Spielregeln

- Eigenes Mikrofon: **AUS**
- Eigenes Video: **AN** oder **AUS**
- Buttons:

Alles ok | Halt, ich komme nicht mit! | langsamer | schneller

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Praxislabor Digitale Geisteswissenschaften

Einführung in die Korpusanalyse (AntConc/Voyant)

Helene Schlicht
h.schlicht@ub.uni-frankfurt.de

Ablauf

- Was ist eine Korpusanalyse?
- Was ist ein Textkorpus?
- Wie müssen Texte aufbereitet werden?
- Woher erhalte ich Texte?
- Die Methoden
  - Wortliste/Worthäufigkeiten (Voyant: Cirrus)
  - KWIC (Keyword in Kontext) (Voyant: Kontexte)
  - Concordance Plot (Voyant: Bubblelines)
- Die Tools
  - AntConc
  - Voyant

Was ist eine Korpusanalyse

Eine Korpusanalyse ist ein Verfahren der empirisch arbeitenden Sprachwissenschaft mit dem Ziel, einen konkret umrissenen Sprachgebrauch methodisch abgesichert zu beschreiben und Hypothesen bezüglich der Merkmale dieser Sprache zu bilden oder zu prüfen.

Die digital gestützte Korpusanalyse, auch als Text Mining bezeichnet, ermöglicht die **maschinelle Sprachverarbeitung und -auswertung** großer Mengen an Textdaten. Diese werden **mit statistischen und linguistischen Mitteln** miteinander verglichen und nach verborgenen Bedeutungsstrukturen untersucht. Beim Text Mining handelt es sich um **softwarebasierte** Methoden, die quantitative aber auch qualitative Verfahren anwenden. Nötig sind hierfür Textdaten in einem bearbeitbaren Format. Diese werden vor der Analyse strukturiert, um ihre Erschließung zu ermöglichen.

Was ist ein Textkorpus?

Sammlung von schriftlichen Texten oder textlich aufgezeichneten mündlichen Äußerungen einer bestimmten Sprache und/oder Textgattung.

Für die digitale Korpusanalyse:
Sammlung an digital vorliegenden Textdaten, die für die Analyse aufbereitet werden müssen.
Häufig zusammengestellt mit einer bestimmten Forschungsfrage im Hinterkopf.

Verschiedene Korpustypen

Referenzkorpus

Ein Referenzkorpus hat einen festen Inhalt und ist häufig öffentlich verfügbar. Referenzkorpora sollen **umfassende Informationen über eine Sprache** (zu einer bestimmten Zeit) liefern und diese umfassend abdecken. Daher können sie als Referenz herangezogen werden.

Monitor corpora

Ein monitor corpus ist ein Korpus, das mit der Zeit immer weiter anwächst und womit sich bspw. **Sprachgebrauch und sprachliche Veränderungen über die Z**eit verfolgen und abbilden lassen.

Verschiedene Korpustypen

Balanced corpus

Ein balanced corpus, auch als sample corpus bezeichnet, versucht eine bestimmte Art von Sprache zu einer bestimmten Zeit zu repräsentieren. Hier ist Repräsentativität das Ziel. Ein sample corpus nimmt eine Vielzahl von Textgattungen in die Stichprobe auf, um ein Maximum an Ausgewogenheit und Repräsentativität zu erreichen.

Thematische Korpora/Spezialkorpora

Viele Korpora fallen in keine der genannten Kategorien und folgen keiner rigorosen Systematik. Oftmals repräsentieren sie „nur“ die Daten, die zu einer bestimmten Zeit für eine bestimmte Forschungsfrage verfügbar waren.

Mögliche Forschungsfragen

Die Korpusanalyse erlaubt es uns, Muster in großen Textmengen zu erkennen, die uns beim Lesen nicht unbedingt auffallen würden. Dazu gehören:

- Häufig wiederkehrende Sprachmuster
- Typische Sprachgebrauchsmuster einer*s bestimmten Autor*in
  - Autorschaftsbestimmung (Bsp. Shakespeare)
- Grammatische Verwendungsmuster
- Sprachliche Varianz bestimmter Autoren (Bsp. Rap-Musik)

Mögliche Forschungsfragen

- Welche Begriffe werden im vorliegenden Korpus am häufigsten verwendet?
- In welchem Kontext kommen ausgewählte Wörter vor?
- Auf welche Weise ballen sich Begriffe in einer Textsammlung zusammen?
- Finden sich sprachliche Muster in allen Texten des Korpus wieder?

