KI-Workloads.

DOKPLOY

Zur Überwachung und Verwaltung der Services wurde im Laufe der Pilotphase die grafische Administrationsoberfläche *Dokploy* eingesetzt, was das Deployment vorhandener und neuer Services massiv vereinfacht und auch bei einer stetig wachsenden Anzahl von Services für Übersichtlichkeit sorgt.

GITHUB CI/CD

Da die Hauptservices – die Forschungsdatenbank und LexoTerm – dauerhafte Bestandteile der ADL-Infrastruktur sein werden, wurde bei ihrer Umsetzung bereits ganz zu Beginn der Entwicklung darauf geachtet, Industrie-Standards in Quellcode-Strukturierung und Qualitätssicherung anzuwenden und gleichzeitig den Open Source-Gedanken des Projekts in Form von offenem, über Github geteiltem

Quellcode umzusetzen. Dabei werden vor jedem System-Update automatisierte Tests durchgeführt und Deployment-Workflows angestoßen, die ein reibungsloses Zusammenspiel mit dem o.g. Dokploy bieten. In der nächsten Projektphase, wenn weitere Entwickler*innen an der Codebasis arbeiten, werden diese Automatisierungen und Coding-Standards weiter konkretisiert werden.

LOADTESTS

Auch für die Entwicklungsumgebungen wurden bereits Lasttests mithilfe der Testumgebung *K6* durchgeführt. Für die insbesondere in der Pilotphase zu erwartenden Nutzerzahlen steht mehr als genug Reserve zur Verfügung. In seiner aktuellen Ausbauform, die sich am unteren Ende der verfügbaren Hardware bewegt, bestehen Reserven für bis zu 60 konkurrente Nutzer (nur oberstes Perzentil mit Ladezeiten > 800ms). Dieses Gesamtbild ändert sich mit der Integration weiterer Services und aufwändigerer Seitenbestandteile – im Gegenzug besteht noch sehr viel Spielraum bei der Skalierung der VMs, bis hin zum Einsatz mehrerer VMs im Verbund.

Schnittstellen

APIs

Ein leichter Datenzugang für legitimierte Benutzer ist ein zentrales Element der FDB bzw. von LexoTerm. Der Zugriff wird daher auch per API möglich sein, wobei diese umfangreiche ausgestaltet sein soll und nicht nur einfache Basisdaten liefert. Auch die Werkzeuge verfügen in vielen Fällen über API-Schnittstellen, die dann auch Zugriff auf dedizierte Hardware erlauben.

LEX-FCS-ENDPUNKT

Für die interdisziplinäre Zusammenarbeit ist es entscheidend, Daten und Informationen in einer Form anzubieten, die in den jeweiligen Fachbereichen gebräuchlich ist bzw. vom Gegenüber erwartet wird. Daher wird auch diese Form von Schnittstellen Teil von LexoTerm sein. Beispielhaft sind sie bereits in einem Endpunkt realisiert, der die BDO-Daten für die föderierte Suche von Text+ bereitstellt.

Technische Arbeiten: Werkzeuge & Algorithmisches Rechnen

Auf den folgenden Seiten werden die bisher realisierten Werkzeuge ausführlich dargestellt. Alle Werkzeuge stehen unter den angegebenen Links auch als lauffähige Applikationen zur Verfügung.

Bei der Entscheidung, welche Werkzeuge realisiert werden sollen, gab es zwei Impulsrichtungen: Die "schwergewichtigeren" Tools sind aus Anregungen aus den Wörterbuchredaktionen und Fachabteilungen an der BAdW heraus entstanden, und sollen unmittelbaren Nutzen für die tägliche lexikographische Arbeit entfalten. Die "spielerischen" oder "funktionalen" Tools sind kleinere Proofs of Concept, oder erfüllen bestimmte technische Anforderungen.

Alle Werkzeuge sind so konzipiert, dass sie grundsätzlich von einem möglichst großen Nutzerkreis verwendet werden können. Es gibt keine enge Beschränkung auf bspw. BDO-Daten, wenngleich einzelne Funktionen spezifisch auf Eigenheiten der BDO-Wörterbücher abgestimmt sind; dies lässt sich aber für weitere Wörterbücher problemlos anpassen.

Alle Werkzeuge sind mit modernen, benutzerfreundlichen grafischen Interfaces ausgestattet. Sie sollen damit eine möglichst niedrige Nutzungsschwelle aufweisen und schnell im Redaktionsalltag einsetzbar sein, von Mitarbeiter*innen jeglichen technischen Levels. Im Umkehrschluss wurde auf Kommandozeilen-Funktionen weitgehend verzichtet. Ein Nachteil dieses Ansatzes ist, dass auch relativ simple Funktionen durch das erforderliche GUI schnell ein großes Volumen an Dateigröße und Code-Umfang erreichen. Dem Endandwender, der nur eine Web-App nutzt, kann das gleich sein – die Entwickler stellt das vor gewisse Herausforderungen hinsichtlich Wartung und Pflege.

REPOS AUF GITHUB

Die meisten Code-Projekte des Pilotprojektes sind frei auf Github verfügbar. Im Regelfall sind diese zudem mit permissiven Lizenzen ausgestattet, die eine Weiternutzung leicht möglich machen.

github.com/orgs/pdl-lex/repositories

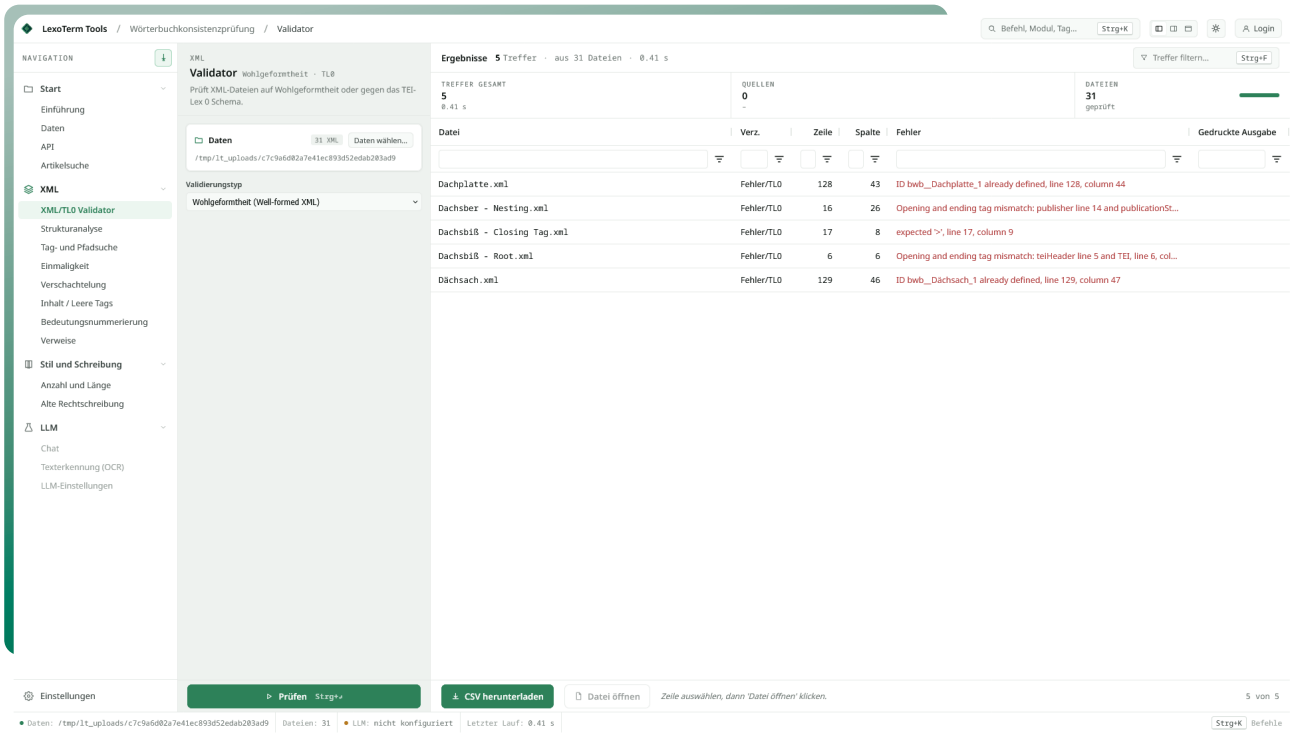
Sofern einzelne Repos nicht öffentlich sind, lohnt eine Kontaktaufnahme – wir geben uns alle Mühe, den Code bereitzustellen.

Alle Werkzeuge, bei denen dies konzeptionell Sinn macht, sollen zudem API-Schnittstellen aufweisen: Sie können also auch im Remote-/Skript-/Batch-Betrieb ohne GUI genutzt werden. Beispielsweise lassen sich alle Prüffunktionen des Wörterbuchkonsistenzprüfungs-Werkzeugs auch über eine API aufrufen.

Die meisten Werkzeuge sind bereits frei zugänglich in Github-Repositories abgelegt und mit permissiven Lizenzen ausgestattet. Einige Repos benötigen vor der Freigabe noch weitere Prüfungen und Umstrukturierungen.

Die Benutzeroberflächen sind weitestgehend ausschließlich in deutscher Sprache realisiert. Mehrsprachigkeit ist nachrüstbar, sofern sich erweisen sollte, dass es dafür Bedarf gibt. Aus dem Projektplan ADL ergibt sich diese Anforderung jedoch nicht. Auch die begleitenden Dokumentationen sind in deutscher Sprache verfasst. Allgemeiner Konvention folgend sind der Code und die unmittelbaren Code-Kommentare jedoch zumeist in englischer Sprache gehalten.

Wörterbuchkonsistenzprüfung



ÜBERBLICK

Das Wörterbuchkonsistenzprüfungs-Tool ist das umfangreichste und komplexeste Werkzeug der bisherigen Entwicklung. Es verarbeitet beliebige XML-Daten, unabhängig vom spezifischen XML-Format oder der Sprache. Für den Standard TEI Lex-0 werden erweiterte Prüffunktionen angeboten.

Es können einzelne oder mehrere Dateien, komplette Ordner oder ZIP-Archive hochgeladen werden. Alternativ kann der in einer angebundenen Datenbank vorliegende Bestand genutzt werden. Ein Nutzer- und Rechtemanagement erlaubt es zudem, die Bestände nutzerbezogen freizugeben – so können auch noch unveröffentlichte Daten geprüft werden.

Die Prüffunktionen teilen sich grob in zwei Gruppen auf: Technische Kriterien (wie Wohlgeformtheit, Schemavalidität, Vorkommen bestimmter Inhalte, Tags und Attribute, Suche, Nummerierung, Verweise) und inhaltliche Kriterien, wobei diese deutlich schwieriger zu prüfen sind. In diesem Bereich laufen auch Tests zum Einsatz künstlicher Intelligenz, aber noch sind die Ergebnisse nicht hochwertig und konsistent genug.

