WebLicht als korpuslinguistisches Analyseinstrument für studentische Forschungsarbeiten – am Beispiel von Vorwissenschaftlichen Arbeiten

Andrea Steiner-Cardell (Universität Wien)
Betreuerin: Mgr. Michal Dvorecky, PhD (Universität Wien)

Abstract:
Für die Analyse großer Datenmengen von gesprochener oder geschriebener Sprache bietet das Forschungsfeld der Korpuslinguistik viele Möglichkeiten: Einzelne Elemente natürlicher Sprachen können untersucht werden, aber auch eine Analyse der Gesamtstruktur der Sprache ist möglich. Liegen Sprachdenkmäler in digitalisierter Form vor, so können auch jene Gegenstand einer korpus-basierten Analyse sein. Die Untersuchungen am Korpus führen dazu, dass durch Rückschlüsse neue Theorien gebildet werden und darauf aufbauend neue Erkenntnisse erlangt werden. So auch die korpusbasierte Untersuchung der Autor*innen Cordula Meißner und Franziska Waller, welche ver-sucht haben, das Repertoire an sprachlichen Formulierungen in geisteswissenschaftlichen Fachdis-ziplinen abzubilden, um so festzustellen, welche Elemente der alltäglichen Wissenschaftssprache angehören und wo Überlappungen zwischen den einzelnen Disziplinen vorkommen.
Liegt der Fokus nun aber nicht auf den Ergebnissen einer solchen Korpusanalyse, sondern vielmehr auf der technischen Umsetzung, gibt es einiges zu beachten: von der Gewinnung der Rohdaten über die Bereinigung und Konvertierung bis hin zur Auszeichnung der Daten mit Metadaten. Exakt jene Schritte werden in nachfolgenden Artikel beleuchtet, wobei es darum geht aufzuzeigen, wel-che Möglichkeiten das Open-Source-Softwareprogramm WebLicht für die Auseinandersetzung mit linguistischen Fragestellungen bietet und was es dabei zu beachten gilt.

Keywords: Korpuslinguistik, Wissenschaftssprache, studierendenorientiert, WebLicht, Methoden in der Schreibwissenschaft

Einleitung

„Forschung heißt, Wissen über etwas zu erzeugen, das vorher noch niemand gewusst hat. In der Praxis bedeutet das, sich in Zonen des Unstrukturierten hineinzubewegen, das bisher noch keinen Namen hatte." (Groebner, 2012, 21) Allerdings reicht es nicht aus, jenes Wissen nur zu erzeugen, sondern dessen Nutzen besteht im Weitergeben, Vertiefen und Verknüpfen, auch mit bereits bestehendem Wissen im interdisziplinä-ren Kontext. Dafür bedarf es einer adäquaten Sprache, die „ein Typ sprachlichen Handelns [ist], in dem Wissenschaft als gesellschaftliche Aufgabe und als Beruf realisiert wird. Dieses Handeln bedient sich allgemeiner und speziell wissenschaftli-cher sprachlicher Mittel in zweckmäßiger Weise." (Graefen, 1997, 73)

Jene Anforderung erfüllt die Wissenschaftssprache: Sie beschreibt Wissenschaft mit dem ihr zugehörigem Vokabular, sie erläutert Zusammenhänge von Erkenntnisgegenständen und Herangehensweisen und gleichsam definiert sie diverse Formen von Wissen (Moll & Thielmann, 2017, 53).

Obwohl jene Definition eindeutig erscheint, gestaltet sich eine konkrete Darstellung von Wissenschaftssprache schwierig. So verfügt sie nicht nur über Gemeinsamkeiten mit der Alltagssprache,

Steiner-Cardell, A. (2020). WebLicht als korpuslinguistisches Analyseinstrument für studentische Forschungsarbeiten - am Beispiel von Vorwissenschaftlichen Arbeiten. zisch: zeitschrift für interdisziplinäre schreibforschung, 3, 28-44.

auch überlappen sich beide hinsichtlich des Wortschatzes, aus dem beide schöpfen. Ehrlich (1999) meint dazu, dass „wissenschaftliches Arbeiten zentral gebunden ist an jene alltägliche Sprache, die das unumgängliche Verständigungsmittel der ganzen Sprachkommunität ist“ (Ehrlich, 1999, 7). Das heißt, dass ohne die Basis der Alltagssprache eine wissenschaftliche Verständigung nicht möglich ist: Eine Vorlesung, ein Laborgespräch, eine Seminardiskussion sind nur deshalb möglich, weil neben den fachsprachlichen Elementen, welche nur einen Teil des Diskurses bzw. des Textes ausmachen, auch alltägliche Sprachelemente zur Realisierung enthalten sind (Ehrlich, 1999, 9).

Weiter wird Wissenschaftssprache als universell verwendbares Konzept verstanden, d.h. einzelne Fachdisziplinen können aus ihrem Wortschatz schöpfen, ohne dass dieser Wortschatz als Teil der Fachsprache der jeweiligen Forschungsrichtung verstanden wird. Konkret bedeutet dies, dass Wissenschaftssprache „sich auf einen Sprachgebrauch bezieht, der als typisch wissenschaftlich be-zeichnet werden kann und aufgrund den allen Wissenschaftsdisziplinen eigenen Handlungsberei-chen auch in allen wissenschaftlichen Disziplinen zu finden sein müsste.“ (Deml, 2015, 25)

Im Zuge eines Projektes zur Bestimmung des Inventars der allgemeinen Wissenschaftssprache der Geisteswissenschaften haben die Autor*innen Cordula Meißner und Franziska Wallner ein Korpus zusammengestellt, welches darauf abzielt, eine lexikalische Grundlage für die wissenschaftspropädeutische Sprachvermittlung zu sein. Jenes soll Aufschluss über

> das Repertoire an Formulierungsroutinen zur Realisierung der jeweiligen Sprachhandlungen [und die sprachlichen Elemente der] Wissensgewinnung in geisteswissenschaftlichen Diszip-linen [geben, denn] hier ist jener Wortschatz zentral, der Ausdrucksmittel für wissensmetho-dologische Inhalte bereitstellt – der Bereich der fächerübergreifend gebrauchten Lexik, [die alltägliche] Wissenschaftssprache. (Meißner & Wallner, 2019, 10)

Um den fächerübergreifenden Aspekt in der Untersuchung gewährleisten zu können, wurden für das Korpus „Gemeinsames sprachliches Inventar der Geisteswissenschaften“, kurz GeSIG, 197 Dissertationen aus 19 geisteswissenschaftlichen Fachbereichen herangezogen.

Meißner und Wallner vertreten in ihrer Studie die Ansicht, dass gerade Studierende von Geisteswissenschaften ein besonderes Bewusstsein für Sprache aufbauen (müssen), denn neben ihrer engen Verbundenheit mit der Alltagssprache und ihrer Funktion als Instrument des Erkenntnisgewinns sowie dessen Präzisierung und Weiterentwicklung bewirkt sie eine Weiterentwicklung und Innovation der Sprachverwendung selbst, das heißt, dass die Autoren*innen mit ihr und über sie ihren individuellen Schreibstil definieren und kreieren. Der Sinn und Zweck besteht demnach nicht nur darin, Wissen zu verstehen, sondern es auch zu gebrauchen und sprachlich adäquat zu be-schreiben (Meißner & Wallner, 2019, 16-18)

Jenes sich aus 19 Teilkorpora zusammensetzende Korpus wird als repräsentative Quelle für den natürlichen und authentischen Sprachgebrauch angesehen, da die korpuslinguistische Bearbeitung erst nach Beendigung des Schreibprozesses durchgeführt wurde. Die Autorinnen haben sich für die induktive Vorgehensweise entschieden: Im Vorfeld der Analyse wurden keinerlei Hypothesen erstellt oder auf anderen Theorien basierende Ansätze übernommen. Vielmehr wurden ausgehend von den Ergebnissen theoretische Rückschlüsse gezogen.

