Published in: Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology, Sep 30–Oct 2 2015, University of Duisburg-Essen, Germany - Gesellschaft für Sprachtechnologie and Computerlinguistik e. V., pp. 162-169

IDaSTo – Ein Tool zum Taggen und Suchen in historischen Paralleltexten

Rahel Beyer
Institut für luxemburgische Sprach- und Literaturwissenschaft
Universität Luxemburg
L-4366 Esch-Belval, Luxemburg
rahel.beyer@uni.lu

Abstract

Ein integriertes Datenbank-, Such- und Tagging-Tool (IDaSTo) wird vorgestellt, das sich besonders für Variablenanalysen, für Paralleltexte und für diachronische Untersuchungen eignet. Relevante Kategorien bzw. Variablen können individuell definiert, Tags frei im Text und auf verschiedenen Wegen gesetzt und ihre Häufigkeiten in den verlinkten Statistiken direkt abgerufen werden.

1 Einleitung[1]

Die historische Soziolinguistik greift zunehmend auf (große) Korpora zurück und führt an ihnen variablenanalytische Untersuchungen u.a. im Kontext der Erforschung der Sprachstandardisierung durch (vgl. z.B. Durrell et al., 2008; Elspaß, 2005; Vosters et al., 2012). Zu den Arbeitsschritten gehört es, Varianten in den Texten ausfindig zu machen, diese zu sammeln und auszuzählen. Von besonderem Interesse sind dabei die Entwicklungen von Strukturen auf allen sprachlichen Ebenen im Laufe von teils großen Zeiträumen und ihr soziohistorischer Kontext. Dementsprechend wichtig ist es, das Gesamtkorpus auf der Grundlage von Subperioden zu analysieren und ggf. Ergebnisse verschiedener Textsorten voneinander getrennt zu halten, jedoch jeweils die übergreifenden Verhältnisse präsent zu haben. Zu einem solchen Kontext gehören auch funktionale Aspekte bzw. Metadaten den Text als Ganzes bzw. seine äußeren Merkmale betreffend wie Entstehungsjahr, Drucker(haus), Unterzeichner u.ä. Diese Rückbindung ist v.a. dann von Relevanz, wenn es um Sprachwandel und die Auswirkungen von und Wechselwirkungen zwischen sprachlichen und gesellschaftspolitischen Faktoren geht.

Außerdem gewinnen in der historischen Soziolinguistik Paralleltextkorpora immer mehr an Aufmerksamkeit (vgl. Claridge, 2008). Diese sind für Untersuchungen in Sprachkontaktkontexten umso attraktiver als dass gerade auf der Ebene von grammatischen Phänomenen ein Nachweis von kontaktinduzierten Veränderungen immer wieder als problematisch diskutiert wird (vgl. Heine, 2009). In Paralleltexten lassen sich direkte Abgleiche zwischen zwei oder mehreren Sprachversionen vornehmen und auf diese Weise z.B. Konvergenzen erkennen. Aber auch im Hinblick auf funktionale Erkenntnisinteressen bieten sie vielversprechende Ansatzpunkte für Untersuchungen. Wurden beide Sprachversionen auf ein Papier gedruckt, so ist z.B. ihre Abfolge von Bedeutung, d.h. welcher Text links (bei horizontaler Anordnung) bzw. oben (bei vertikaler Anordnung) steht.

Aus der beschriebenen Vorgehensweise bei der Variablenanalyse lässt sich der Mehrwert maschineller Unterstützung unschwer ableiten. Gerade für eine datengesteuerte und explorative Identifizierung von Variablen bedarf es jedoch flexibler, nicht voreingestellter Annotationskategorien. Bestehende Annotationsprogramme fokussieren sich in der Regel auf spezifische sprachliche Aspekte bzw. lassen sich toolabhängig für die Aufbereitung nur bestimmter sprachlicher Phänomene bzw. vordefinierter Ebenen (mittels vorinstallierter Tagsets) einsetzen. Web-Anno (Yimam et al., 2013) bietet zwar auch die Einrichtung von benutzerdefinierten Annotationsebenen an, allerdings beschränkt sich die Funktionalität – wie bei vielen Annotationsprogrammen – auf eine Anreicherung mit linguistischen Informationen innerhalb von Texten. Ihre Auswertung muss dementsprechend extern

[1] Ich danke den anonymen Gutachtern sowie Instituts- und Projektkkollegen für hilfreiche Kommentare und Hinweise.