MERKMALE

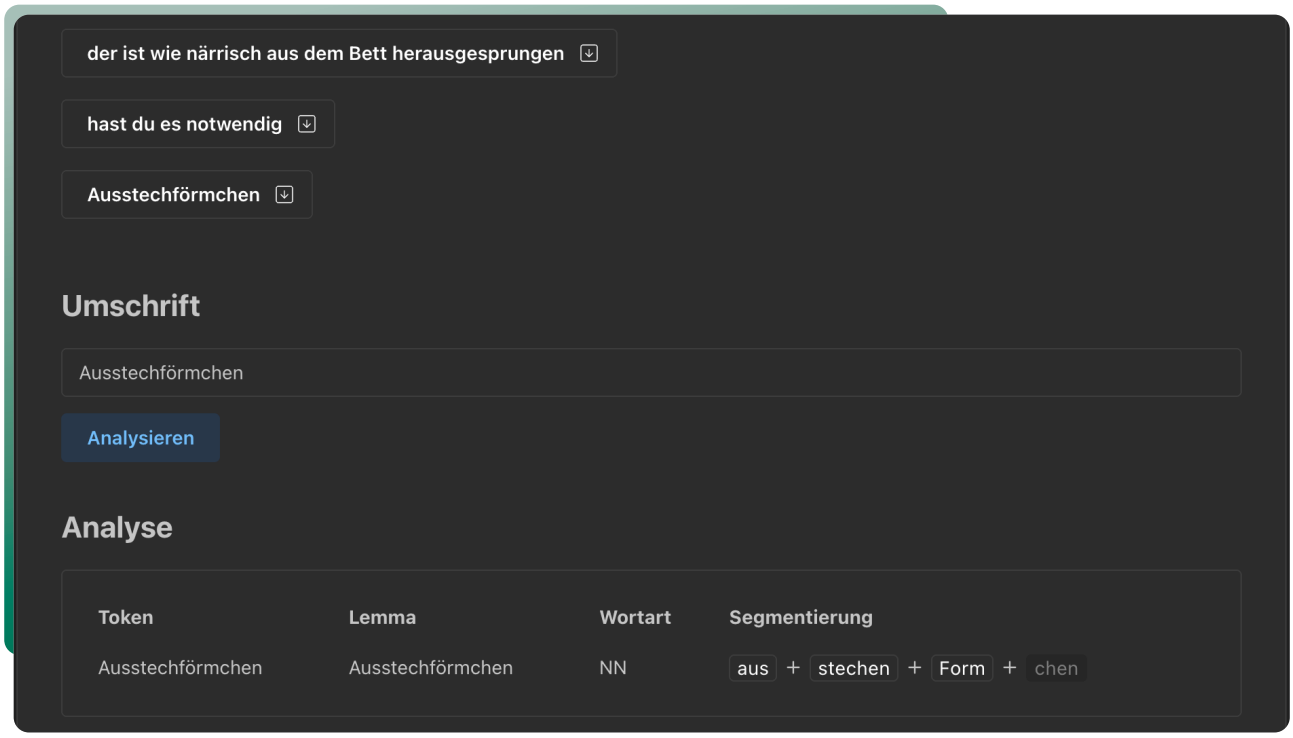
- » Artikelsuche
- » XML und TEI-Lex-0-Validator
- » Strukturanalyse aller ausgewählten Artikel
- » Tag- und Pfadsuche
- » Einmaligkeit
- » Verschachtelung
- » Inhalte und leere Tags
- » Bedeutungsnummerierung
- » Ungültige Verweise
- » Anzahl und Länge (von Bedeutungen)
- » Alte Rechtscheibung
- » In der Erprobung: LLM-Schnittstelle

- » Komplette API-Schnittstelle
- » Vielfältige Upload-Möglichkeiten
- » Datenbank-Anbindung
- » Nutzer- und Rechtekonzept

URL

https://dictconsistency.tools.lexoterm.de

Morphologische Segmentierung



ÜBERBLICK

Die Aufbereitung von dialektologischen Fragebögen im Fränkischen Wörterbuch beinhaltet neben der händischen Erfassung und Digitalisierung von eingescannten Bögen auch die Aufbereitung der dort enthaltenen Wortbedeutungen. Diese kurzen Definitionen werden dafür in einer Excel-Tabelle u.a. in Einzelwörter zerlegt, die anschließend gefiltert, lemmatisiert, mit Wortarten versehen und morphologisch analysiert werden. Da viele dieser Schritte mit hohem Genauigkeitsgrad automatisiert werden können, wurde in einem Demo-System konkret getestet, inwiefern sich einer der Schritte – die Analyse von Nominal-Komposita – mittels eines regelbasierten Systems (*SMOR: Stuttgarter Morphologisches Analysewerkzeug*) automatisieren lässt. Dafür wurde das Java-basierte Tool als eigener Docker-Service umgesetzt und in Kombination mit einer Python-basierten NLP-Pipeline als morphologische Analyse-API zur Verfügung gestellt. Im weiteren Verlauf soll die Pipeline erweitert werden, um vollständige Datensätze zu verarbeiten, wofür noch einige Vorarbeiten zur Optimierung der Systemparameter und der Benutzeroberfläche nötig sind.

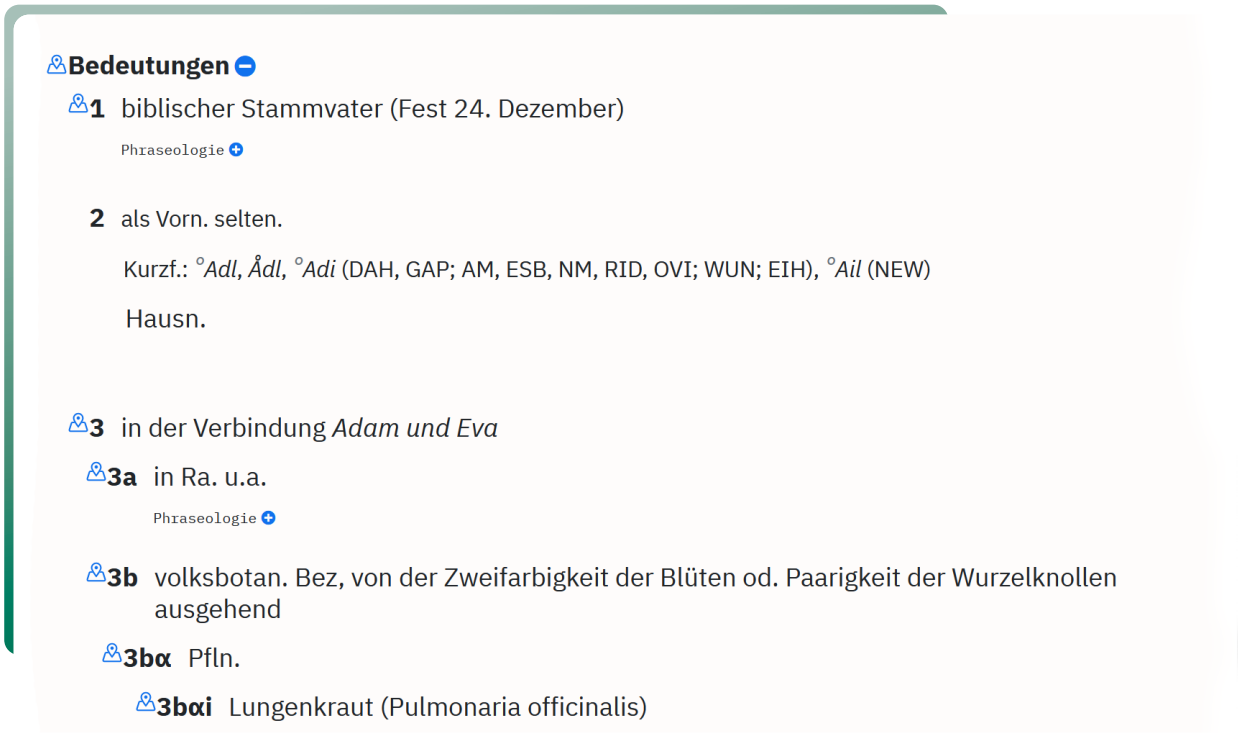
MERKMALE

- Testimplementierung für morphologische Segmentierung
- Basierend auf SMOR: Stuttgarter Morphologisches Analysewerkzeug
- API für Einbindung der Analyse in redaktionelle Abläufe/Tools

URL

ggw. nur intern verfügbar

Bedeutungsnummerierung



ÜBERBLICK

Bei der Arbeit mit den Daten des Bayerischen Wörterbuchs sind einige Inkonsistenzen in der Nummerierung von Wortbedeutungen und -unterbedeutungen zu Tage getreten. Um diese systematisch zu erfassen, wurde ein Prüfmechanismus umgesetzt, der die gesamten BDO-Daten einliest und für jedes Lemma prüft, ob die Bedeutungsnummerierungen in richtiger Reihenfolge und Einrückungstiefe vorliegen. Da das korrekte Format bei der Eingabe derzeit noch nicht forciert wird, wurden dadurch Fehler und Inkosistenzen im dreistelligen Bereich identifiziert.

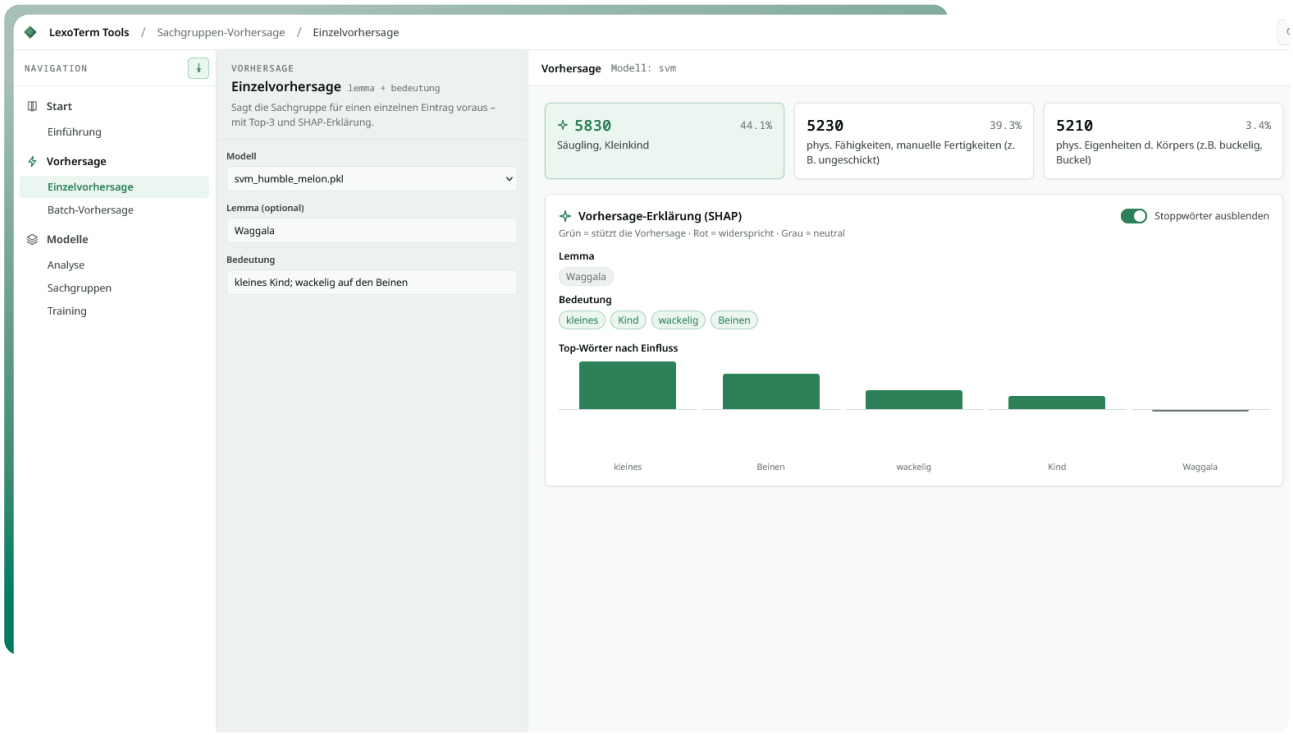
MERKMALE

- Prüfung auf Nummerierungskonsistenz
- Entwickelt für BDO-Wörterbücher
- Grundlage der Nummerierungsprüfung im WBKP-Tool

URL

als Skript verfügbar

Sachgruppenvorhersage



ÜBERBLICK

Das Sachgruppenvorhersage-Tool entstand aus der Idee heraus, die Vorhersage-Genauigkeit klassischer ML-Algorithmen zu überprüfen, z.B. Support Vector Machines, Random Forest, Extreme Gradient Boost oder einfache neuronale Netze.