So haben Meißner und Wallner das „GeSIG“ etwa nach Wortarten, Häufigkeitsklassen und

autor*innenübergreifender Lexik untersucht. Weiter haben sie sich auch mit der funktionalen Erschließung des „GeSIG“ Inventars auseinandergesetzt, was am Beispiel des Verbs „darstellen“ exemplarisch aufgezeigt wurde. Mit ihrer Studie konnten sie nicht nur neues Wissen erzeugen und so eine bis dahin bestehende Forschungslücke füllen, sondern sie regten damit zu weiterführenden Fragestellungen und Forschungsvorhaben an – so auch zum Inhalt dieses Artikels, welchem die Frage vorausgegangen ist, wie eine solche korpusbasierte Untersuchung aussehen könnte, wenn jene im Zuge eines studentischen Forschungsprojektes zur Wissenschaftssprache durchgeführt würde. Das bedeutet: Welche Überlegungen gilt es zu treffen und welche Kriterien zu definieren? Welche Einzelschritte sind im Zuge der Korpusbearbeitung notwendig, um eindeutige Ergebnisse zu erhalten und wie sind jene Ergebnisse zu lesen bzw. inwiefern können diese als Basis für weiterführende Fragestellungen gelten?
Die Analyse des Korpus bzw. die daraus resultierenden Ergebnisse stehen nicht im Vordergrund. Vielmehr geht es um das Forschungsdesign und die technische Umsetzung unter Berücksichtigung der Überlegung, dass das analysierte Korpus für weiterführende Forschungsvorhaben verwendet werden könnte. Ein mögliches Forschungskonzept könnte sein, ein ähnliches Projekt wie jenes von Meißner und Wallner durchzuführen, bei dem anstelle von Dissertationen etwa Vorwissenschaftli-che Arbeiten als Textgrundlage verwendet und jene hinsichtlich spezifischer Kriterien untersucht werden, ggf. mit Fokus auf die darin enthaltenen Elemente der alltäglichen Wissenschaftssprache.

Was ist ein Korpus und warum Korpuslinguistik?

Ein Korpus ist grundsätzlich eine digitalisierte Sammlung von geschriebener oder gesprochener Sprache, welche in ein (bestimmtes) kontextuelles Umfeld eingebettet ist. Jene kontextuelle Einbettung ist genau genommen eine Mindestanforderung, welche ein Textkorpus erfüllen muss, das heißt im Gegenzug, dass eine bloße Ansammlung von Wörtern, selbst wenn es sich beispielsweise um eine Liste von Verben mit dem Präfix ver- handelt, noch kein Korpus darstellt. Mit dieser Forde-rung geht das Kriterium der Authentizität einher: Je geringer die Beeinflussung des Texterstellungsprozesses, desto höher der Grad der Authentizität. Allerdings darf in diesem Zusammenhang das Forschungsvorhaben nicht außer Acht gelassen werden, denn jenes bedingt zusätzlich, inwieweit von außen auf die Textproduktion eingegriffen wird bzw. diese von außen reguliert wird (Hirsch-mann, 2019, 2-3). Das übergeordnete Ziel sollte es dennoch sein, die chomsky‘sche Prämisse zu wahren, die besagt, dass „ein Korpus niemals repräsentativ für Sprache sein kann, da eine Sprache unendlich ist, während ein Korpus immer nur einen endlichen Sprachausschnitt darstellt“ (Lenz, 2000, 10).
Unabhängig dessen gilt generell: je umfassender das Korpus, desto größer dessen Aussagekraft.
In ihrer Konzeption werden Korpora als wiederverwertbare Ressource verstanden, was bedeutet, dass mehrere Untersuchungen an ein und demselben Korpus möglich sind. Gleichsam kann ein Korpus auch als Vergleichsobjekt selbst gesehen werden. Somit ergibt dies die Möglichkeit, ähnliche Korpora miteinander zu vergleichen (Lemnitzer & Zinsmeister, 2015, 13). Darüber hinaus tragen korpuslinguistische Analysen dazu bei, „eine empirische Basis für andere Forschungszweige im Bereich [der] Sprachtechnologie zu schaffen und der damit verbundene Gedanke der Standardisiertheit und der Wiederverwendtbarkeit [fungiert]“ interdisziplinär. (Lenz, 2000, 8)

Grundsätzlich sind korpuslinguistische Untersuchungen eher quantitativ als qualitativ und gleichsam werden jene aus einer induktiven als aus einer deduktiven Perspektive her betrachtet. Induktiv ist die Perspektive deshalb, weil aus den analysierten Daten auf Gesetzmäßigkeiten rückgeschlossen wird. Es wird darauf verzichtet, die Daten vorab zu strukturieren, lediglich die empirischen Beobachtungen sind von Bedeutung, anders als bei der deduktiven Methode, wo bereits vor der Ana-lyse explizit festgehalten wird, worauf fokussiert werden soll und welche vorab definierte Hypothese es zu beantworten gilt. Dabei wird eine strikte Trennung dieser Vorgehen nicht als sinnvoll erachtet, denn „induktiv abgeleitete Regeln müssen sich bewähren, indem sie deduktiv an anderen Daten getestet werden." (Bubenhofer, 2009, 17). Demnach bedeutet dies, dass Korpuslinguistik „die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind." (Lemnitzer & Zinsmeister, 2015, 14) ist.

WebLicht – Web-based Linguistic Chaining Tool

Das webbasierte Analyseprogramm WebLicht ist eine sogenannte serviceorientierte Architektur (SOA), die

> an execution environment for automatic annotation of text corpora [is]. Linguistic tools such as tokenizers, part of speech taggers, and parsers are encapsulated as web services, which can be combined by the user into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format. (WebLicht, n.d.)

Entwickelt wurde WebLicht im Oktober 2008 als Teil der Deutschen Sprachressourcen-Infrastruktur (D-SPIN), was wiederum ein Teilprojekt des europäischen CLARIN – Common Language Resources and Technology Infrastructure – Projektes ist (WebLicht, n.d.).

Aus zweierlei Gründen wurde dieses Programm zum Auszeichnen der Daten herangezogen: Zum einen werden Wörter und Satzzeichen als Einzelelemente definiert, was ein essenzielles Auswahlkriterium darstellte. Zum anderen ist es das Grundverständnis von CLARIN, digitale Sprachressourcen und computerbasierte Bearbeitungswerkzeuge als Open-Source-Software für Lehrende, Forscher*innen, Student*innen und Citizen Scientists verschiedenster Disziplinen anzubieten, was die Benutzung der Software – Login mittels Universitätsaccount – absolut unkompliziert gestaltet (CLARIN, n.d.).

Eine Installation auf dem Computer ist nicht notwendig, was den Vorteil mit sich bringt, dass eine ortsunabhängige Bearbeitung des Korpus möglich ist. Weiter ist es auch nicht nötig, dass andere Softwareprogramme bzw. zusätzliche Dateien für die Bearbeitung des Korpus installiert werden, da WebLicht bereits über eine Reihe von Tools verfügt. Darüber hinaus ist WebLicht an kein spezielles Betriebssystem gebunden; es wird lediglich ein Standard-Webbrowser dafür benötigt (Web-Licht, n.d.).