geschehen.[2] Eine variablenanalytische Untersuchung bedarf jedoch einer zentralen Verwaltung inklusive einer Übersicht der Phänomene (Variablen) und ihrer Varianten sowie der Häufigkeiten, von der aus zudem Belegstellen nach Bedarf aufgerufen werden können. Auch Standardmethoden der Korpuslinguistik wie Frequenzlisten und Konkordanzen sollten unmittelbar mit einem variablenanalytischen Tool verknüpft sein. Eine solche Kombination verschiedener Funktionen in einem integrierten Tool gibt es jedoch bislang nicht.[3] Die Suche in Paralleltexten bringt weitere Anforderungen mit sich, die von gängigen Konkordanzprogrammen ebenfalls nicht erfüllt werden. tlCorpus[4] z.B. kann weder sprachspezifisch suchen noch den Suchbegriff als Variante eines (sprachlichen) Merkmals auszeichnen. Außerdem ist das Filtern der Texte, die durchsucht werden sollen, aufwändiger und wenig flexibel.

Der vorliegende Beitrag stellt eine Anwendung vor, die nun den genannten Bedürfnissen Rechnung trägt. Dabei handelt es sich weniger um ein klassisches computerlinguistisches Tool (z.B. stehen weder Lemmatisierung noch Parts-of-speech- noch Morphologie-Tags o.ä. zur Verfügung), vielmehr wurde das Programm speziell auf variablenanalytisches Arbeiten zugeschnitten, bei dem abgrenzbare Konstruktionen und Phänomene in den Fokus genommen werden (z. B. orthografische Variation, afinite Nebensätze, der realisierte Kasus nach bestimmten Präpositionen oder Entwicklungen im Wortschatz). Zudem wurde es zu Beginn eines historisch-soziolinguistischen Projekts entwickelt, so dass eine zügige Einsatzmöglichkeit erforderlich war. Letzten Endes konnte eine pragmatische Lösung gefunden werden, die jedoch hochspezifiziert für diachrone Variablenanalyse an Paralleltexten ist. Gegenstand des besagten Projekts ist die Standardisierung des Deutschen in einem Mehrsprachigkeitskontext. Variation bzw. Variantenreduktion soll hier v.a. anhand von historischen, zweisprachigen Paralleltexten untersucht werden.[5] Durch die Nutzung parallel zur Entwicklung konnten konkrete Bedürfnisse ermittelt und im Programm berücksichtigt werden.

[2] Vgl. etwa die Annotation in CorA und die anschließend notwendige Importierung der Transkripte in ANNIS zur Durchsuchung und Visualisierung (Bollmann et al., 2014).
[3] Vgl. z.B. die Zusammenstellung unter https://www.linguistik.hu-berlin.de/institut/professuren/korpus-linguistik/links/software. Aus Platzgründen kann eine ausführliche Evaluation jedes der dort aufgeführten Tools an dieser Stelle nicht stattfinden.
[4] http://tshwanedje.com/corpus/.
[5] S. das Datenbeispiel in Abb. 2.

2 Beschreibung des Tools

Grundsätzlich handelt es sich um eine Mischung aus Datenbank, Suchprogramm und Tagging-Tool. Neben der Funktion als Datenbank, aus der entsprechend den eingegebenen Filterkriterien verschiedene Datensätze aufgerufen und angesehen werden können, sollten auch in den einzelnen Texten Tags frei verteilt werden, Tokens im Fließtext gesucht und die Suchergebnisse ebenfalls getaggt werden können.

Außerdem galt es zu berücksichtigen, dass das Projekt an verschiedenen Standorten bearbeitet wird. Dementsprechend wird eine web-basierte Architektur[6] verwendet, d.h. alle Daten sowie die Anwendung werden auf einem Server gespeichert, der für alle Benutzer über einen beliebigen Internetbrowser zugänglich ist. Dadurch kann problemlos von jedem Computer darauf zugegriffen werden, es bedarf keiner lokalen Installation und alle Benutzer arbeiten an derselben Version.

2.1 Datenbank

Von der Anmeldung gelangt man zunächst automatisch auf die Startseite, auf der alle Datensätze aufgelistet werden.[7] In dem konkreten Projekt[8] handelt es sich dabei um öffentliche Bekanntmachungen, die mehrheitlich als zwei, parallel auf einem Dokument angeordnete Sprachversionen vorliegen. Diese wurden zunächst mit einem Großformatscanner erfasst. Informationen über die Bilddateien sowie deren Metadaten (Signatur, Datum, Titel, Verantwortlicher und gebrauchte Sprach(en)) wurden ebenfalls in einer Tabellenkalkulation festgehalten. Anschließend wurden die Dateien manuell text-digitalisiert und in ein XML-Format überführt. Dabei blieb die Originaltextstruktur durch Taggen der essentiellen Merkmale wie Überschriften, Sprachenwechsel, Schriftarten (Französisch ist in Antiqua, Deutsch meistens in Fraktur) und Groß- und Kleinschreibung nach den internationalen Standards der Text Encoding Initiative (TEI)[9] erhalten.