Der entscheidende Vorteil dieser Verfahren: Sie sind schnell zu trainieren, liefern blitzschnelle Inferenz, benötigen keine teure Hardware oder Cloud, verursachen keine Abo-Kosten, funktionieren deterministisch, und sind theoretisch, konzeptionell und in ihren Ergebnissen gut verstanden.

Die entstandenen Modelle, die testhalber auf den WBF- und DIBS-Daten trainiert wurden, erreichen Vorhersagegenauigkeiten zwischen rund 50% und 85%, je nach Parameterwahl und Rahmenbedingungen. Eine detaillierte Auswertung zeigt zudem zahlreiche Einschränkungen und Caveats auf.

In einer vergleichenden Analyse wurde dieses Verfahren auch einem LLM-basierten Ansatz sowie einem BERT-Embedding gegenübergestellt. Details stehen auf Anfrage zur Verfügung.

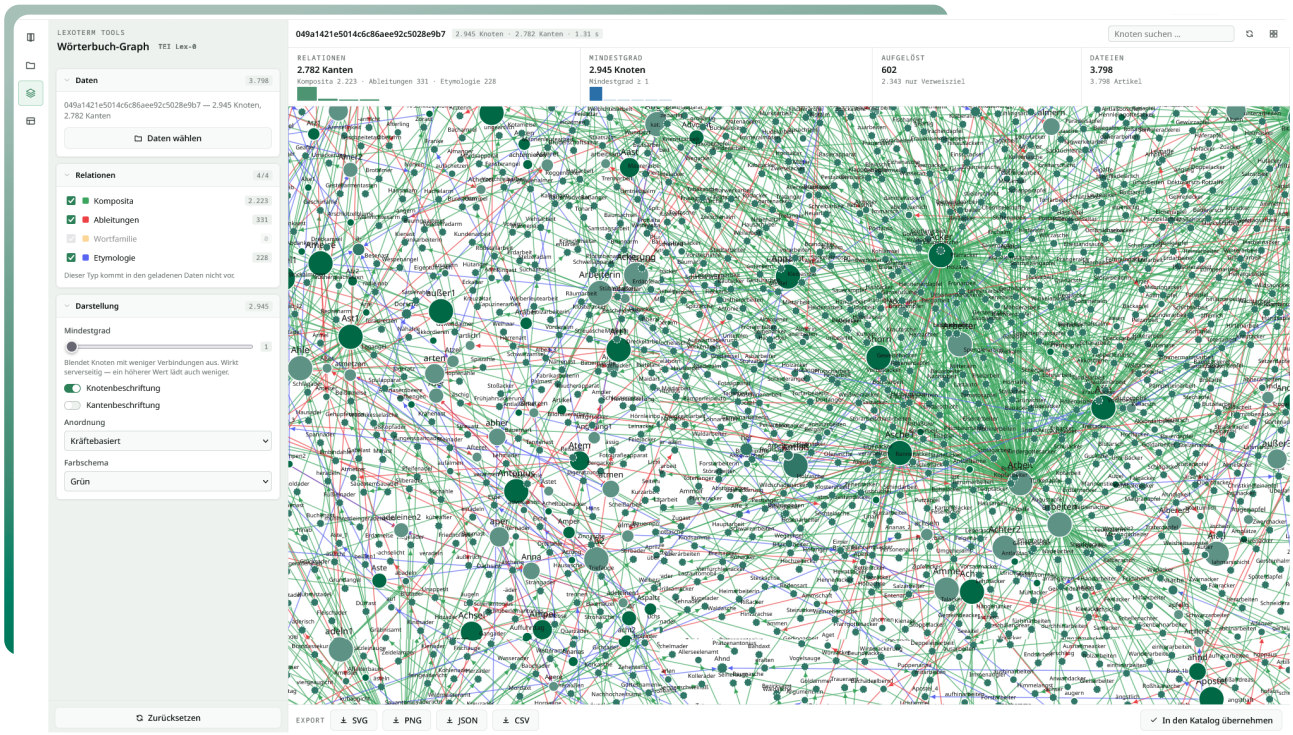
MERKMALE

- frei wählbare Input-Daten, nur String und Klasse (Sachgruppe) erforderlich
- flexibel für beliebige Wörterbücher nutzbar
- komfortables Interface für Trainings mit logistischer Regression, linearen Support Vector Machines, Random Forest, XGBoost sowie neuronalen Netzen mit variabler Layer und Neuronenzahl
- Hyperparameter-Tuning via GUI
- Semantische Anreicherung (spaCy, Dornseiff)
- Cross- und Batch-Trainings
- Ergebnisanalyse mit detaillierten Reports
- Klassenbezogene Auswertungen
- Vorhersage inkl. Konfidenzen
- Batch-Vorhersage
- SHAP-Analyse
- Komplette API-Schnittstelle

URL

https://sgpredict.tools.lexoterm.de

Wörterbuchgraph



ÜBERBLICK

Der Wörterbuch-Graph-Generator gehört für den Moment eher in die Kategorie der spielerischen Anwendungen, wobei es viel Freude bereitet, sich durch den Graphen und seine Vernetzungen zu bewegen.

Die technische Grundlage sind XML-Dateien im TEI Lex-0-Format, da dieses die fest definierten Relationstypen "etymology", "compound", "derivative" und "family" kennt. Sind diese in den Daten vorhanden, können sie dargestellt werden.

Wörterbuchartikel bzw. Lemmata werden so zu Knoten und Verweise zu gerichteten Kanten. Die entstandenen Nodes-Listen können im JSON-Format exportiert und importiert werden.

Für die Verwendung der Visualisierungen ausserhalb des Tools stehen zudem Export-Funktionen als SVG oder PNG bereit. Auch können verschiedene Farbmodi gewählt werden, darunter auch Schwarz-Weiss für einen effizienten Druck.

Das Graph-Tool soll die Grundlage für ein noch folgendes Tool zur graphbasierten-Darstellung komplexer Etymologien liefern.

MERKMALE

» Verarbeitung beliebiger XML-Wörterbuchdaten im TEI Lex-0-Format

» Upload von einzelnen oder mehreren Dateien, Ordnern, ZIP-Files, auch Graph-JSONs

» Vorberechnete Graphen

» Datenbankanbindung

» Kräftebasierte und radiale Anordnung

» Mindestgrade

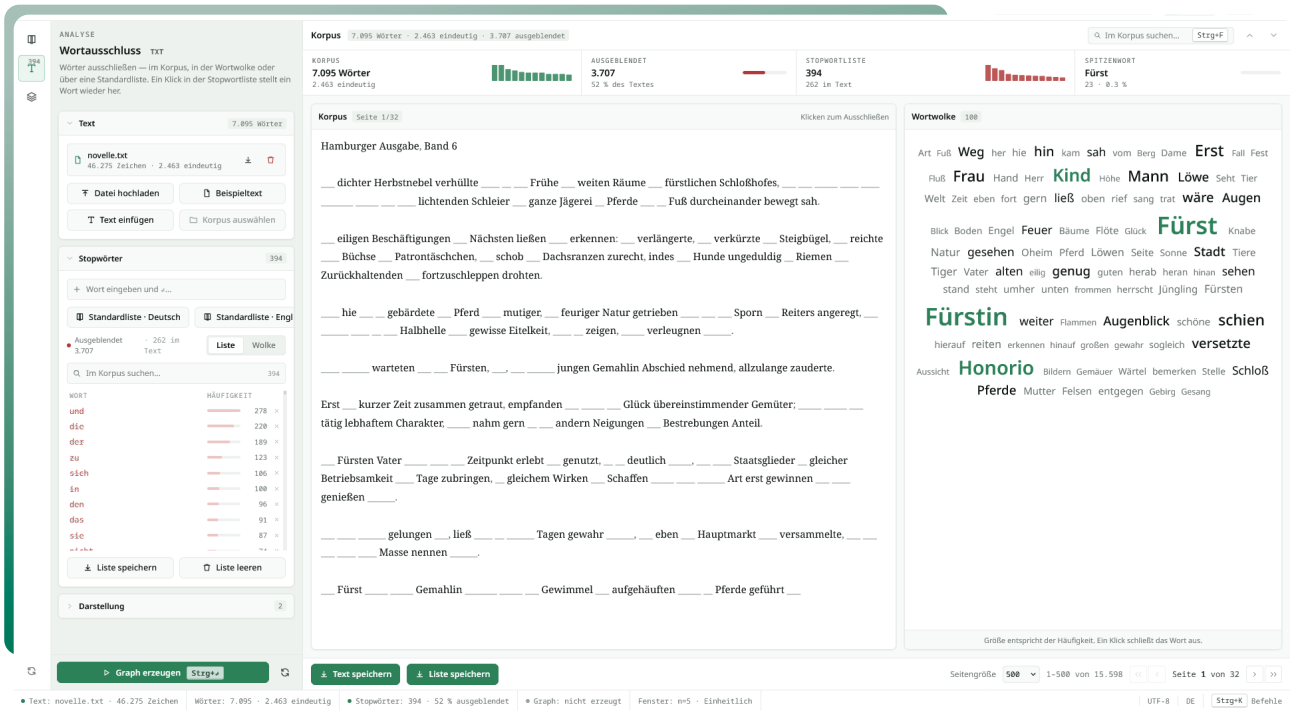
» Export als SVG, PNG, CSV und JSON

» Verschiedene Farbmodi

URL

https://dictgraph.tools.lexoterm.de

Visueller Wortauschluss



ÜBERBLICK

Die Entfernung sogenannter Stopwords, also kaum Sinn tragender Wörter wie etwa Artikel, Pronomen, Konjunktionen oder Hilfsverben, gehört zum Standardvorgehen in vielen NLP-Projekten. Üblicherweise wird dazu auf vorgefertigte Listen zurückgegriffen, die kaum je modifiziert werden.

Möchte man jedoch individuelle Listen von auszuschließenden Wörtern erstellen, bleibt meist nur das mühsame Durcharbeiten langer Listen.

Hierfür bietet das Tool "Visueller Wortauschluss" eine Lösung: In verschiedenen Modi können Wörter durch einfaches Klicken aus- bzw. eingeschlossen werden: Volltext, frequenzbasierte Wortwolken, oder Graphen.

Dieser Ansatz bietet sich gerade auch im Dialektbereich an, weil für Dialekte in den üblichen Programmbibliotheken keine vorgefertigten Listen angeboten werden.