WebLicht folgt nicht nur dem Open-Source-Konzept, sondern auch seine Konstruktion impliziert das Verständnis von data sharing:

> To make sure that all tools are interoperable, WebLicht has implemented a common inter-change format for digitized texts: The Text Corpus Format […]. TCF is used exclusively as in-ternal processing format designed to support efficient data sharing and web service execu-tion. Using TCF ensures interoperability between WebLicht tools and resources." (WebLicht, n.d.)

TCF ist ein XML-basiertes Format, welches mit seiner schichtartigen Struktur garantiert, dass die verschiedenen Annotationsmöglichkeiten in WebLicht als eigenständige Ebenen (layer) konzipiert und so keine anderen Ebenen oder auch Metadaten überschrieben werden (WebLicht, n.d.).

Neben WebLicht gibt es noch andere Programme zur Korpusanalyse. Eines davon ist EXMARaLDA. Es handelt sich dabei um ein flexibles Annotationsprogramm, welches keine vordefinierten Annotationsarten enthält und einzelne standardisierte Untersuchungsprozesse bis zu einem gewissen Grad an das jeweilige Forschungsvorhaben anpassen lässt. Daher wird jenes besonders gerne für kleinere Korpora und für Video- bzw. Audiodaten verwendet (Hirschmann, 2019, 23). Neben dem Transkriptions- und Annotationseditor (Partitur-Editor) enthält es auch ein Verwaltungswerkzeug (Corpus-Manager) sowie ein Such- und Analysewerkzeug (EXAKT) (EXMARaLDA, n.d.).

EXMARaLDA ähnlich sind die Suchtools IMS Open Corpus Workbench (CWB), in welches auch neue Korpora eingepflegt werden können, um diese hinsichtlich Wortart und Syntax (Chunking) zu analysieren. Dieselben Analysemöglichkeiten (z.B. Wortartenannotation, Abbildung von Kollokationen) bietet das Korpustool AntConc, welches vom englischen Korpuslinguisten Laurence Anthony entwickelt wurde. Ebenso stellt TIGERSearch ein Einstiegsprogramm in die Korpuslinguistik dar. Dieses Programm kann Teilkorpora bilden, für die dann spezifische Suchabfragen generiert werden kön-nen (Lemnitzer, 2015, 94).

Forschungsdesign

Bevor nun die zu bearbeitenden Textquellen entweder mittels Copy&Past eingefügt oder als Datei in WebLicht hochgeladen werden können, müssen die einzelnen Texte vor der Analyse bereinigt und konvertiert werden, siehe Arbeitsschritt „Datenaufbereitung“ in der nachfolgenden Grafik.

Grundsätzlich ist die Textanalyse in drei Arbeitsblöcke – Datenerhebung, -aufbereitung und -auswertung – unterteilt, welche Teilarbeitsschritte enthalten. Erst wenn jene abgeschlossen wur-den, kann mit dem nächsten Arbeitsschritt begonnen werden.

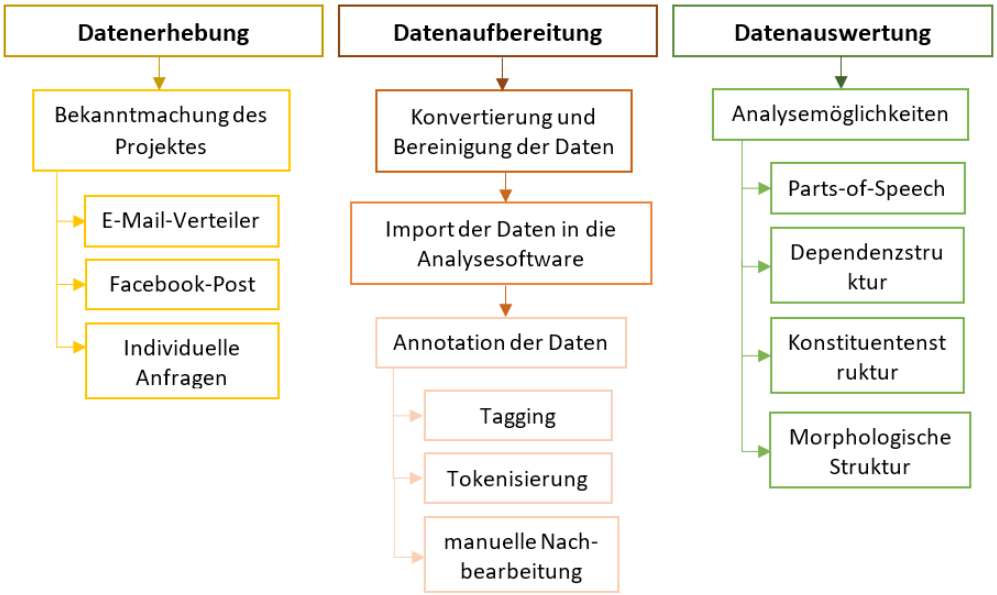


Abbildung 1: Prozessverlauf (eigene Abbildung)

Datenerhebung

Im Zuge der Projektkonzeption wurde die Entscheidung getroffen, Vorwissenschaftlichen Arbeiten (VWA) aus geisteswissenschaftlichen Schulfächern für die Korpusanalyse zu verwenden. Zum einen spielte die Zugänglichkeit über direkte Anfragen (Mailverteiler, Facebook-Post, individuelle Emailanfragen) eine Rolle. Zum anderen werden diese bereits als Gegenstand von studentischen Untersuchungen herangezogen, d.h. Forschungsinteresse und Relevanz bestehen.

Darüber hinaus bestand die Überlegung, dass weiterführende Auseinandersetzungen mit dem in VWA enthaltenen Wortschatz äußerst spannend sind, etwa die Verwendung einzelner Verben oder Verbklassen, der Einsatz von Konjunktionen und deren Umgebung im Satz oder welche Elemente der alltäglichen Wissenschaftssprache in der VWA enthalten sind, zumal es eine Anforderung der VWA ist, wenn auch unabhängig von ihrer fachlichen Zuordnung „inhaltliche Kenntnisse […], Kommunikations- und Diskursfähigkeit, logisches und kritisches Denken sowie eine klare, angemessene Sprachform." (Karmasin & Ribing, 2018, 10) Das bedeutet, dass auch hier die alltägliche Wissenschaftssprache zum Tragen kommt, insbesondere da die VWA „in die Grundlagen des wissenschaftlichen Arbeitens einführen [soll] und […] als Vorbereitung für ein Universitäts- oder Hochschulstudium [dient]." (Karmasin & Ribing, 2018, 9)

Dem Prinzip des höchstmöglichen Maßes an Authentizität folgend, wurden nur jene VWA verwendet, welche zum Zeitpunkt der Analyse bereits fertiggestellt waren. Ergo lassen sich die daraus gewonnen Daten als nicht-elizitiert und nicht-introspektiv bezeichnen, da sie authentischen Produktionskontexten entnommen sind (Hirschmann, 2019, 5).

Datenaufbereitung

Die ursprünglichen PDF-Dateien wurden in Word-Dateien konvertiert, um sämtliche irrelevanten und auswertungsverzerrenden Dokumententeile zu entfernen. Darunter fielen Verzeichnisse, Abbildungen und sonstige nicht als Primärtext geltende Textteile, also jene Textelemente, die keinen wissensverarbeitenden Charakter aufweisen. Besonderer Fokus lag dabei auf den personenbezogenen Daten. Es wurde den Autoren*innen der VWA im Vorfeld versichert, dass keinerlei personenbezogene Daten mit in die Untersuchung einfließen oder im Artikel veröffentlicht werden.

Im Anschluss an die Bereinigung und Konvertierung der Primärdaten wurden diese in WebLicht eingespielt, um Metadaten, also Daten zu den Primärdaten, zu gewinnen. Hierfür war es notwendig, die Primärdaten mittels spezieller Tools auszuzeichnen – zu annotieren –, um sie dann wiederum in ein lesbares Format zu transformieren.