Durch das Ausfüllen der Datenfelder am Anfang der Seite können jeweils bestimmte Datensätze herausgefiltert werden.

[6] Dabei wurde in PHP programmiert und auf das Webentwicklungs-Framework Symfony 2 bzw. für die Benutzerschnittstelle auf Bootstrap und jQuery zurückgegriffen.
[7] Hier wurde MariaDB als Datenbank-Verwaltungssystem eingebaut.
[8] S. auch http://infolux.uni.lu/standardization/.
[9] TEI Consortium (2015).

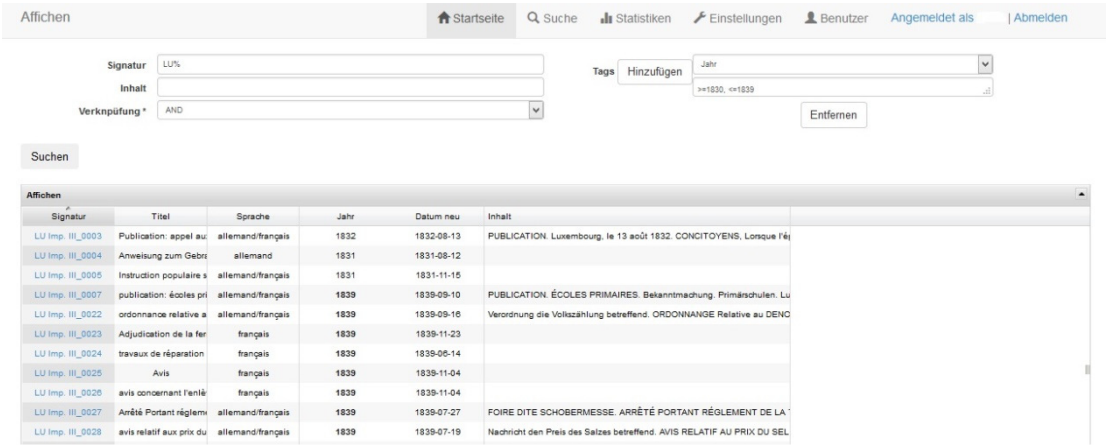

Abb. 1: Screenshot der Startseite

So können z.B. mithilfe des Wildcard-Zeichens ‚%' alle Dokumente mit demselben Signaturanfang, d.h. eines Subkorpus' selektiert und aufgelistet werden. Zur Filterung können jegliche Dokumentenmerkmale, d.h. Metadaten herangezogen werden, die in einem vorangegangenen Schritt definiert und ihre Ausprägungen für die einzelnen Dokumente notiert wurden.[10] Mithilfe von Vergleichsoperatoren können außerdem Zeitspannen ausgewählt werden. Ferner können mehrere Auswahlkriterien miteinander kombiniert werden. In Abbildung 1 bspw. wurden alle Dokumente der Signatur „LU%" aus den Jahren 1830-1839 gesucht.

2.2 Taggen in einzelnen Dokumenten

Durch Anklicken der Signatur öffnet sich das jeweilige Dokument in einer neuen Registerkarte. Hier sind der Original-Scan, die Metadaten zu dem Dokument (grundlegende Merkmale wie Signatur, Sprachen, Titel, Erstellungsjahr usw.) und der Text einzusehen. Werte von Metadaten können an dieser Stelle korrigiert oder neue Metadaten hinzugefügt, d.h. der Text als Gesamtes getaggt werden. Die Metadaten aus dem Kopf der Dokumente finden sich in den Spalten der Datenbank auf der Startseite wieder.[11]

Vor dem Hintergrund der zweisprachigen Ausgabe der Texte und des intendierten (punktuellen) Abgleichs der beiden Sprachversionen[12] war die Trennung und parallele Darstellung der beiden Sprachen von entscheidender Bedeutung. Dieser Schritt steigert die Leserfreundlichkeit und beschleunigt damit den Bearbeitungvorgang. Zu diesem Zweck wurden die Texte zuvor absatzweise auf einem N-Gram-basierten Verfahren automatisch kategorisiert,[13] d.h. in diesem Fall einer Sprache zugeordnet. Für den Fall, dass eine automatische Zuordnung nicht möglich war oder nicht der tatsächlich geschriebenen Sprache entspricht, können die einzelnen Absätze innerhalb des Tools von Hand korrigiert werden. Beide Sprachversionen können wahlweise untereinander in einer Spalte oder in getrennten Spalten nebeneinander, wie in Abbildung 2, angezeigt werden. Für die Auswahl gibt es eine entsprechende Schaltfläche in der rechten Leiste. In den Texten selbst werden per Mouseover auf die Token Links zu verschiedenen Wörterbüchern angeboten.