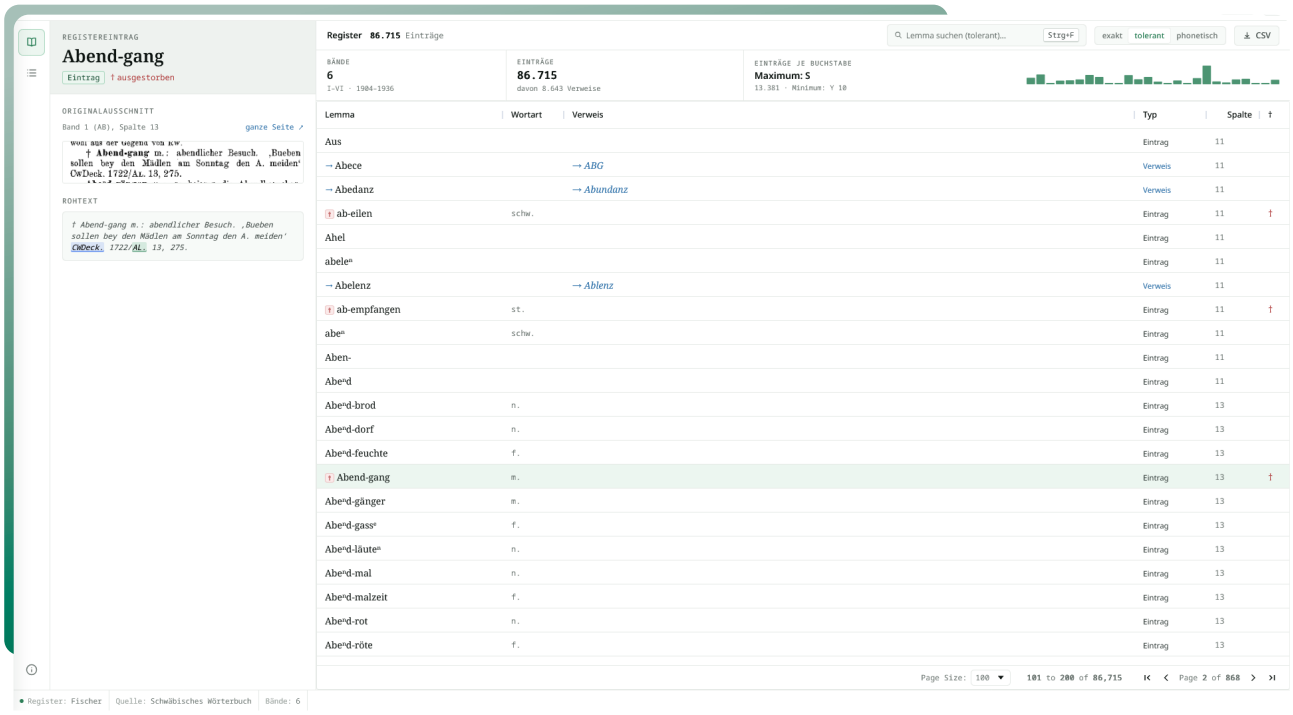
MERKMALE

- Verarbeitung beliebiger Rohtexte
- Ein- und Ausschluss durch einfaches Klicken in verschiedenen Ansichtsmodi
- Laden vorgfertigter Listen
- Speichern erzeugter Listen
- Graph-Darstellung
- Geplante Weiterentwicklung: Integration von POS- und NER-Analyse zur gezielten Filterung

URL

https://stopgraph.tools.lexoterm.de

Fischer-Register



ÜBERBLICK

Das sogenannte Schmeller-Register zu Johann Andreas Schmellers Bayerischem Wörterbuch ist eine der beliebtesten Ressourcen unter den Online-Publikationen der BAdW.

Vor diesem Hintergrund entstand in der Redaktion des DIBS der Wunsch, ein in gleicher Weise ausgeführtes Fischer-Register für das historische "Schwäbische Wörterbuch" zu schaffen.

Während das Schmeller-Register in jahrzehntelanger Handarbeit erstellt wurde, war dies für das Fischer-Register keine realistische Option. Auch die vorhandenen OCR-Schichten der digitalen Ausgabe waren für diesen Zweck zu fehlerbehaftet.

Daher wurde für dieses Projekt ein neues OCR-Modell nachtrainiert, auf Basis eines Fraktur-Kraken-Modells und des eScriptorium-Frameworks. Mit diesem individuellen Modell wurde der gesamte "Fischer"-Volltext neu erfasst. Ein aufwändiges Parsing der Seiten und Einträge erschließt die einzelnen Einträge, aus denen das Register erzeugt wird. Komfortfunktionen wie ein integriertes Abkürzungs- und Siglenverzeichnis stehen ebenfalls zur Verfügung.

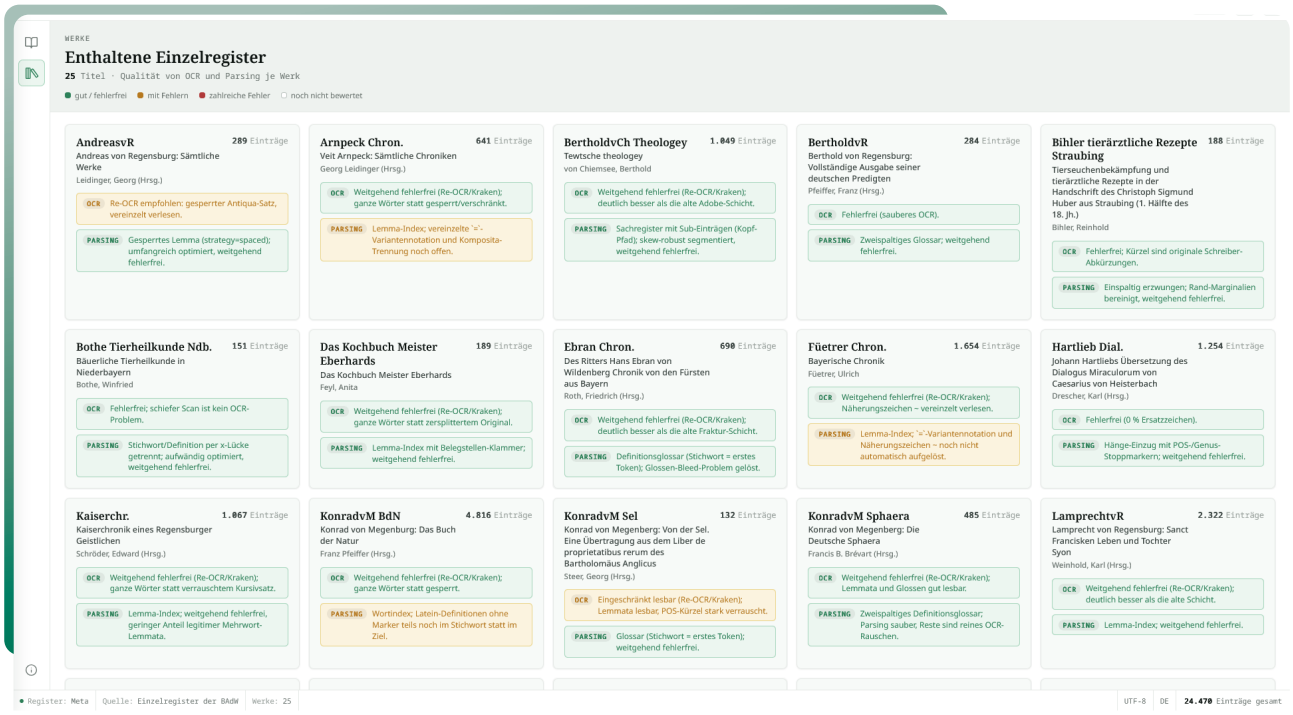
MERKMALE

- Kompletter Inhalt des "Schwäbischen Wörterbuchs" inkl. Artikeltexten, insgesamt 86.715 Einträge
- Anzeige des Original-Ausschnitts bzw. der Original-Scans
- Neue Textschicht, basierend auf einem eigenes OCR-Modell, getuned auf handtranskribierten Seiten, Fehlerrate (CER) ~ 2%.
- Lemma, Wortart und Volltext durchsuchbar
- exakte, tolerante und phonetische Suche
- Abkürzungs- und Siglen-Verzeichnis mit Inline-Auszeichnung
- Korrekturmodus zur Nachbearbeitung
- Geplant: Veröffentlichung auf dem Publikationsserver der BAdW

URL

https://fischer.register.lexoterm.de

BWB-Meta-Register



ÜBERBLICK

Inspiriert durch das Fischer-Register wurde für die Redaktion des BWB ein Meta-Register umgesetzt: Es vereint die Einträge von 25 Einzelregistern, die nun nicht mehr separat durchsucht werden müssen, sondern in einem Durchgang erschlossen werden können.

Die besondere Herausforderung bestand in diesem Projekt darin, dass jedes der 25 Register eine andere Konvention nutzt, um die Registereinträge zu kennzeichnen bzw. im Layout darzustellen: einspaltig, zweispaltig, dreispaltig; eingerückt, gesperrt, kursiv, unterstrichen, fettgedruckt; mit Wortartangaben und ohne, mit wiederkehrenden Trennzeichen und ohne, und so weiter.

Es wurden verschiedende Erkennungsstrategien entwickelt, die sich je nach Druckwerk kombinieren lassen. Für ausgewählte Verzeichnisse werden auch hierarchische Gliederungen unterstützt, um verschachtelte Listen abzubilden.

Aufgrund der gleichen technischen Basis bietet das Metaregister dieselben Komfortfunktionen wie das Fischer-Register.

MERKMALE

» Enthält 25 verschiedene Einzelregister aus historischen Werken

» Anzeige des Original-Ausschnitts bzw. der Original-Scans

» Weitgehende Neuerstellung der Textschichten auf Basis der existierenden eScriptorium-Pipeline

» Lemma, Werk und Volltext durchsuchbar

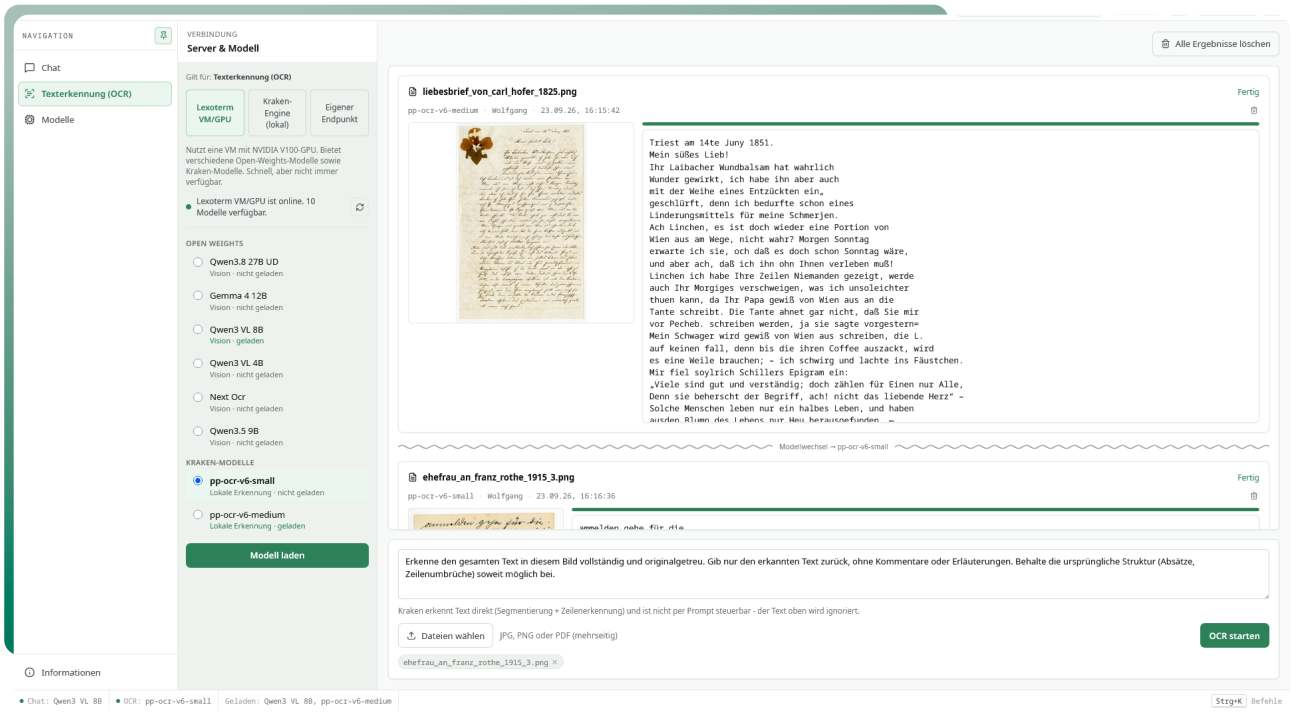
» exakte, tolerante und phonetische Suche

» Korrekturmodus zur Nachbearbeitung

URL

https://bwbmeta.register.lexoterm.de

Modellspielplatz



ÜBERBLICK

Gegenwärtig erscheinen täglich zahlreiche neue KI-Modelle zur Generierung von Text, zur Texterkennung (OCR), zur Bildgenerierung, etc. Erfahrene Nutzer können diese Modelle mit dafür geeigneten Programmen laden und ausführen, je nach Art des Modell mehr oder weniger komfortabel.