Nachdem die Konvertierung der VWA von PDF in Worddokumente abgeschlossen war und jene in eine Gesamtdatei zusammengefügt wurden, konnte das Korpus nach Auswahl der Sprache (Deutsch) und des Modus – easy oder advanced – in WebLicht importiert und ausgezeichnet werden.

Für die Bearbeitung des Korpus wurde der easy Modus verwendet. In dieser Einstellung bietet WebLicht als serviceorientierte Architektur (SOA) vordefinierte Analyseprozesse an, welche unter-schiedliche Aspekte des Korpus beleuchten: Part-of-Speech Tags and Lemmas, Morphology, Constituent parses, Dependency parses, oder Named entities. Anders als im Advanced Modus, wo Benutzer*innen die

Analysewerkzeuge selbst wählen können, d.h. das Programm bietet einzelne Tools an, die beliebig strukturiert werden können (WebLicht, n.d.).

Der Annotationsprozess

Im Zuge des Annotierens, welches automatisch oder durch Benutzer*innen gesteuert werden kann, werden die Daten getaggt, d.h. die Daten werden mit morphosyntaktischen Informationen versehen. Demnach erfolgt beim Part-of-Speech (POS)-Tagging die Zuweisung der jeweiligen Wortart (Substantiv, Verb, Adjektiv, etc.) zu den einzelnen Elementen im Korpus (Lemnitzer & Zinsmeister, 2015, 63).

Jene Zuteilung der Kategorien bzw. Merkmalklassen erfordert ein auf bestimmten Richtlinien basierendes Verfahren, welches das Prinzip der Konsistenz gewährleistet, also gleichartige Phänomene nach denselben Regeln auszeichnet bzw. definiert. Auf diese Weise unterliegt der Annotationsprozess festen Regeln, wodurch wiederum die Aussagekraft der Ergebnisse garantiert wird. (Hirschmann, 2019, 22f)

Für die Annotation der Daten wird ein sogenanntes „Tagset" verwendet, welches eine Liste aller der von der Analysesoftware verwendeten Wortarten darstellt. Gängige Tagsets umfassen zwischen 50 und 150 Einzeltags (Lemnitzer & Zinsmeister, 2015, 63).

Für Korpusanalysen im deutschsprachigen Raum gilt das Stuttgart-Tübingen Tagset (STTS) als Standard-Tagset. Jenes 1999 veröffentliche Tagset ist das Resultat der Kombination zweier Tagsets, eines von der Universität Stuttgart und das andere von der Universität Tübingen stammend, welches die Annotation deutschsprachiger Textkorpora vereinfachen sollte. Diese Liste setzt sich aus insgesamt 54 Tags zusammen, 48 reine Tags und 6 zusätzliche Tags für etwa fremdsprachliches Material, Satzzeichen und Nichtwörter wie etwa XY. Diese Tags spiegeln „die Hauptwortarten und ihre Unter-kategorien […] wider. Die tags bestehen aus möglichst selbsterklärenden Buchstabensequenzen, die von links nach rechts gelesen zuerst die Hauptwortart und dann die Unterwortart kodieren, also von der allgemeinen Information zur spezifischen hinführen." (Schiller, 1999, 4)

Weitere Elemente, die neben der Wortart mittels des STTS ausgezeichnet werden, sind syntaktische Positionen, Distribution des Wortes, grammatische Funktionen sowie die Morphologie bzw. die Semantik betreffende Eigenschaften. Darüber hinaus werden auch Satz- wie Sonderzeichen als solche erkannt. (Lemnitzer & Zinsmeister, 2015, 63)

Folglich ist zwischen zweierlei Arten von Annotation zu unterscheiden: einerseits das Wortarten-Tagging (POS-tagging) und andererseits die Annotation von Flexionsmorphologie, bei der das Token analysiert und dessen Grundform, das Lemma, gebildet wird, d.h. nicht konjugierte (Infinitivform des Verbs), nicht deklinierte (Nominativ Singular des Nomens) und flexionslose (Adjektiv-)Form des Wortes. (Hirschmann, 2019, 35-38)

Parallel zum Tagging verläuft der Arbeitsschritt der „Tokenisierung". Darunter wird das Zerlegen des Korpus in die kleinste zählbare Einheit – „Token" – verstanden, wobei die Definition, was als kleinste zählbare Einheit gilt, vom Forschungsvorhaben abhängt. In den meisten Fällen einer Kor-pusanalyse markieren Leerzeichen die Grenzen zwischen den einzelnen Token, was bedeutet, dass jedes Wort und jedes Satzzeichen als Token verstanden wird. Ferner kann eine Tokenisierung auf Satz-, Morphem-,

Zeichen- oder Lautebene durchgeführt werden. Essenziell ist nur, dass – wie auch bei der Annotation selbst – die Tokenisierung einheitlich und das gesamte Korpus betreffend erfolgt. Abgesehen von der allgemein gültigen Definition, dass ein Wort und ein Satzzeichen Token sind und jene durch Leerzeichen voneinander getrennt sind, besteht natürlich die Möglichkeit bzw. die Notwendigkeit, weitere Regeln und Richtlinien für die linguistische Aufbereitung zu definieren, etwa wie mit Eigennamen, Abkürzungen oder auch der Spatienschreibung im Zuge der Analyse verfahren werden soll. (Hirschmann, 2019, 32-38)

Ist es nicht möglich, ein Token im Zuge des automatisierten Annotationsprozesses eindeutig zu taggen, kann dies durch eine manuell angefertigte Regel oder einen Wahrscheinlichkeitswert im Nachhinein gemacht werden. Daher wird in der Korpuslinguistik auch zwischen regel- und wahrscheinlichkeitsbasiertem Annotieren differenziert. (Hirschmann, 2019, 41)

Dies hat zur Folge, dass nach dem automatischen POS-Tagging und der Lemmatisierung ein Zwischenarbeitsschritt vonnöten ist, bevor mit der Auswertung der Ergebnisse begonnen werden kann. Somit erfolgt zusätzlich zur (halb)automatischen Annotation eine manuelle Nachbereitung. Lemnitzer & Zinsmeister (2015, 69) bilden den Prozess des Taggings folgend ab:

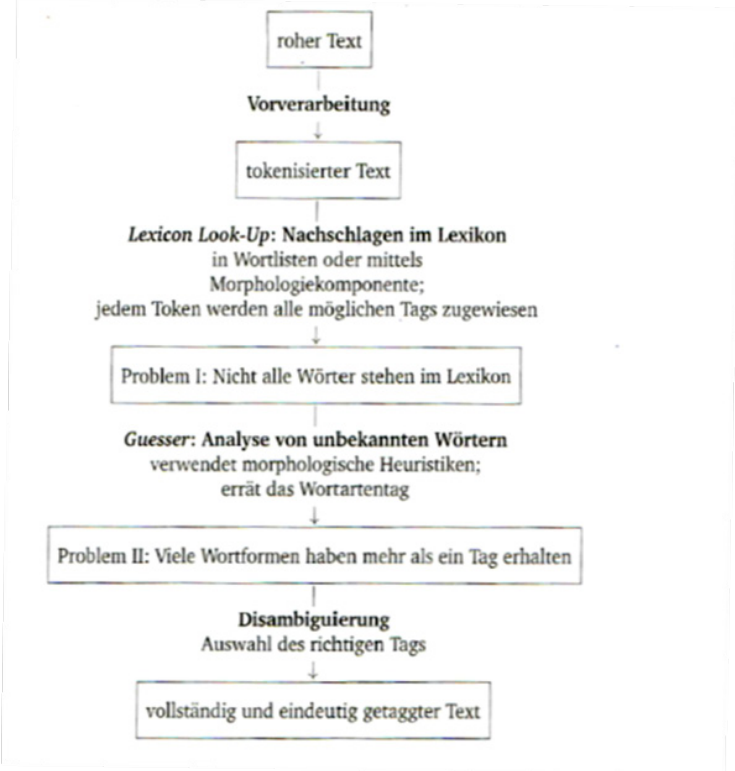


Abb. 2: POS-Tagging und Tokenvergabe (Lemnitzer & Zinsmeister, 2015, 69).