Vor allem aber können sämtliche Token (d.h. Wortformen sowie Interpunktionszeichen) des Textes ausgewählt und anschließend mit verschiedenen Informationen und hinsichtlich verschiedener Aspekte getaggt werden. Bezogen auf das korpuslinguistische bzw. variablenanalytische Vorgehen kann also das Vorkommen einer Variante eines relevanten sprachlichen Phänomens (Variable) mittels Tags dokumentiert werden (z.B. die <ä>-Variante von z.B. *Kerker*). Genauso können jedoch auch durch Aktivieren der Steuerungstaste mehrere Token gleichzeitig ausgewählt und somit mehrteilige Ausdrücke oder Phrasen, die z.B. eine bestimmte Spracheinstellung zum Ausdruck bringen (z.B. „unsere geliebte Muttersprache"), getaggt werden. Ein Tag ist folglich immer als Merkmal-Wert-Paar aufgebaut, bei dem der Tagname bzw. die Variable dem Merkmale entspricht und als Wert der

[10] Diese Meta-Daten können entweder aus einer Tabellenkalkulation importiert oder in IDaSTo eingegeben werden (s. Abschnitt 2.2).

[11] Die Anzeige der Spalten kann individuell nach Bedarf des Benutzers in den „Einstellungen" angepasst werden, d.h. Spalten mit Metadaten können hinzugefügt oder ausgeblendet werden.

[12] Zur Aufdeckung von Replikationen und Transferenzen.

[13] Der Algorithmus wurde basierend auf Cavnar und Trenkle (1994) gebildet.

konkrete Beleg bzw. die vorzufindende Variante eingesetzt wird. Es gibt keine vordefinierten oder obligatorischen Tags oder Tagkategorien; vielmehr können Tagname (z.B. „<ä>-<e>-Variation“ oder „Spracheinstellung“) und zugehöriger Wert (d.h. Variante, z.B. <ä> oder „emotionales Motiv“) individuell und flexibel erstellt werden. Da für den Wert ein freies Textfeld zur Verfügung steht, können z.B. auch notizartige Teilsätze als Werte eingegeben und die Tagfunktion kann als eine Art Lesezeichen eingesetzt werden. Bereits im Laufe des Projekts erstellte Tagkategorien können indes aus einem Dropdown-Menü ausgewählt werden. Außerdem wird der zuletzt vergebene Tag (d.h. Tagname und Wert) im Tagging-Fenster aufgeführt und kann direkt angewendet werden (s. Abbildung 3). Dabei kann jedes Token mit beliebig vielen Tags ausgezeichnet werden. Bereits ausgezeichnete Wortformen sind grün unterstrichen und bei ihrer Auswahl werden vergebene Tags inklusive Werten in der rechten Leiste angezeigt (vgl. Abbildung 2).

Nicht zuletzt kann für ein ausgewähltes Token direkt aus dem Dokument eine Suche erstellt werden, d.h. ein mit dem betreffenden Token als Suchbegriff, ihm zugewiesene Tags und zugeordnete Sprache vorausgefülltes Suchformular öffnet sich in einem neuen Tab.