Das Werkzeug "Modellspielplatz" bietet für technisch weniger versierte Nutzer die Möglichkeit, viele dieser Modelle ebenfalls zu nutzen, und zwar in einer vertrauten Benutzeroberfläche.

Für einen konkreten Anwendungsfall wurden HTR-Modelle zur Handschrifterkennung von Fragebögen in Sütterlin getestet. Zwar war die Erkennungsleistung in diesem speziellen Fall aufgrund verschiedener Faktoren unbefriedigend, aber das Werkzeug funktioniert unabhängig davon für viele verschiedene Anwendungsfälle.

Neben offenen und lokalen Modellen können über eine Endpunkt-Schnittstelle auch kommerzielle Cloud-Anbieter innerhalb des Tools genutzt werden.

MERKMALE

» Komfortable Testumgebung für Modell-Inferenz-Leistung

» Kompatibel mit Modellen mit OpenAI-Schnittstelle, d.h. nahezu allen Modellen zur Textgenerierung und Bilderkennung, die in Ollama, llama.ccp oder LM Studio genutzt werden können.

» Kompatibel mit Modellen der Kraken-Familie (weit verbreitet für historische Texte)

» Anbindung an eine dezidierte VM mit NVIDIA V100-GPU vorhanden/möglich

» Lokale Ausführung möglich, je nach Hardware

» Integration weiterer Endpunkte (z.B. OpenAI oder Anthropic-APIs)

» Textgenerierung (Chat) und Texterkennung (OCR/HTR)

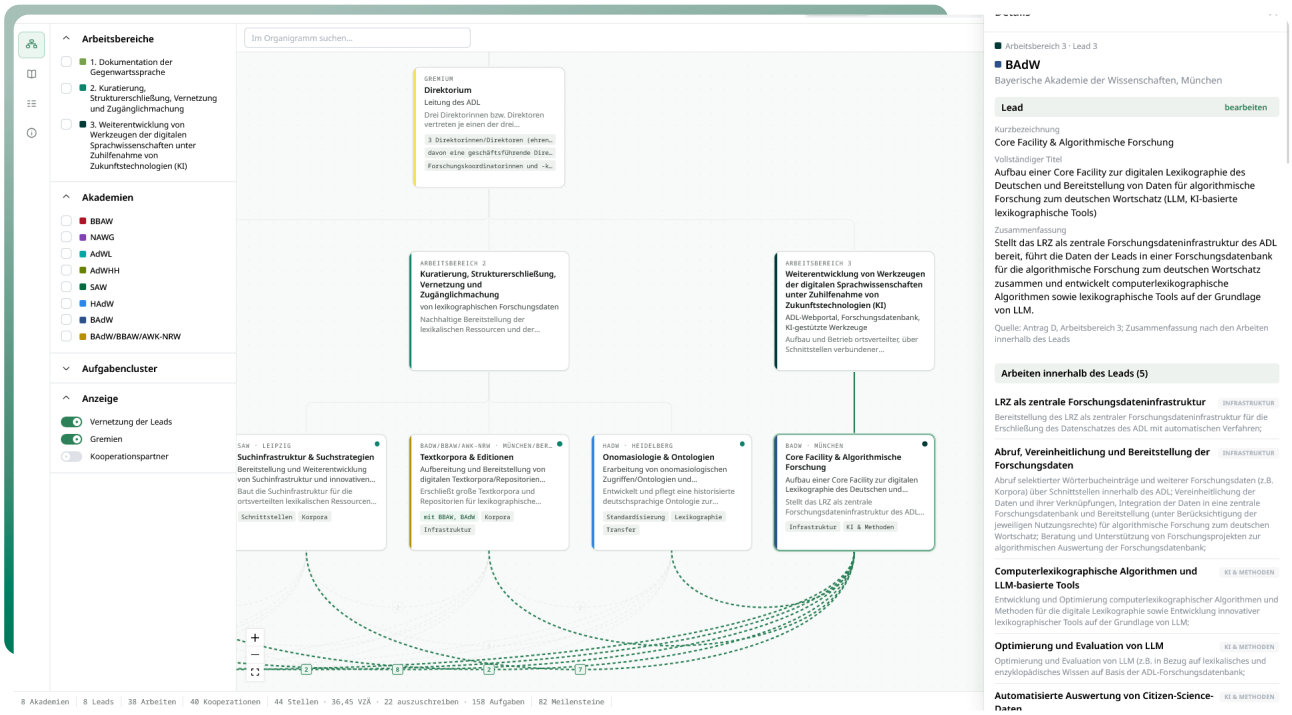
» Modellwechsel in kontinuierlicher Ergebnisanzeige möglich

» Geplant: Möglichkeit zur Reservierung der GPU bzw. des geladenen Modells für einen wählbaren Zeitraum

URL

https://modelplayground.tools.lexoterm.de

Interaktives ADL-Organigramm



ÜBERBLICK

Der Antrag bzw. Projektplan des zu gründenen "Akademienzentrums Digitale Lexikographie des Deutschen" ist ein umfangreiches, komplexes Dokument. Nicht nur werden die Aufgaben der beteiligten Akademien detailliert beschrieben und in Meilensteine gegliedert, es werden auch zahlreiche Kooperationen und Bezüge zwischen den Projekten ausgeführt.

Das interaktive ADL-Organigramm übersetzt den Antrag in verschiedene Darstellungsformen. Im Mittelpunkt steht ein Organigramm, das weiter erforscht werden kann: Zu jedem Eintrag stehen umfangreiche Details zu Verfügung, immer basierend auf dem Originaltext. Es werden die wechselseitigen Beziehungen innerhalb der Leads dargestellt sowie Meilensteine und Stellen gelistet.

Darüber hinaus kann die Zusammenarbeit auch als Netzwerk, Matrix oder über die Aufgaben innerhalb der Stellen erschlossen werden.

Zusätzlich ist ein Verzeichnis aller behandelten Wörterbüchern und der umfangreichen Abkürzungen integriert.

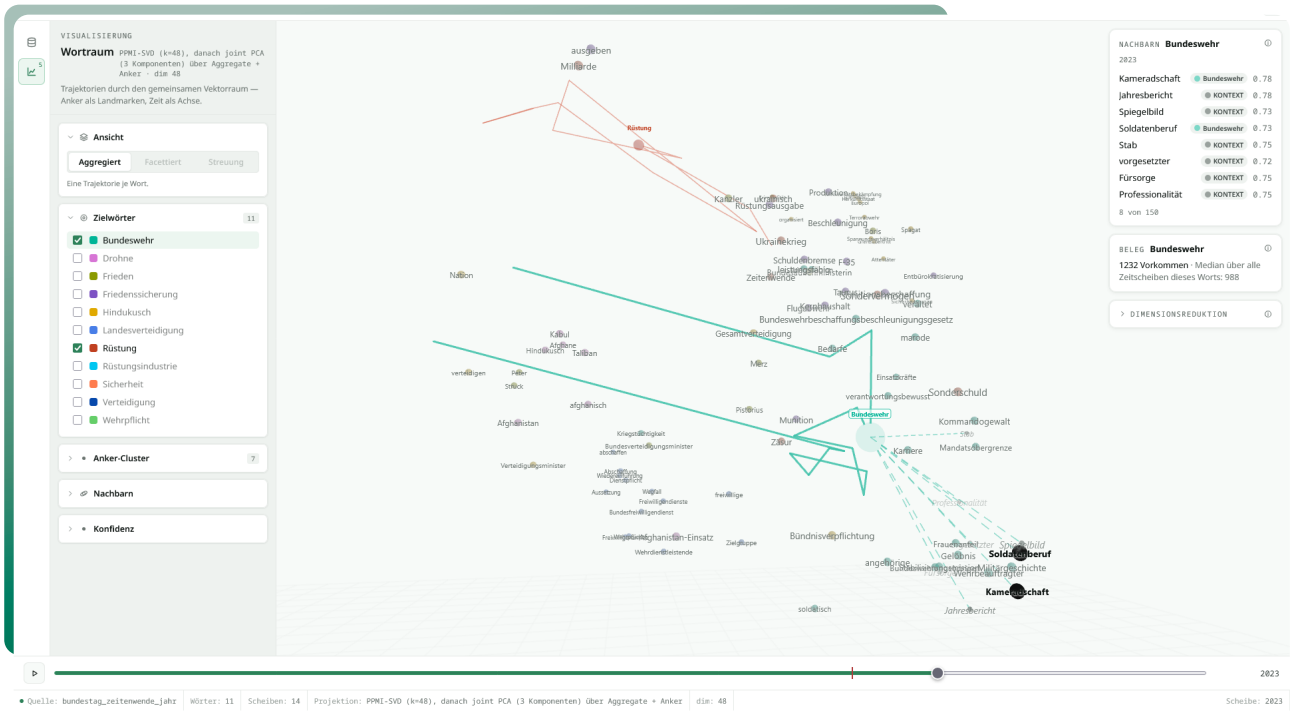
MERKMALE

- Darstellung der Arbeiten innerhalb des Projekts ADL als Organigramm
- Alternativ Darstellung als Netzwerk, Matrix oder stellenbasiert
- Details zu jedem Partner bzw. Lead
- Auflistung der wechselseitigen Nennungen in Kooperationen
- Zugang ggw. nur mit Zugangsdaten möglich

URL

https://adlorg.tools.lexoterm.de

Visualisierung Semantischer Vektoren



ÜBERBLICK

Für das DNB-Forschungsprojekt "Zeitenwende im Wortfeld" wurde früh auch die Bedeutung einer gelungenen Visualisierungskomponente erkannt.

Angelehnt an Googles *Tensorflow Projector* übernimmt das aufwändige Tool "Visualisierung Semantischer Vektoren" (VSV) die Darstellung von Wort-Koordinaten aus hochdimensionalen Embeddings im 3D-Raum.

Ein definiertes, inhaltsneutrales Datenformat erlaubt die Darstellung beliebiger Embeddings. Auch die Methode der Generierung der Embeddings kann frei gewählt werden.

Mitgeliefert wird eine Beispiel-Pipeline zur Datenverarbeitung, die Bundestagreden (als CSV-Volltexte) schrittweise mittels PPMI-SVD bzw. SGNS umwandelt und die benötigten Koordinaten und Informationen zu nächsten Nachbarn bereitstellt.

Ergänzt werden die Darstellung der Zeit-Trajektorien pro Wort durch Ähnlichkeitsmatrizen, Drift-Kurven, Nachbar-Analysen, und Wort-zu-Wort-Vergleiche.