Datenauswertung

Im Folgenden werden die einzelnen Analysemöglichkeiten in WebLicht abgebildet.

Das Korpus setzt sich mit neun Einzeldateien wie folgt zusammen:

- VWA 1: 32 Seiten, Entstehungsort/-jahr: Wien, 2015
- VWA 2: 25 Seiten, Entstehungsort/-jahr: Wien, 2016
- VWA 3: 31 Seiten, Entstehungsort/-jahr: Scheibbs, 2017
- VWA 4: 31 Seiten, Entstehungsort/-jahr: Wien, 2019
- VWA 5: 42 Seiten, Entstehungsort/-jahr: Wien, 2019
- VWA 6: 32 Seiten, Entstehungsort/-jahr: Wien, 2019
- VWA 7: 44 Seiten, Entstehungsort/-jahr: Wien, 2019
- VWA 8: 32 Seiten, Entstehungsort/-jahr: Wien, 2019
- VWA 9: 40 Seiten, Entstehungsort/-jahr: Wien, 2019

Nach der Aufbereitung der Daten ergab sich eine Gesamttokenanzahl von 64.523 Token. Für die Bearbeitung wurde das Korpus in die Bearbeitungsoberfläche Chain 1 eingepflegt. Chain deshalb, weil darunter die einzelnen Tool Chains – Verarbeitungsketten – verstanden werden, welche gleichzeitig, auf verschiedenen Ebenen, das Korpus bearbeiten: “The tools take the output of the preceding tool as their input, add information to these input data in a cumulative way, and do not alter either the primary data nor the data which were added by preceding tools.” (WebLicht n.d.)

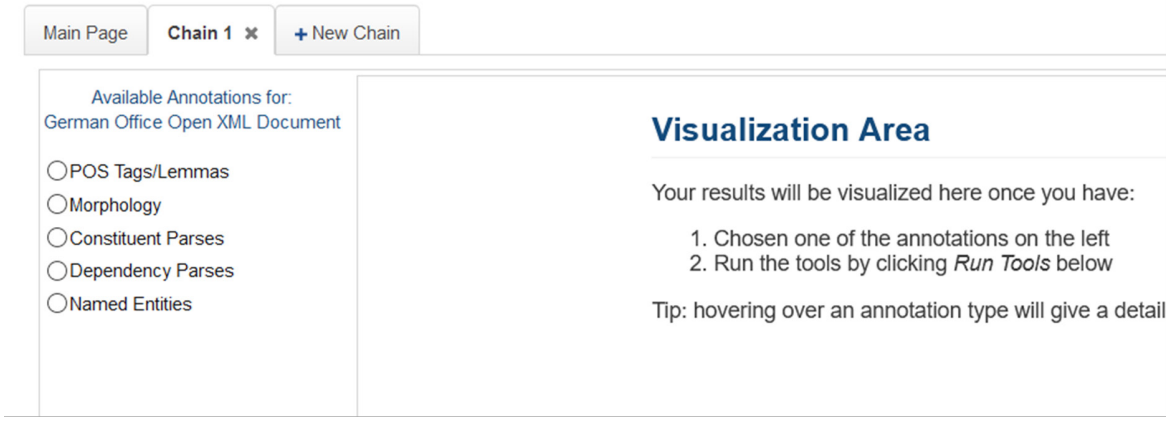


Abb. 3: Easy Modus in WebLicht. Screenshot.

Die Verarbeitungsketten sind am unteren Bildschirmrand ersichtlich. So können User*innen eigenhändig entscheiden, ob und wann welche Annotation am Korpus durchgeführt werden soll. Defaultmäßig wird der importierte Text in das zuvor erwähnte TCF umgewandelt. Die Segementierung des Korpus in einzelne Sätze und Tokens bildet den nächsten Arbeitsschritt, welchem das Tagging durch den TreeTagger folgt.

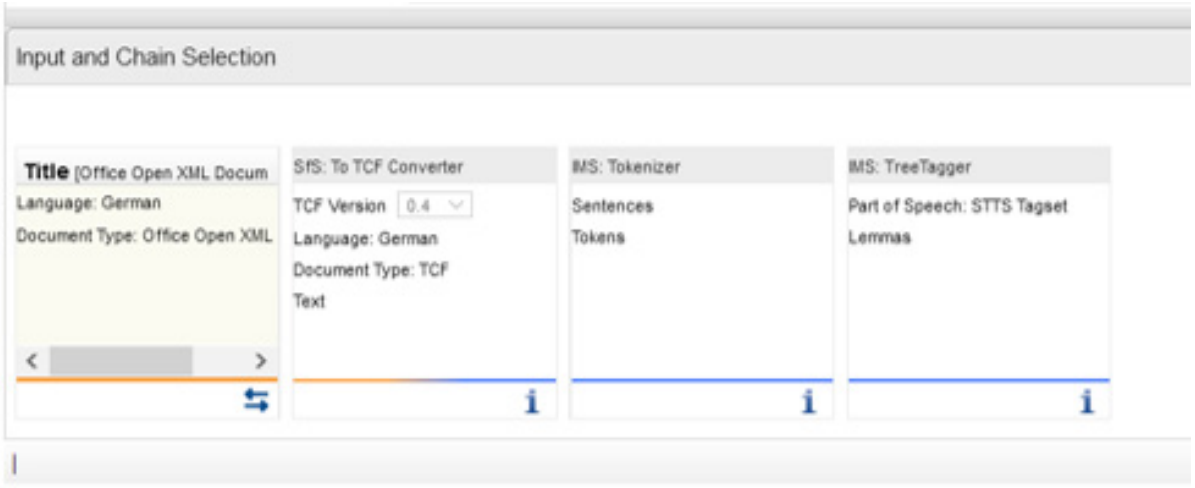
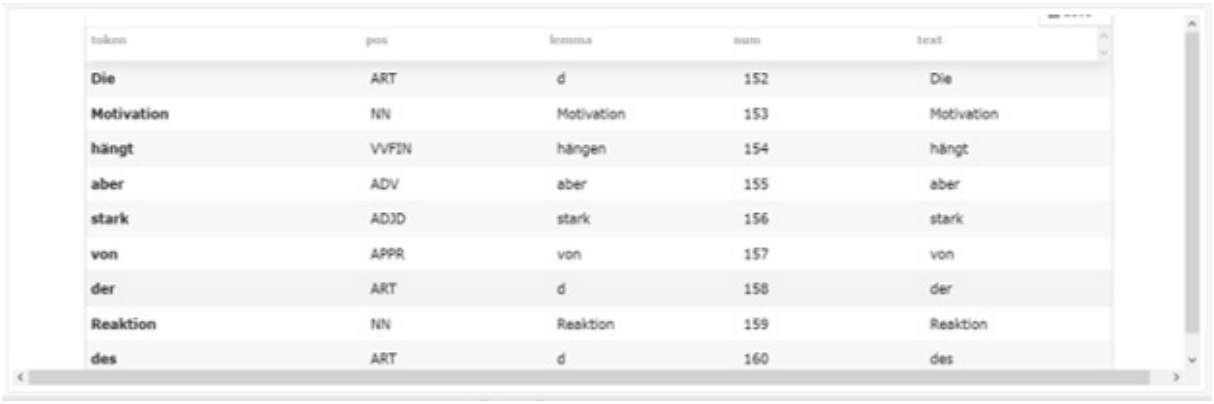