Abb. 2: Screenshot der Dokumentenansicht

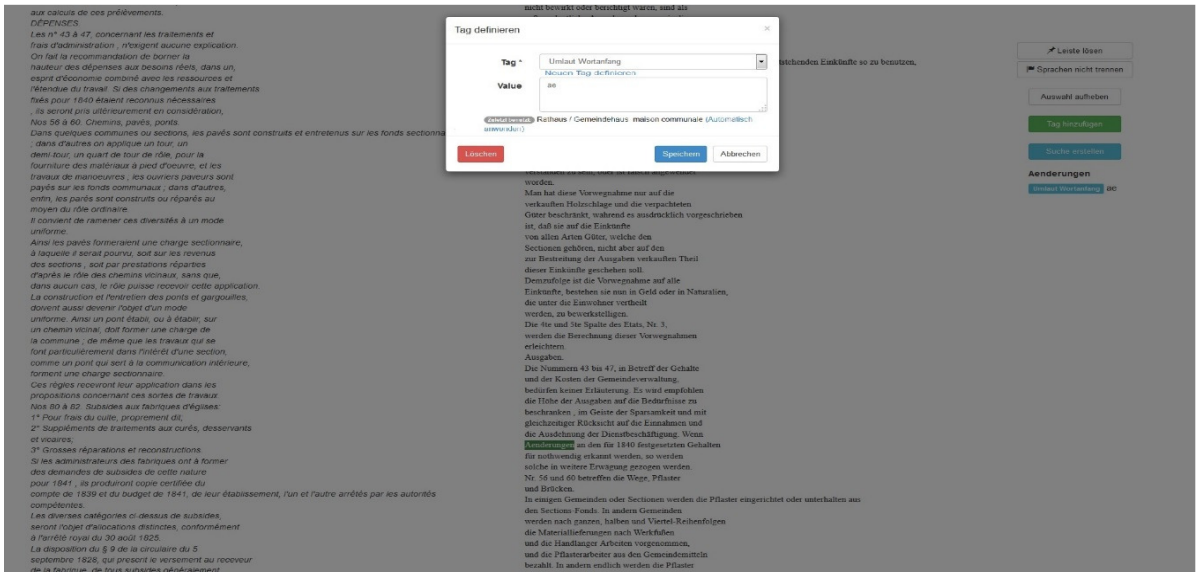

Abb. 3: Screenshot des Tagging-Fensters

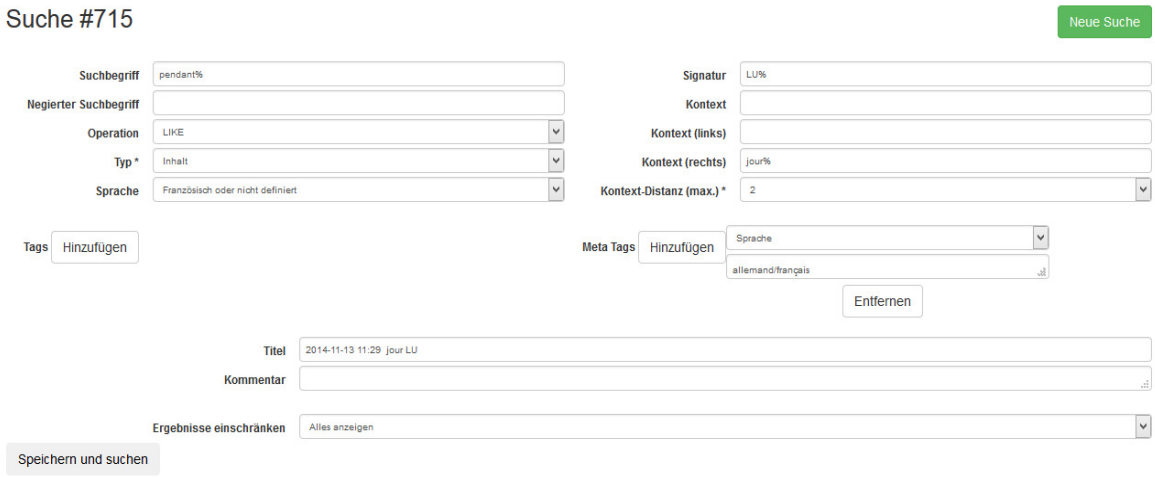

Abb. 4: Screenshot der Suchspezifizierungen

2.3 Das Suchmenü

Unter dem Menüpunkt „Suche“ werden zunächst alle bereits ausgeführten Suchen aufgelistet. Alternativ kann eine neue Suche gestartet werden.

Der Suchbereich für eine beliebige Zeichenkette kann hinsichtlich verschiedener Aspekte eingeschränkt bzw. präzisiert werden. Darunter befinden sich Sprache, Subkorpus (d.h. Signaturanfang), evtl. gespeicherte Tags sowie flexibel zuwählbare Metadaten. Bezüglich der Operationen stehen die exakte Suche nach einem String und die Suche mit vordefinierten Zeichen (LIKE [nur mit Wildcard-Zeichen ‚%‘] oder reguläre Ausdrücke) zur Auswahl. Des Weiteren kann man sich für verschiedene Inhaltstypen entscheiden („Inhalt“ [ignoriert sowohl Groß- und Kleinschreibung als allerdings auch Sonderzeichen, z.B. Umlaute], „Groß-/Kleinschreibung beachten“ sowie „Normalisierter Inhalt“).

Um über einzelne Wörter hinaus Wortfolgen ausfindig machen zu können, wurde eine Suche im (rechten, linken oder unspezifizierten) Kontext des eigentlichen Suchbegriffs implementiert. Diese Option wird relevant, wenn z.B. die Realisierungen bestimmter Mehrwortlexeme oder Nomen-Verb-Verbindungen überprüft werden sollen. Vorteil dieser Lösung ist, dass die Distanz zwischen Such- und Kontextbegriff über ein entsprechendes Datenfeld einzelfallspezifisch bestimmt werden kann. So kann z.B. nach der Konstruktion *pendant X jour* gesucht werden, wobei *pendant%* Suchbegriff ist und *jour%* im Kontext mit einer maximalen Distanz von zwei Wörtern zum Suchbegriff stehen soll (s. Abbildung 4).