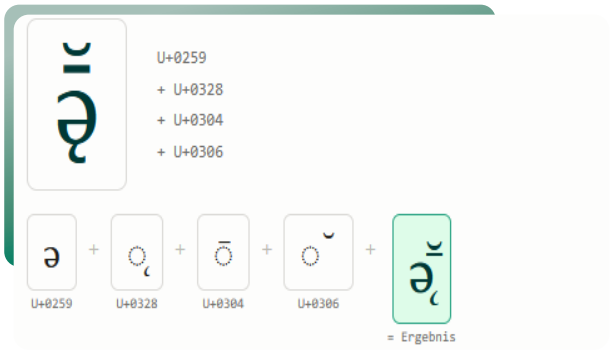
MERKMALE

- Visualisiert Wörter im 3D-Raum, basierend auf hochdimensionalen Embeddings
- Zeigt semantische Verschiebungen über die Zeit in Form von Trajektorien an
- Kuratierbare Ankerwörter und Cluster liefern feste Bezugspunkte
- Nächste Nachbarn werden im hochdimensionalen Raum berechnet und pro Zeitscheibe angezeigt (Anzahl wählbar)
- Granularität der Zeitscheiben frei wählbar
- Ähnlichkeitsmatrizen (nach Kiyama)
- Driftkurven (nach Hamilton)
- Wort-zu-Wort-Vergleiche (cos)
- Veränderungen der nächsten Nachbarn
- Interpretationshilfen in der App

URL

https://vsv.tools.lexoterm.de

Hilfsprogramme

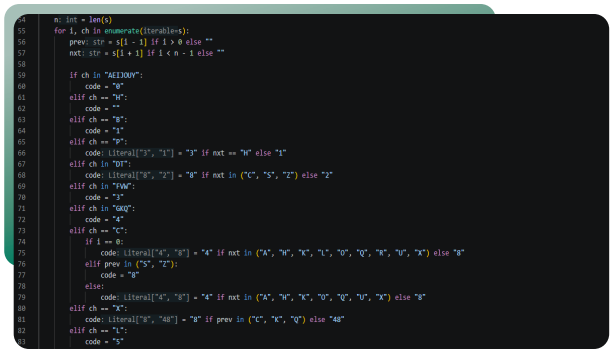


UNICODE-HOCHSTAPLER

Zur Unterstützung der manuellen Transkription im Fischer-Register wurde ein kleines Tool realisiert, mit dem sich mehrfache Diakritika und Sonderbuchstaben zu gültigem Unicode kombinieren lassen. Im "Fischer" werden diese Zeichen zur Markierung der Aussprache breit verwendet.

URL

https://unicodestacker.tools.lexoterm.de



REGISTERSUCHE

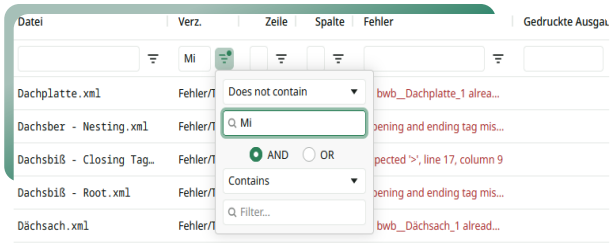
Für die Register wurde eine einheitliche Suchkomponente entwickelt, die drei Modi bietet:

» exakte Anfangssuche

» fehlertolerante Suche (Prefix-Levenshtein)

» Kölner Phonetik

Die Suche mittels Kölner Phonetik erlaubt eine Suche nach Ausspracheähnlichkeit und eignet sich besonders für historische Texte. Beispielsweise findet *minne* auch *mynne* und *Mayr* findet *Meier*.



AGGRID-WRAPPER FÜR REFLEX

Frühere Tool-Versionen nutzten das Reflex-Framework, das eine Programmierung auch des Web-Frontend rein in Python erlaubt, dieses aber in ein React-Frontend wandelt.

AG Grid ist das leistungsstärkste Tabellentool auf dem Markt - steht aber standardmäßig nicht für Reflex zur Verfügung. Der Wrapper liefert die JS-Version auch für das Python-Framework aus.

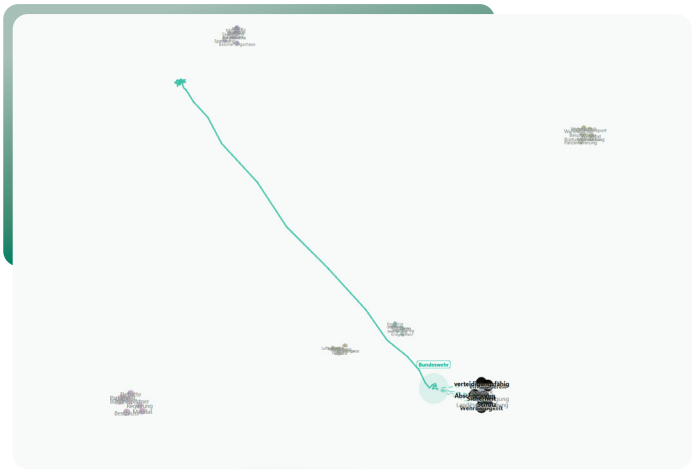
Forschung

Einzelne Projekte des Pilotprojekts eignen sich auch zur wissenschaftlichen Vertiefung, mit der Möglichkeit zur Konferenzeinreichung oder Veröffentlichung eines Aufsatzes. Im Nachfolgenden sind diese kurz beschrieben.

DNB-Forschungsprojekt "Zeitenwende im Wortfeld"

Die Deutsche Nationalbibliothek lädt seit geraumer Zeit Forschende aus dem Bereich Digital Humanities dazu ein, sich mittels vorwiegend neuartiger korpuslinguistischer Ansätze mit den Beständen der DNB auseinanderzusetzen und diese zu erschließen. Ein besonderer (An-)Reiz besteht darin, dass diese Bestände normalerweise nicht ohne Weiteres zugänglich oder analysierbar sind, vorwiegend aus urheberrechtlichen Gründen. Da im Projekt ADL vielfach korpusanalytische Aufgaben skizziert sind, ergab sich somit eine günstige Gelegenheit.

Die zentrale Frage des Forschungprojektes lautet: Lässt sich ein semantischer Wandel von Begriffen aus dem Wortfeld *Militär* im Zeitraum 2014 - 2025 nachweisen? Das Jahr 2014 markiert mit der Besetzung der Krim-Halbinsel durch Russland den Beginn einer Periode, die ihren Höhepunkt in der Invasion der Ukraine und der darauf folgenden Zeitenwende-Rede des damaligen Bundeskanzlers Olaf Scholz im Jahr 2022 findet. Eine zu beweisende Vermutung wäre, dass sich Begriffe wie *Bundeswehr, Wehrpflicht* oder *Verteidigung* zuvor in einem eher negativ konnotierten Umfeld (*Skandal, Abschaffung, Kostenexplosion*) befinden, sich dann aber in eine eher positiv wahrgenommende Richtung bewegen (*Sicherheit, Wehrhaftigkeit, Landesverteidigung*).



Visualisierung der Bedeutungsverschiebung über die Zeit des Wortes "Bundeswehr" auf Basis synthetischer Daten

Diese vermutete Bedeutungsverschiebung soll anhand von Embeddings in Form von Vektordistanzen bzw. über die Zeit betrachtet, semantischen Vektortrajektorien gemessen werden. Für das Embedding selbst kommen zwei aktuelle Verfahren zum Einsatz: PPMI-SVD sowie SGNS mit anschließendem Alignment der unabhängigen, hochdimensionalen Vektorräume. Beide Verfahren weisen Stärken und Schwächen auf, deren genaue Auswirkung noch analysiert werden muss.

Grundlage der Analyse ist ein Korpus bestehend aus 23 sorgfältig ausgewählten Tageszeitungen aus 12 Jahren. Diese werden in unterschiedlich granulare Zeitscheiben geclustert und facettiert nach Publikation ausgewertet.

Eine Besonderheit des Projektes stellt die Entwicklung einer Visualisierungsumgebung dar, die vielfältige visuelle Analysehilfen bietet. Neben der räumlichen Darstellung werden beispielsweise auch Selbstähnlichkeitsmatrizen und Drift-Kurven integriert (siehe auch: Werkzeug "Visualisierung semantischer Vektoren")

Die Datenanalyse, die vor Ort am Standort der DNB erfolgen muss, ist für Ende November terminiert, die vorbereitenden Arbeiten sind im vollen Gange. Der Projektabschluss wird im Frühjahr 2027 erwartet.

Kooperationsvereinbarung Staatsbibliothek

Im erweiterten Kontext der Bewilligung des Akademienzentrums Digitale Lexikographie wurden durch die Leitung der BAdW die Kontakte zur Bayerischen Staatsbibliothek revitalisiert. Anknüpfend an frühere Kooperationen wurde vereinbart, thematisch passende Ausschreibungen des BMFTR im Rahmenprogramm "Orientierung für eine Welt im Wandel" für die Geistes- und Sozialwissenschaften auf Potential für ein gemeinschaftliches Projekt hin zu prüfen. Hierfür könnten ggf. korpusanalytische Methoden aus dem Projekt "Zeitenwende im Wortfeld" zum Einsatz kommen. Auch auf fachlicher Ebene, beispielsweise im Bereich des korpusbasierten Modelltrainings scheint ein engerer Austausch reich an Potential.

Organisatorisches

Stellenbesetzungen

Der ADL-Förderantrag enthält bereits präzise ausgearbeitete Stellen- und Aufgabenbeschreibungen. Mit der Erfahrung aus einem Jahr praktischer Arbeit im Pilotprojekt wurden diese minimal überarbeitet: Der Aufgabenbereich Front-End-Entwicklung bekommt etwa eine stärkere Betonung. Mit Blick auf den Beginn des ADL im ersten Quartal 2027 wird eine zeitnahe Ausschreibung angestrebt.

Zusammenarbeit BDO

Die Zusammenarbeit mit den Wörterbuchredaktionen BWB, WBF und DIBS gestaltet sich weiterhin sehr angenehm und produktiv. Für das WBF-Team konnte das Tool zur morphologischen Segmentierung entwickelt werden, für das DIBS das Fischer-Register, und für das BWB das Meta-Register. Alle Redaktionen können die Tools zur Wörterbuchkonsistenzprüfung oder zum Testen neuer KI-Modelle nutzen.

Ebenso gestaltet sich der Austausch mit den Kolleg*innen aus der Abteilung Digital Humanities produktiv. Für das kommende Jahr wird eine Zusammenarbeit im Bereich des Modelltrainings angestrebt.

Projektwebseite / Kommunikation

Die Projektwebseite (pdl.badw.de) informiert auf aktuellem Stand über Projektergebnisse und nutzbare Tools.

Für den schnellen Überblick stehen weiterhin zwei Flyer über das Projekt zur Verfügung. Diese beschreiben die Projektziele und -inhalte jeweils in kurzer Form (Mini) bzw. ausführlicher (Midi). Beide Flyer sind über die Projektwebseite abrufbar.

Auf darüber hinausgehende Kommunikation, etwa im Social-Media-Bereich wurde bisher verzichtet, nicht zuletzt weil die Bewilligung des Akademienzentrums abzuwarten war.