Abb. 4: Chain Selection in WebLicht (Screenshot)

Mittels des Befehls Run Tools (rechter Bildschirmrand unter der Visualization Area) beginnt Web-Licht mit der Annotation der Daten. Beim POS-Tagging werden die einzelnen Sätze hinsichtlich ihrer Tokens (Wort), POS (Wortart), Lemmata (Grundform des Wortes), Nummerierung (Position) im Korpus und dem Text selbst aufschlüsselt.



token	pos	lemma	num	text
Die	ART	d	152	Die
Motivation	NN	Motivation	153	Motivation
hängt	VVFIN	hängen	154	hängt
aber	ADV	aber	155	aber
stark	ADJD	stark	156	stark
von	APPR	von	157	von
der	ART	d	158	der
Reaktion	NN	Reaktion	159	Reaktion
des	ART	d	160	des

Abb. 5: POS-Tagging in WebLicht (Screenshot)

Äußerst relevant für die Berechnung von Korrelationen ist die Abbildung der Dependenzstruktur des Korpus, auch Dependency Phrases genannt. Dabei basiert die Satzhierarchie auf den Abhängigkeiten (Dependenzen) der einzelnen Satzglieder zueinander. Es gibt ein „Regens“ und ein davon abhängiges „Dependens“. (Lemnitzer & Zinsmeister, 2015, 73)
Zur graphischen Darstellung verwendet WebLicht die funktionale Dependenzstruktur (Pfeile), wie in der nachfolgenden Abbildung zu sehen ist. Zusätzlich zu den Pfeilen, welche „Regens“ und „Dependens“ abbilden, annotiert WebLicht die einzelnen Satzglieder: Im nachfolgenden Satzbeispiel ist wurden etwa als Satzwurzel (root) definiert, welches beispielsweise von Frauen regiert wird, dessen grammatische Funktion wiederum das Subjekt des Satzes darstellt und somit das Kürzel SUBJ erhält.

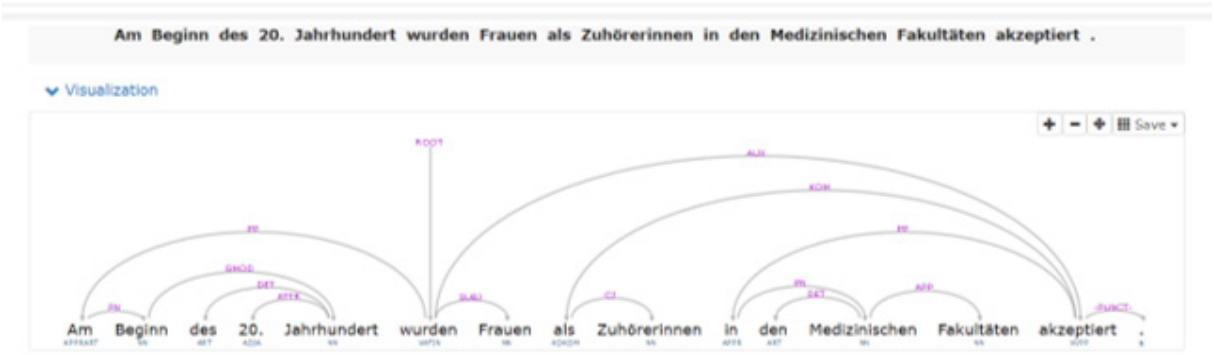


Abb. 6: Dependency Phrases in WebLicht (Screenshot)

Neben der Dependenzstruktur bietet WebLicht die Möglichkeit die Konstituentenstruktur, Constitutent Phrases, abzubilden. Die Software bedient sich dafür des Berkely Phraser Analysers. Diese linguistische Theorie, auf den amerikanischen Strukturalismus zurückgehend, definiert Sätze als hierarchisch geschachtelte Konstrukte, die aus Untereinheiten bestehen, welche Sequenzen von zusammenhängenden Wörtern ergeben. Gängige Darstellungsmöglichkeiten sind etwa die Klammerstruktur.

Ähnlich verhält es sich bei der Baumstruktur: dabei steht der sogenannte Wurzelknoten (root) an der Spitze. Von diesem aus verlaufen die Äste (Kanten bzw. edges) zu verschiedenen Knoten (node), die jeweils einem höher gelegenen Knoten zugeteilt sind. Das Wort selbst wird als terminaler Knoten bezeichnet. Darüber hinaus können in einem Korpus die Phrasen (Nominalphrase, Verbalphrase, etc.) oder Chunks (prosodische Einheiten) abgebildet werden. (Lemnitzer & Zinsmeister, 2015, 71-75)

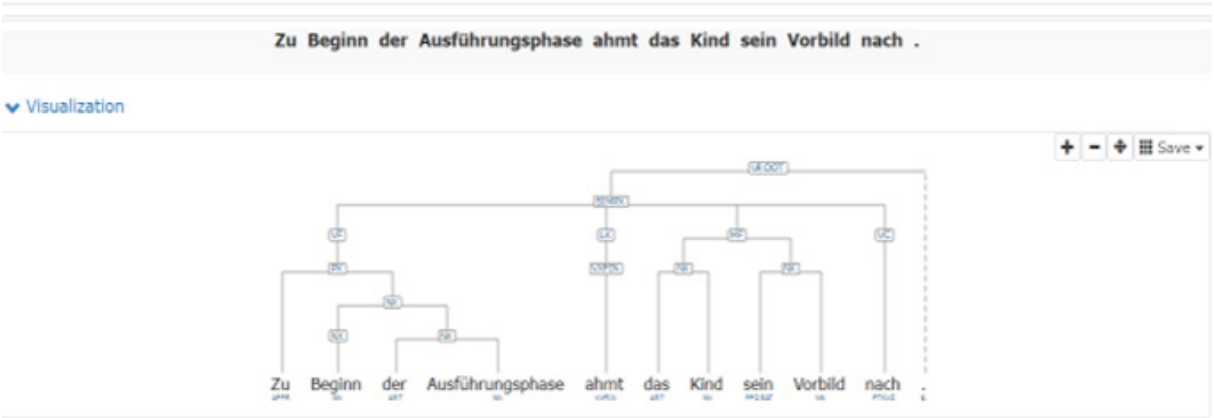


Abb. 7: Constituent Phrases in WebLicht (Screenshot)

Eine vierte Darstellungsmöglichkeit ist die Aufsplittung des Korpus in seine morphologischen Einheiten. Im Zuge jener Analyse werden den einzelnen Token Kategorien wie etwa Kasus, Genus, Nume-rus, Person, Tempus und Modus zugeordnet. Kann einem Token ein zweideutiger morphologischer Wert zugeordnet werden, so vergibt das STTS einen Asterisk. (Lemnitzer & Zinsmeister, 2015, 68)

Sidra	nominative	proper na...			singular			43002	Sidra
ist		verb		true	singular	3	present	43003	ist
jetzt		adverb						43004	jetzt
neun		cardinal						43005	neun
Jahre	accusative	regular no...			plural			43006	Jahre
alt		adjective						43007	alt
und		conjunction	true					43008	und
ist		verb		true	singular	3	present	43009	ist
körperlich	nominative	regular no...	masculine		singular			43010	körpe
altersentsprech...		adjective						43011	alters

Abb. 8: Morphology in WebLicht (Screenshot)

Nach erfolgter Bearbeitung kann das annotierte Korpus mit dem Befehl Download Chain heruntergeladen werden, etwa als XML- oder CSV-Datei. Ohne weitere Probleme soll es möglich sein, die annotierte Datei in andere Programme einzuspeisen bzw. in andere Formate zu konvertieren.