Unter den Datenfeldern erscheinen nach Beendigung des Suchlaufs die Suchergebnisse in einem integrierten Fenster (s. Abbildung 5). Das Fenster selbst entspricht dem typischen Aufbau von Konkordanzprogrammen. Zusätzlich lassen sich auf der äußersten rechten Seite einzelne Suchergebnisse deaktivieren bzw. nach Deaktivierung bei Bedarf wieder aktivieren. Auf diese Weise können Suchresultate, die nicht dem relevanten Phänomen entsprechen, aussortiert und von einer weiteren Verarbeitung ausgeschlossen werden. Optional können in den „Einstellungen“ weitere Spalten mit Metadaten dazugeschaltet werden.

Die Suchergebnisse lassen sich über drei Wege weiterverarbeiten, d.h. taggen. Erstens kann jede der Belegstellen des Suchbegriffs direkt im Fenster mit den Suchresultaten einzeln ausgewählt und wie in Abschnitt 2.2 beschrieben getaggt werden. Zweitens gelangt man durch das Anklicken der Signatur zum Belegdokument, das beim Öffnen direkt zur Fundstelle des Suchbegriffs springt. Dort kann man ebenfalls nach demselben Verfahren Tags vergeben. Für den Fall, dass alle (nach der ‚Bereinigung‘ übriggebliebenen) Suchresultate denselben Tag und dazugehörigen Wert bekommen, kann die Option „Tag auf die Suchergebnisse anwenden“ gewählt werden. Beide Vorgänge können genauso für Meta-Tags durchgeführt werden.

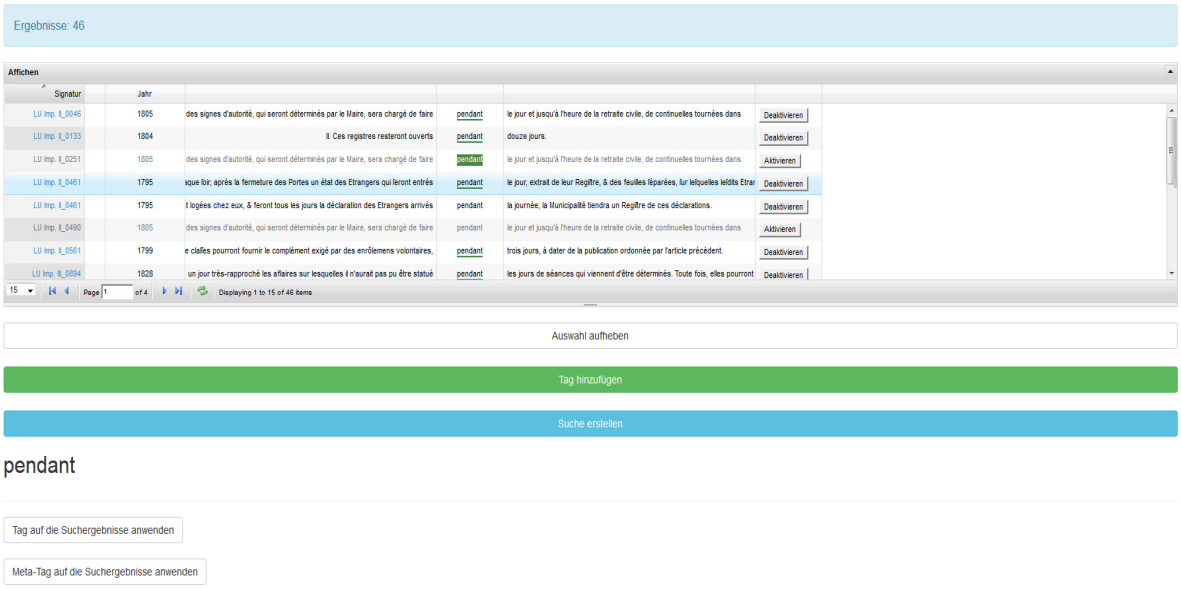

Abb. 5: Screenshot des integrierten Fensters mit der Auflistung der Suchergebnisse