Projektflyer

Ausblick

Anlauf ADL & Abstimmung Projektpartner

Für Januar 2027 plant die BBAW mit einer Auftaktveranstaltung zur Gründung des Akademienzentrums Digitale Lexikographie des Deutschen. Nach diesem Startschuss dürften sich die Abstimmungen zwischen den Projektpartnern intensivieren: Das gesamte Projekt fusst auf vielfältigen Kooperationen zwischen den Akademien in ganz unterschiedlichen Arbeitsbereichen. Der BAdW kommt dabei eine zentrale Rolle zu: Mit Unterstützung des LRZ soll sie einen wichtigen Teil der technischen Infrastuktur bereitstellen, insbesondere für die geplante Forschungsdatenbank und die KI-Nutzung.

Bereits im Pilotprojekt konnte der fachliche Austausch mit mehreren Akademien eingeleitet werden:

» Die Berlin-Brandenburgische Akademie der Wissenschaften stellt mit dem DWDS einen zentralen Bezugspunkt dar, beispielsweise für die Artikelverlinkung – hier ist auch eine formlose Kooperation im Bereich Linked Open Data anvisiert. Darüber hinaus besteht ein regelmäßiger Austausch zu verschiedenen Fragestellungen rund um das Thema ADL. Schließlich wurden durch das DWDS-Team Testdatensätze zur Integration in LexoTerm bereitgestellt.

» Mit der Sächsischen Akademie der Wissenschaften konnte bereits in zwei Bereichen erfolgreich kooperiert werden: Für die BDO-Daten wurde eine LexFCS-kompatible Schnittstelle geschaffen, die die Daten für die föderierte Suche im Rahmen von Text+ zur Verfügung steht. Zudem hat auch die SAW freundlicherweise Testdatensätze aus dem Althochdeutschen Wörterbuch bereitgestellt.

» Um frühzeitig auch Aspekte der Multimodalität von Daten zu berücksichtigen, wurde der Austausch mit der Akademie der Wissenschaften in Hamburg initiiert. Die Bereitstellung von Daten wurde begleitet von fachlichem Austausch hinsichtlich der Besonderheiten der Deutschen Gebärdensprache.

» Auch mit weiteren Akademien bestand bereits fachlicher Austausch, der in der Folge weiter intensiviert werden wird. In einigen Fällen wird dies erst nach Schaffung entsprechender Stellen möglich sein.

Kooperation ELEXAI / Linked Open Data

Im EU-Projekt ELEXAI kooperieren 26 Einrichtungen aus ganz Europa zu verschiedenen Fragestellungen rund um Sprache und KI. Die Kernfragen lauten: Wie kann lexikographische Arbeit von KI profitieren – und wie kann KI von lexikographischem Wissen profitieren?

Eine mögliche Antwort besteht in der Nutzung von Linked Open Data (LOD). Grundsätzlich lassen sich nahezu alle Dinge (Objekte, Ideen, Konzepte, Erscheinungen, etc.) als Teil einer Ontologie beschreiben, die wiederum die Beziehungen zwischen den einzelnen Elementen definiert. Konkret ist es ein Ziel der lexikographischen Arbeit, Wortbedeutungen miteinander zu verknüfen, und zwar in verschiedenen Dimensionen: hierarchisch, etymologisch, über Sprachgrenzen hinweg, etc. Gäbe es einen solchen vollständigen Katalog aller Wörter und Bedeutungen in allen Sprachen, so ließen sich verschiedenste Fragestellungen der Sprachwissenschaft auf sehr elegante Weise angehen. Auch LLMs hätten damit eine ganze neue, "wahre" Wissensbasis, und müssten nicht allein auf Basis von Wahrscheinlichkeiten operieren.

Weder die BAdW, noch PDL oder ADL sind formal Teil von ELEXAI. Jedoch wurde mit der BBAW, die Projektpartner des Konsortiums ist, eine lose Zusammenarbeit im Bereich LOD vereinbart. Mögliche Bereiche der Kooperation: Berücksichtigung eines Formats wie Triple Stores als Teil der ADL-Forschungsdatenbank sowie Vernetzung der Dialektwörterbücher und des DWDS auf Bedeutungsebene.

Konzept Inhalte Forschungsdatenbank

Als eine der komplexesten Aufgaben des Pilotprojektes erweist sich die Definition der Inhalte der geplanten Forschungsdatenbank. Diese sollen einerseits verschiedene Formen der algorithmischen Analyse ermöglichen; andererseits sollen die Wörterbuch-Datenbestände an den jeweiligen Akademien verortet bleiben. Welche Inhalte hat die Forschungsdatenbank also?

Diese Frage ist nicht nur technischer oder akademischer Natur, sondern auch beispielhaft für unterschiedliche Sichtweisen zu Daten-, Speicher- und Austauschfragen an den beteiligten Akademien. In der Folge erscheint auch eine einseitige Definition oder Beantwortung dieser Frage als wenig zielführend. Vielmehr muss zeitnah im gemeinsamen Austausch und Verständnis aller Akademien eine wenigstens vorläufige Definition der Inhalte erfolgen. Viele weitere Arbeiten wie etwa das weitere Datenbankdesign bauen darauf auf.

Damit lautet das leistbare, naheliegende Ziel des Pilotprojektes: Problembeschreibung, Kommunikation, Konzeptentwürfe zur Datenbank.

Ausbau LexoTerm

Wie genau Forschungsdatenbank und zugehöriges Webportal weiter ausgebaut werden, liegt ebenso vorwiegend an gemeinsamen Richtungsentscheidungen der ADL-Partner. In seiner gegenwärtigen Form kann es beliebige Wörterbücher aufnehmen, ist gerüstet für die Erweiterung auf Korpora, und bietet eine Plattform für algorithmische Werkzeuge zu lexikographischen und korpuslinguistischen Analyse.

Drei Erweiterungen stehen davon unabhängig noch in den nächsten Monaten an: Die Verfeinerung der Suche, die testweise Einbindung von LLM-Funktionen, und die Integration einer Zugangskontrolle über akademische Credentials.

ERWEITERE SUCHFUNKTIONEN

Aufgrund des für Analysen optimierten Datenformats erlaubt es die Forschungsdatenbank, in LexoTerm Suchfunktionen umzusetzen, die weit über Standard-Algorithmen hinausgehen. Dazu zählen insbesondere linguistisch motivierte und KI-gestütze Suchstrategien, die einige Kern-Herausforderungen der lexikographischen Arbeit massiv erleichtern können:

» Unscharfe Suche: Obwohl die Suche nach ähnlichen (und nicht exakten) Begriffen nicht neu ist, können in LexoTerm linguistisch motivierte Ähnlichkeitssuchen umgesetzt werden, die z.B. bekannte Regelmäßigkeiten in Lautwandel oder regionalen Unterschieden nutzen, um gezielter mögliche Verwandtschaften zwischen Einträgen über verschiedene Wörterbücher hinweg zu identifizieren.

» Semantische Suche: Traditionell wird in der Lexikographie ein großer Aufwand betrieben, um Lemmata über Wörterbücher hinweg auf Bedeutungsebene miteinander zu vernetzen, wofür bspw. mit manuellen Sachgruppen-Annotationen gearbeitet wird, die konzeptbedingt nur zum Teil funktionieren. Als Ergänzung bieten moderne, Transformer-basierte, distributionelle semantische Verfahren (Embeddings mit BERT u.ä.) einen alternativen Ansatz, um semantisch ähnliche Bedeutungen unüberwacht (d.h. ohne menschliche Annotation) zu ermitteln. Durch das Datenformat und die Einbindung verschiedener Ressourcen soll LexoTerm es möglich machen, diese Art der Suche wörterbuch-übergreifend anzubieten. Auch für die Untersuchung einzelner Wörterbücher bietet semantische Suche ein immenses Forschungspotenzial, um ohne Kenntnis der konkreten Formen nach Konzepten zu suchen.

» LLM-basierte Suche: LexoTerm bietet alle nötigen Voraussetzungen, um das Potenzial generativer Sprachmodelle für die lexikographische Arbeit auszuschöpfen. Denkbar ist zum einen eine KI-gestütze Suche, die natürlichsprachliche Eingaben in strukturierte Suchabfragen umwandelt;

so könnten Nutzer:innen z.B. fragen, “Welche Diminutiv-Endungen gibt es im Bayerischen?”, was das Modell auf Basis der vorliegenden Daten zu beantworten versucht. Auch komplexe Suchanfragen können so informell gestellt werden, ohne sich bspw. mit den Details regulärer Ausdrücke beschäftigen zu müssen.

EINBIDNUNG VON LLM-FUNKTIONEN

Ferner können auf Basis der Forschungsdatenbank Ko-Piloten und LLM-basierte Plugins entwickelt werden, die Lexikograph*innen bei ihrer Arbeit zu unterstützen, indem sie z.B. Formulierungen für Wörterbuchartikel schreiben, Auto-Vervollständigung anbieten oder Teile von Artikeln (insbesondere stark formularische Segmente) verfassen, die von den Wörterbuch-Redaktionen nur noch geprüft werden müssten. Zur Unterstützung dieser Aufgabenstellung sind auch insbesondere Instruct-Modelle besonders geeignet – siehe nächstes Kapitel.

ZUGANG ÜBER AKADEMISCHE CREDENTIALS

Um den Zugang einerseits so frei und einfach zu machen wie möglich, andererseits aber die hinterlegten Ressourcen vor Missbrauch zu schützen, wird LexoTerm ein Zugangssystem erhalten, das auf akademischen Credentials beruht. Die Mitwirkung in einer forschenden Einrichtung muss daher nachgewiesen werden, um Zugang zu erlangen. Sobald ein Zugangs- und Session-System etabliert ist, können grundsätzlich auch noch weitere Berechtigungsstufen integriert werden.

Training eines Instruct-Modells auf BDO-Daten

Der ursprüngliche PDL-Projektplan skizzierte auch die Möglichkeit des Trainings eines eigenen Sprachmodells, wobei diese Idee aufgrund des Aufwandes und der benötigten Ressourcen zunächst zurückgestellt wurde.

Um dieses spannende, zukunftsweisende Feld nicht unbearbeitet zu lassen, prüfen wir ggw. die Idee, zumindest ein sogenanntes Instruct-Modell zu trainieren und mit einer RAG-Pipeline zu kombinieren.

Instruct-Modelle können über Finetunings größerer Modelle trainiert werden, und sind damit deutlich weniger aufwändig. Sowohl aufgrund des nicht abgeschlossenen Charakters der zugrundeliegenden BDO-Wörterbücher wie auch konzeptionell sind sie nicht dafür auslegt, vollständige Texte (z.B. in Dialekt) zu formulieren. Dennoch würde solch ein Modell Wissen über die zum Training verwendeten Daten erlangen: dialektale und historische Schreibvarianten, lexikographische Terminologie oder auch Artikelaufbau. Mögliche Einsatzgebiete wäre die Unterstützung bei der Formulierung von Definitionen, Inhaltsparaphrase, oder evtl. die automatisierte Transformation in strukturierte Elemente (z.B. TEI).

Um maximale Faktentreue zu erreichen, ließe sich ein Instruct-Modell zudem mit einem RAG-Ansatz kombinieren. Im Zuge der Retrieval Augmented Generation werden vorgängig passende Wissenseinheiten, beispielsweise Bedeutungsdefinitionen, vektorisiert. Diese werden dann über Ähnlichkeitsmessungen dem Ursprungsprompt beigefügt und können die Antwort durch geprüfte Verweise auf existierende Originalartikel präzisieren.