Einspielen des mit WebLicht annotierten Korpus in die Open-Source-Statistiksoftware RStudio

Dieser Annahme wurde im nächsten Schritt nachgegangen, als versucht wurde, die aus WebLicht heruntergeladene Datei in die Software RStudio einzupflegen und mittels einzelner Befehle der Programmiersprache R weiterzubearbeiten. Es wollte festgestellt werden, ob ein reibungsloses Einspielen und Weiterbearbeiten der Datei möglich ist, oder ob es zu Problemen kommt und wenn ja, wie diese aussehen.

RStudio ist „an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debug-ging and workspace management.“ (RStudio, n.d.)

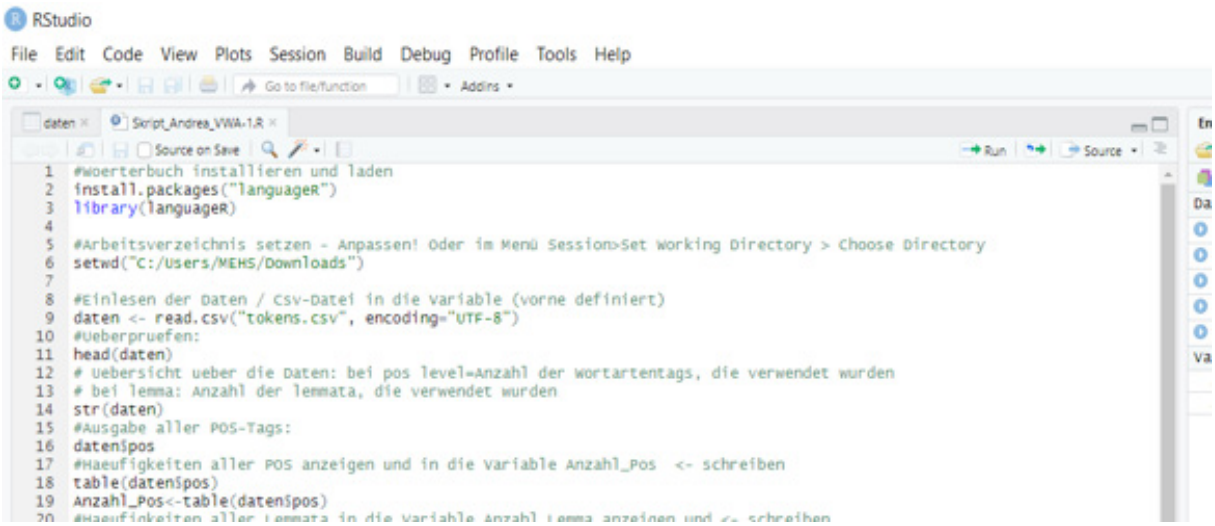


Abb. 9: RStudio Desktop Surface Structure (Screenshot)

Anders als bei WebLicht musste die Software lokal auf dem Rechner installiert werden. Erst danach konnte die Datei aus WebLicht importiert und in R bearbeitet werden. Einfache Befehle wie die Abfrage der Anzahl der getaggten Wortarten bzw. der Lemmata oder die Abbildung der Häufigkeit aller POS und Lemmata sowie die prozentuellen Angaben zu POS und Lemmata sind problemlos möglich.
Wobei an dieser Stelle darauf verwiesen werden sollte, dass ein grundlegendes Verständnis von RSTudio gegeben sein sollte, d.h. ein Einlesen bzw. Einarbeiten in die Software notwendig ist. Besonders für die Darstellung von Grafiken (Plots), welche nach der sogenannten „Grammatik von Grafiken" realisiert werden. (Wickham, n.d.)
Nachfolgende Grafik wurde direkt in RStudio erstellt und zeigt etwa die die Häufigkeiten der einzelnen Tags.

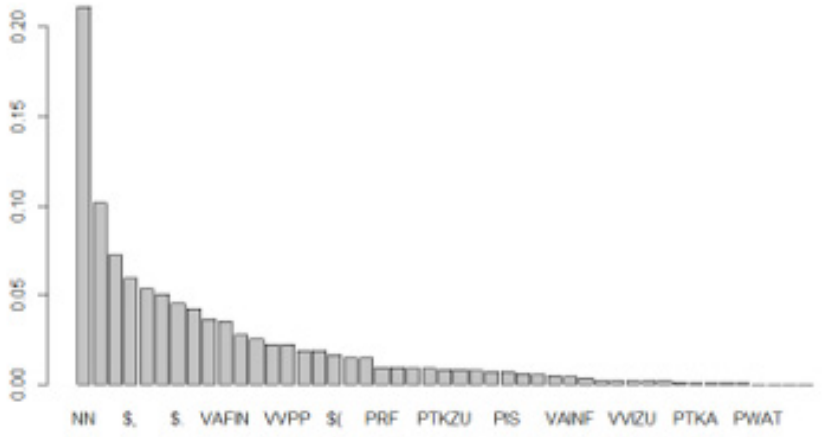


Abb. 10: Häufigkeiten (Tags), eigene Abbildung aus dem beschriebenen Datensatz, erstellt in RStudio

Wenig überraschend ist, dass „normale Nomen" (NN), gefolgt von Satzzeichen „Komma" (S,) und „satzbeendende Funktion" (S.) die drei häufigsten Tags darstellen und an vierter wie fünfter Stelle das „finite Voll- oder Kopularverb" (VAFIN) und das „partizipiale Vollverb (Partizip II)" (VVPP) ste-hen.
Im Anschluss daran wurden die XML-Dateien, Lemmata und POS-Tags gesamt aus RStudio in CSV-Dateien umgewandelt, etwa zur weiteren Bearbeitung mit Excel. Diese sieht für die POS-Tags im Korpus der VWA hinsichtlich ihrer Häufigkeit so aus:

1	NN (normales Nomen)	13635
2	ART (Artikel)	6546
3	APPR (Präposition	4662
4	$, (Komma)	3838
5	ADJA (attributives Adjektiv)	3440
6	ADV (Adverb)	3274
7	$. (satzbeendende Interpunktion)	2923
8	VVFIN (finites Vollverb)	2732
9	KON (nebenordnende Konjunktion	2360
10	VAFIN (finites Voll- oder Kopulaverb)	2270

Die Tabelle zeigt, welche zehn Tags in welcher Häufigkeit im Korpus vorkommen. Es ist nicht verwunderlich, dass Nomen und Artikel erst- und zweitgereiht sind, gefolgt von Präpositionen und dem Satzzeichen „Beistrich".

Der Import des in WebLicht bearbeiteten VWA-Korpus in RStudio sowie die weiterführenden Arbeitsschritte gefolgt vom Export und der Konvertierung in Excel waren problemlos möglich. Das edeutet, beide Programme sowie deren Auswertungen miteinander kompatibel sind, was für Projekte vorteilhaft ist, wo nicht nur mit einem einzigen Softwareprogramm gearbeitet wird oder wo es notwendig ist, Ergebnisse so kompatibel als auch nachhaltig wie möglich zu gestalten.

Resümee

Bereits vor der hier durchgeführten Analyse am VWA-Korpus war das allgemeine Verständnis gegeben, dass jene technischen Konzepte und Methoden, die im Forschungsprojekt von Meißner und Wallner verwendet wurden, problemlos auf andere, weniger umfangreiche Textkorpora angewandt werden können, zumal korpusbasierte Annotationsprogramme unabhängig des verwendeten Korpus nach einheitlichen Schemata vorgehen.