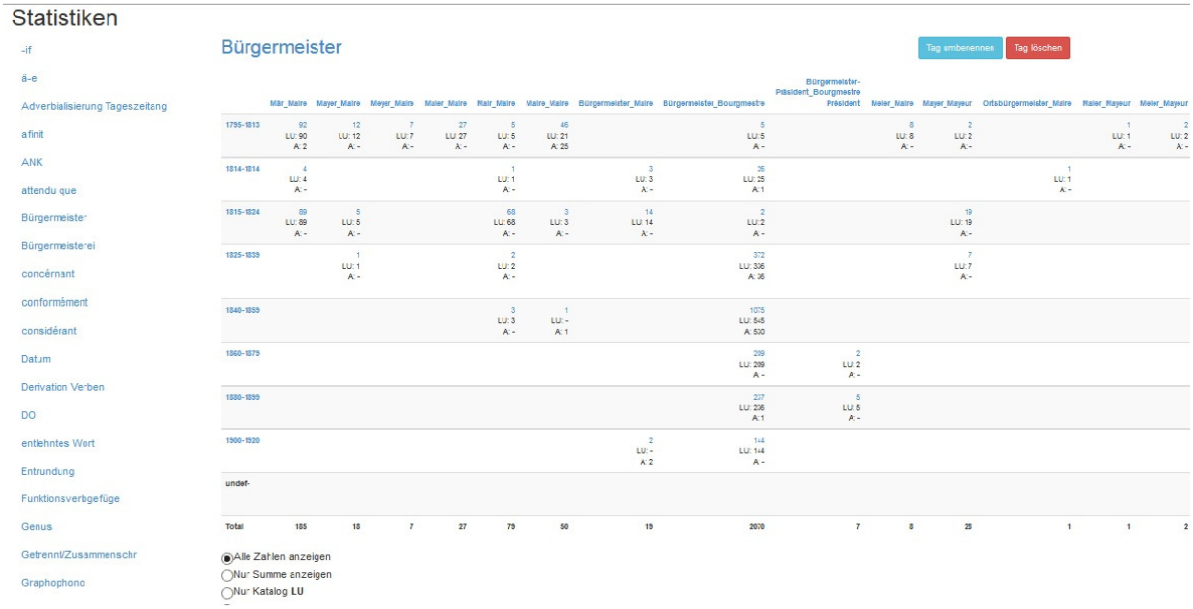

Abb. 6: Screenshot der Statistik für die Variable BÜRGERMEISTER

2.4 Statistiken

Im Menüpunkt „Statistik“ werden definierte Tagkategorien (ähnlich einem Inhaltsverzeichnis) aufgelistet, vergebene Werte gezählt und für jede Tagkategorie (d.h. Variable) in einer Tabelle präsentiert.[14]

Auf diese Weise sind die Tags systematisch gesammelt und müssen nicht von Hand, Dokument für Dokument herausgesucht und ausgezählt werden. Da diese Funktion in das Tagging-Tool integriert ist, ist somit kein Export der getaggten Texte erforderlich. Dieser Menüpunkt liefert somit automatisch die quantitative Auswertung der Variablenanalyse. Da nicht nur die Gesamtzahl aller vergebenen Tags angegeben wird, sondern ihre (absolute) Häufigkeiten nach Zeitintervallen aufgeschlüsselt werden, kann der Statistik die qualitative und quantitative Verteilung von Varianten im Verlauf des Untersuchungszeitraums direkt entnommen werden. Somit wird ein entscheidender Schritt diachroner Analysen maschinell unterstützt. Die Statistik in Abbildung 6 bspw. führt sämtliche Bezeichnungen des Amtes des Bürgermeisters, die in den Bekanntmachungen gefunden und als Werte dieser Variable via Taggen festgehalten wurden, in der Kopfzeile und darunter ihre Belegzahlen auf.

[14] Für die Tabellen wurde das Plugin Flexigrid verwendet.

Die einzelnen Zeitabschnitte können dabei individuell in den „Einstellungen“ bestimmt und geändert werden.

Standardmäßig werden alle Zahlen angezeigt, d.h. sowohl getrennt für die jeweiligen Korpora als auch die Summen für jede Zelle. Unterhalb der Tabelle gibt es jedoch die Möglichkeit, die Zahlen nur eines der Korpora (d.h. nur einen der Signaturanfänge) oder nur die Summen dargestellt zu bekommen. Die Tabelle kann des Weiteren via Copy&Paste-Verfahren in ein Programm zur Tabellenkalkulation übertragen werden.

Außer zur Übersicht über die Zahlen gelangt man über das Statistik-Menü zu den Belegstellen der Varianten. Die Werte (Varianten), Zeitabschnitte sowie Summen jeder einzelnen Zelle sind anklickbar und führen zu den jeweiligen Beleglisten. Diese entsprechen grundsätzlich dem Fenster mit den Resultaten des Suchmenüs. Durch das Anklicken der Signatur kommt man wiederum zum Belegdokument, das beim Öffnen direkt an die Stelle des Tokens springt. Hier ergibt sich ggf. die Möglichkeit zur Korrektur. Andererseits ist die Auflistung der Belegstellen eine wichtige Funktion für die (qualitative) Interpretation der quantitativen Ergebnisse. So lässt sich z.B. feststellen, ob bei dem einen oder anderen Phänomen lexemspezifische Realisierungen bzw. Entwicklungen vorzufinden sind.