Regelwerk Datenaustausch

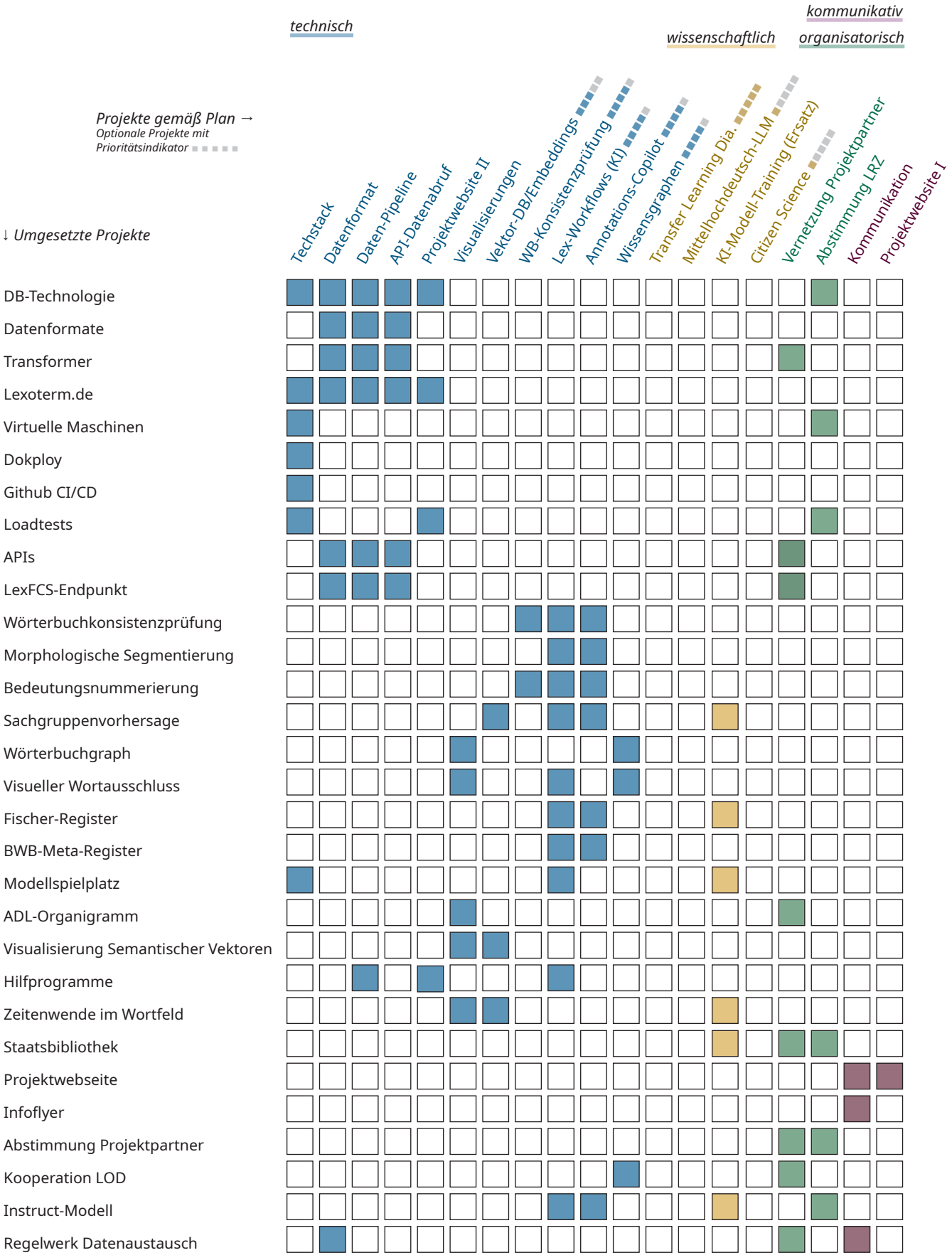
Entscheidend für den Erfolg des Akademienzentrums Digitale Lexikographie des Deutschen wird ein neuer Umgang im Austausch, in der Bereitstellung und Veröffentlichung von Daten sein. Die Lexikographie kann von einer besseren Vernetzung stark profitieren. Während beispielsweise die Naturwissenschaften umfangreiche, zugängliche Datenbanken pflegen, von Gensequenzen bis Radioteleskop-

signalen, ist das Teilen von Forschungs- und Primärdaten im Bereich der Wörterbücher keineswegs Standard.

Auf Initiative des Akademiepräsidenten Prof. Dr. Schwaiger wurde im Rahmen des Pilotprojekts mit der Ausarbeitung eines Entwurfs für ein Regelwerk zum vertrauensvollen Datenaustausch im ADL-Netzwerk begonnen. Während Teile davon bereits vorliegen – etwa die Beschreibungen der Selbstverpflichtungen der Akademien, aber auch ein konkreter Vorschlag für verschiedene Umfänge mit Bezug auf Wörterbuchartikel – fehlt noch ein übergreifender Modus sowie eine Definition von jeweiligen Rechten und Pflichten. Ähnlich wie bei der Inhaltsdefinition der Forschungsdatenbank erscheint es angemessen, diese zentralen Passagen im gemeinsamen Einvernehmen mit den anderen Akademievertretern zu erarbeiten. Vor diesem Hintergrund sind die Arbeiten vorerst bis zum Anlauf des ADL zurückgestellt.

PROJEKTPLAN-MATRIX

Am Anfang war der Plan: Dieser Matrix-Abgleich zeigt, wie sich die aktuellen und zukünftigen Projekte in den usprünglichen Plan eingliedern, und in wieweit die geplanten Ziele abgedeckt sind.



Fazit

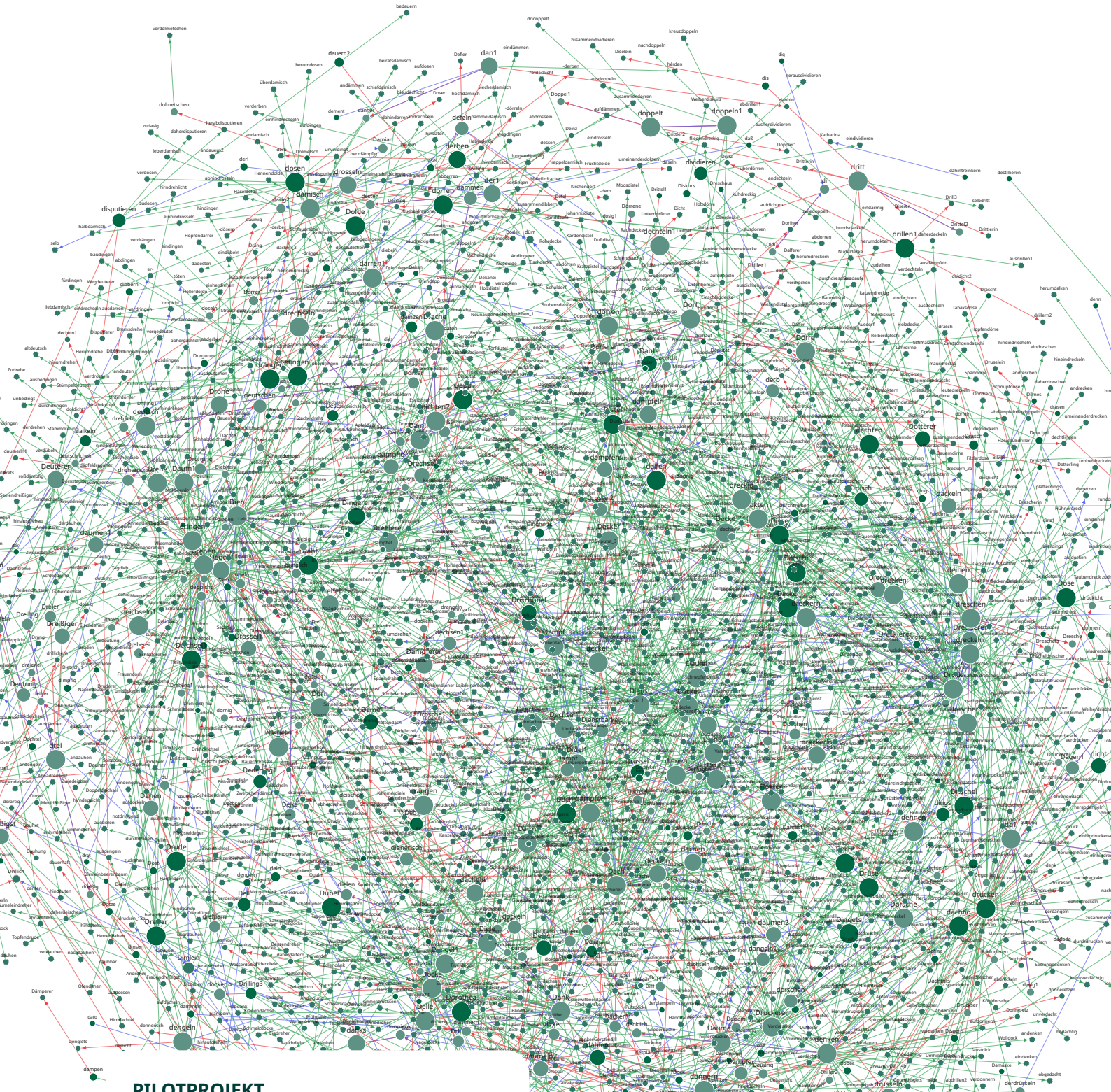
Das erste Jahr des Pilotprojekts "Neue Potenziale für die digitale Lexikographie des Deutschen", das am 30.09.2026 vollendet ist, hat sich als produktive und fruchtbare Standorterkundung erwiesen, mit vielen konkreten Ergebnissen, die sich auch unmittelbar im Redaktionsalltag als nützlich erweisen. Damit zeigt sich auch das Potenzial für weitere Entwicklungen.

Mit der Konstitution des Akademienzentrums Digitale Lexikographie im ersten Quartal 2027 werden sich wichtige Weichenstellungen ergeben, etwa hinsichtlich der Frage des Designs der Forschungsdatenbank. Die Grundlagen für Erweiterungen in vielerlei Richtungen sind gelegt. Von der fortschreitenden Vernetzung mit den Partnern im Akademienzentrum werden auch die Ergebnisse im Pilotprojekt profitieren: durch Klärung und Schärfung, aber auch fachlichen Austausch.

Der zu Beginn des Projektes aufgestellte Projektplan konnte – mit Nachschärfungen – eingehalten werden. Einzig im Bereich des LLM-Trainings verlief die Entwicklung anders als erwartet, weil geplante Kooperationsprojekte bisher nicht realisiert werden konnten. Daher wird für das zweite Jahr das Training eines Instruct-Modells in den Blick genommen. Diese Modellklasse hat sehr nützliche Eigenschaften mit Blick auf die Unterstützung lexikograhischer Workflows, benötigt aber im Vergleich zu LLMs deutlich weniger Rechenaufwand.

Das Forschungsprojekt "Zeitenwende im Wortfeld" befindet sich auf einem guten Weg. Ob die vorbereiteten Algorithmen wirklich wie erhofft den entsprechenden Bedeutungswandel aufzeigen können, wird sich Ende November 2026 zeigen, wenn die Datenanalyse der DNB-Bestände abgeschlossen ist.

Auch LexoTerm als zentrale Anlaufstelle im Web für Forschungsdatenbank und analytische Werkzeuge wird eine weitere Aufwertung durch neue Features erfahren.



PILOTPROJEKT

NEUE POTENZIALE FÜR DIE DIGITALE LEXIKOGRAPHIE DES DEUTSCHEN

Autoren

Wolfgang Huang

Dr. Fabian Simonjetz

Kontakt

pdl.badw.de

lexoterm.de

kontakt@pdl.badw.de

Artikel des Buchstaben "D" des Bayerischen Wörterbuchs, dargestellt als Graph über Etymologie, Ableitung, Wortfamilie und Kompositum