Nachdem es nun nicht primär um die Frage ging, wie denn ein solches Projekt strukturiert sein, welche Forschungsfrage(n) darin untersucht werden oder in welchem Kontext die Ergebnisse behandelt werden könnten, stellte sich die Frage, welche Software man für ein solches Projekt verwenden könne, insbesondere im Kontext einer studentische Auseinandersetzung mit Korpora.

Zwei entscheidende Punkte, die zur Verwendung von WebLicht geführt haben, sind, dass es sich um ein gängiges Programm für Korpusanalysen handelt und bereits aus dem Studienkontext bekannt war und dass es zweitens eine Open-Source-Software ist, was gerade für studentische Arbei-ten von Relevanz ist/sein könnte, da es unter Umständen nicht möglich ist Lizenzen für bestimmte Softwareprogramme zu erwerben.

Neben der einfachen Handhabung und der detaillierten Benutzerbeschreibung von WebLicht selbst überzeugte während dieser ersten Auseinandersetzung auch die einladende und userfreundliche Benutzeroberfläche, welche – auch ohne eine vertiefende Befassung mit der Handhabung – zumeist selbsterklärend ist. Selbst die Bearbeitung von Korpora im advanced Modus sollte für Erstbenutzer*innen keine allzu große Schwierigkeit darstellen, zumal auch hier ausführliche Erklärungen online zu finden sind.

Die diversen Analysemöglichkeiten hinsichtlich der verschiedenen linguistischen Schemata und die schichtartige Vorgehensweise, dass Einzelschritte durchgeführt werden können je nachdem welcher Aspekt beleuchtet und erörtert werden möchte, machen WebLicht zudem attraktiv. Nicht nur die unterschiedlichen Darstellungsoptionen, sondern auch die Komptabilität mit anderen Programmen, wie in Verbindung mit RStudio und Excel, bestärken das webbasierte Tool in seiner Um-setzung.

Unter dem Blickwinkel einer ersten Beschäftigung mit korpuslinguistischen Analysen, den damit einhergehenden methodischen wie strukturellen Überlegungen sowie den für die Datenbearbeitung und -auswertung relevanten technischen Gegebenheiten berücksichtigend, ist zu sagen, dass WebLicht einen einfachen Einstieg mit Option auf Vertiefung und Erweiterung der persönlichen

Anwendungskompetenzen garantiert. Demnach ist es nicht nur spannend, Forschungsprojekte zu konzipieren und diese mit der jeweiligen Software zu bearbeiten. Ebenso besteht die Möglichkeit die Software und die Art und Weise der Durchführung in den Mittelpunkt der Aufmerksamkeit zu stellen. Der Ansatz dieses Artikels, VWA zu verwenden, könnte ausgebaut werden, d.h. ein größeres Korpus erstellt und hinsichtlich bestimmter Aspekte untersucht werden. Ebenso wäre ein Vergleich einzelner Programme zur Korpusannotation, -analyse und -auswertung denkbar, besonders unter den Kriterien von Zugänglichkeit, Nachhaltigkeit, Komptabilität, Benutzerfreundlichkeit oder auch Relevanz im aktuellen Forschungskontext der Computerlinguistik. Unabhängig der spezifischen Forschungsfragen und der bevorzugt verwendeten Software ist klar, dass Computerlinguistik, in welcher Form auch immer, zukünftig an Bedeutung zunehmen wird.

Literatur

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin; New York: W. de Gruyter.

CLARIN (n.d.). *CLARIN in a nutshell*. Abgerufen am 17. Februar 2020 von https://www.clarin.eu/content/clarin-in-a-nutshell.

CLARIN-D AP5 (2012) *CLARIN-D User Guide. Weblicht – A service-oriented architecture for linguistic resources and tools Chapter 8.* Abgerufen am 24. September 2020 von https://media.dwds.de/clarin/userguide/text/web_services_WebLicht.xhtml.

CLARIN-D mit Schmidt, T., K. Wörner, T. Lehmberg, & H. Hedeland. *EXMARaLDA*. Abgerufen am 16. März 2020 https://exmaralda.org/de/ueber-exmaralda.

Deml, I. (2015). *Gebrauchsnormen der Wissenschaftssprache und ihre Entwicklung vom 18. bis zum 21. Jahrhundert.* Dissertation. Universität Regensburg.

Ehrlich, K. (1999). Alltägliche Wissenschaftssprache. *Informationen Deutsch als Fremdsprache*, 26, 1. 3-24.

Graefen, G. (1997). *Der wissenschaftliche Artikel – Textart und Textorganisation*. Berlin [u.a.]: Peter Lang. Abgerufen am 06. September 2020 von https://www.daf.uni-muenchen.de/media/downloads/wiss_artikel_a4.pdf

Graefen, G., & M. Moll (2011). *Wissenschaftssprache Deutsch.* lesen – verstehen – schreiben. Ein Lehr- und Arbeitsbuch. Frankfurt am Main, Wien [u.a.]: Lang.

Groebner, V. (2012). *Wissenschaftssprache. Eine Gebrauchsanweisung.* Konstanz: University Press.

Hirschmann, H. (2019). *Korpuslinguistik. Eine Einführung.* Stuttgart: J.B. Metzler.

Hirschmann, H. (n.d.) *STSS-Tagset gemäß Tiger Annotationsschema.* Institut für deutsche Sprache und Linguistik. Humboldt-Universität zu Berlin. Abgerufen am 17. Februar 2020 von https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger.

Karmasin, M., & R. Ribing (2018). *Die vorwissenschaftliche Arbeit von A bis Z.* Wien: Facultas.

Lemnitzer, L., & H. Zinsmeister (2015). *Korpuslinguistik: eine Einführung*. 3., überarb. und erw. Aufl. Tübingen: Narr.

Lenz, S. (2000). *Korpuslinguistik.* Tübingen: Narr.

Meißner, C. (2014). *Figurative Verben in der allgemeinen Wissenschaftssprache des Deutschen. Eine Korpusstudie.* Tübingen: Stauffenburg-Verlag.

Meißner, C., & F. Wallner (2019) *Das gemeinsame sprachliche Inventar der Geisteswissenschaften. Lexikalische Grundlangen für die wissenschaftspropädeutische Sprachvermittlung.* Berlin: Erich Schmidt Verlag.

Moll, M., & W. Thielmann (2017). *Wissenschaftliches Deutsch: Wie es geht und worauf es dabei ankommt.* Konstanz: UVK.

RStudio (n.d.). *Daten veranschaulichen mit ggplot2. Schummelzettel.* Abgerufen am 24. Oktober 2020 von https://rstudio.com/wp-content/uploads/2015/06/ggplot2-german.pdf.

Schiller, A., S. Teufel, C. Stöckert, & C. Thielen (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technischer Bericht, Universitäten Stuttgart und Tübingen. Abgerufen am 12. März 2020 von https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/TagSets/stts-1999.pdf

The Berkeley NLP Group (n.d.). Abgerufen am 17. Februar 2020 von http://nlp.cs.berkeley.edu/software.shtml.

Software

RStudio v1.2. https://rstudio.com/products/rstudio/. Download am 16.02.2020

H. Wickham (n.d.). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. https://ggplot2.tidyverse.org/authors.html

Schmid, H. (n.d.) *Treetagger.* LMU München. Abgerufen am 16. Februar 2020 von https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

CLARIN (n.d.). *WebLicht 6.1.1.* Abgerufen am 29. September 2020 von https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

Abbildungsverzeichnis