Schließlich dient dieser Menüpunkt in gewisser Weise der Variablenverwaltung. So befindet sich an der linken Seite eine Art Inhaltsverzeichnis, das sämtliche definierte Tagkategorien auflistet. Des Weiteren kann man auf dieser Seite auch Tags umbenennen oder löschen.

3 Zusammenfassung

Es wurde ein Such- und Tagging-Tool vorgestellt, das sich besonders für die Durchführung von quantitativen Analysen, besonders für Paralleltexte und besonders für historische Untersuchungen eignet. Es dient weniger einer linguistischen Aufbereitung im Sinne einer Anreicherung von Texten mit linguistischer Information; vielmehr lassen sich relevante Kategorien bzw. Variablen selbst definieren, Tags frei im Text und auf verschiedenen Wegen setzen sowie ihre Werte frei formulieren. Deren Häufigkeiten können in den verlinkten Statistiken direkt abgerufen werden. Die Anwendung zeichnet sich somit durch Flexibilität und Teilautomatisierung aus. So können Autovervollständigungsfunktionen bei Datenfeldern, (teilweise) vorausgefüllte Suchformulare und die Angabe des zuletzt vergebenen Tags genutzt werden. Die Flexibilität bezieht sich auch auf benutzerspezifische Einstellungen der angezeigten Informationen und der zeitlichen Unterteilung des gesamten Untersuchungszeitraums. Verschiedensprachige Versionen desselben Inhalts können nebeneinander angezeigt werden. Nichtzuletzt können neben systemlinguistischen auch textbezogene (und u.U. auch diskursanalytische) Aspekte bearbeitet werden. Das Programm unterstützt Korpuslinguisten somit auf vielfältige Weise und erleichtert ihnen das empirische Arbeiten in vielerlei Hinsicht.

Literatur

Marcel Bollmann, Florian Petran, Stefanie Dipper und Julia Krasselt (2014): „CorA: A web-based annotation tool for historical and other non-standard language data”. In: *Proceedings of the EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), Gothenburg, Sweden*. Seiten 86-90.

William B. Cavnar und John M. Trenkle (1994): “N-Gram-Based Text Categorization”. In: *Proceedings of SDAIR-94 (3rd Annual Symposium on Document Analysis and Information Retrieval)*. Seiten 161-175.

Claudia Claridge (2008): “Historical Corpora”. In: Anke Lüdeling und Merja Kytö (Hrsg.): *Corpus Linguistics. Handbücher zur Sprach- und Kommunikationswissenschaft*. Berlin: Mouton de Gruyter. Seiten 242-259.

Martin Durrell, Astrid Ensslin und Paul Bennett (2008): “Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert“. In: Werner Besch und Thomas Klein (Hrsg.): *Der Schreiber als Dolmetsch: Sprachliche Umsetzungstechniken beim binnensprachlichen Texttransfer in Mittelalter und Früher Neuzeit*. Berlin: Erich Schmidt. Seiten 263-279.

Stephan Elspaß (2005): *Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert*. Tübingen: Niemeyer.

Bernd Heine (2009): “Identifying instances of contact-induced grammatical replication”. In: Samuel G. Obeng (Hrsg.): *Topics in Descriptive and African Linguistics: Essays in Honor of Distinguished Professor Paul Newman.* Munich: LINCOM EUROPA. Seiten 29-56.

TEI Consortium (Hrsg.) (2015): *TEI P5*: *Guidelines for Electronic Text Encoding and Interchange*. [Version 2.8.0]. http://www.tei-c.org/P5/.

Rik Vosters, Gijsbert Rutten und Wim Vandenbussche (2012): “The sociolinguistics of spelling. A corpus-based case study of orthographical varia-

tion in nineteenth-century Dutch in Flanders". In: Ans M.C. van Kemenade und Nynke de Haas (Hrsg.): *Historical Linguistics 2009: Selected papers from the 19th International Conference on Historical Linguistics, Nijmegen, 10-14 August 2009*. Amsterdam/Philadelphia: John Benjamins. Seiten 253-274.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho und Chris Biemann (2013): "WebAnno: A flexible, web-based and visually supportedsystem for distributed annotations". In: *Proceedings of ACL 2013 System Demonstrations*. Seiten 1-6.