Eine Korpusanalyse ist besonders nützlich, um Intuitionen und Hypothesen über Texte zu entwickeln oder zu testen.

Schritte

- Datenmaterial auswählen
- Datenmaterial für die Analyse aufbereiten
- Datenmaterial analysieren und auswerten

Wie müssen Texte aufbereitet werden?

Die Textdaten, die analysiert werden sollen, können bzw. müssen im Vorfeld mehr oder weniger stark strukturiert und aufbereitet werden, um sinnvoll ausgewertet werden zu können. **Voyant** und **AntConc** arbeiten mit **schwachstrukturierten Textdaten** und sind daher gut für den Einstieg geeignet.

Aufbereitung schwachstrukturierter Texte:

- Geforderte Dateiformate?
- Texte bereinigen: nur inhaltlich relevantes sollte analysiert werden.

Wie müssen Texte aufbereitet werden?

Geforderte Dateiformate:

- AntConc: txt, html, xml (meiner Erfahrung nach funktioniert txt am besten)
- Voyant: u.a. txt, docx, xlsx, csv, Import von URLs mittels Textbox

Texte bereinigen:

- Titel
- Inhaltsverzeichnis
- Überschriften (bei manchen Werken auf jeder Seite)
- Seitenzahlen
- Bei bspw. Theaterstücken: Bühnenanweisungen und Figurenzuordnung
- Ggfls. Normalisierung (Satzzeichen entfernen, Sonderzeichen umwandeln etc.)

Wie müssen Texte aufbereitet werden?

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Ausblick: Starkstrukturierte Texte

Texte aufbereiten:

- Tokenisierung (Satz- und Wortgrenzen identifizieren)
- Part-of-Speech-Tagging (Bestimmung der Wortart einzelner Token)
- Lemmatisierung (Zurückführen auf die Grundform/den Wortstamm)
- Normalisierung (Satzzeichen entfernen, Sonderzeichen umwandeln, historische Schreibweisen überführen etc.)
- Named Entity Recognition (automatische Identifikation und Klassifikation von Eigennamen)
- Textauszeichnung mit Metadaten (XML, TEI)

Woher erhalte ich Texte?

- Deutsches Textarchiv
- Project Gutenberg
  - Projekt Gutenberg(-DE)
- LexisNexis (kostenpflichtig, über die UB zu erreichen, aber: Vorsicht, Lizenzbedingungen)
- Wikisource
- Internet Archive
  - Internet Archive eBooks and Texts
- British National Corpus
- Bibliotheken (bspw. Digitale Sammlungen der UB JCS, Oxford Text Archive)
- English Corpora
- Bundestag Open Data (Plenarprotokolle und Drucksachen)
  - Siehe dazu: Open Discourse – Analyse der Plenarprotokolle des deutschen Bundestages seit 1949

u.v.m.

Restriktionen

- Datenqualität
- Rechtliche Rahmenbedingungen: Urheberrecht, Verwertungsrechte, Lizenzrechte
- Repräsentativität der Stichprobe

Stopwords

- Die in Sprachen am häufigsten vorkommenden Worte
- Sprachabhängig
- im Deutschen: bestimmte (der, die, das) und unbestimmte (einer, eine, ein) Artikel, Konjunktionen (und, oder, aber, weil), häufige Präpositionen (an, in, von), Partikel, Hilfsverben etc.
- Nicht inhaltstragend
- sollten/können aus Korpus entfernt bzw. nicht in Analyse einbezogen werden
- Allgemeine Stoppworte einer Sprache
- Stoppworte in einer gegebenen Dokumentmenge/für einen gegebenen Zweck

- Voyant bietet eine Stoppwortliste mit den gängigsten Begriffen an, diese kann um selbst gewählte Begriffe erweitert werden
- bei AntConc muss man eine Stoppwortliste händisch hinzufügen
- Stoppwortlisten kann man für viele Sprachen Online finden

Wildcards

- Fachbegriff: Trunkierung
- Joker/Platzhalter für ein oder mehrere (auch null) andere Zeichen
- Fragezeichen (?) für genau ein Zeichen: H?cker
- Plus-Zeichen (+) für ein (oder null) Zeichen: H+cker
- Sternchen (*) für beliebig viele (auch null) Zeichen: H*cker
- Vertikaler Trennstrich (|) Suchbegriff links ODER Suchbegriff rechts
- AntConc: At-Zeichen (@) Null oder ein Wort
- AntConc: Raute (#) für genau ein Wort
- AntConc: Ampersand (&) für etwas, das kein Wort ist (Zahlen, Sonderzeichen etc.)

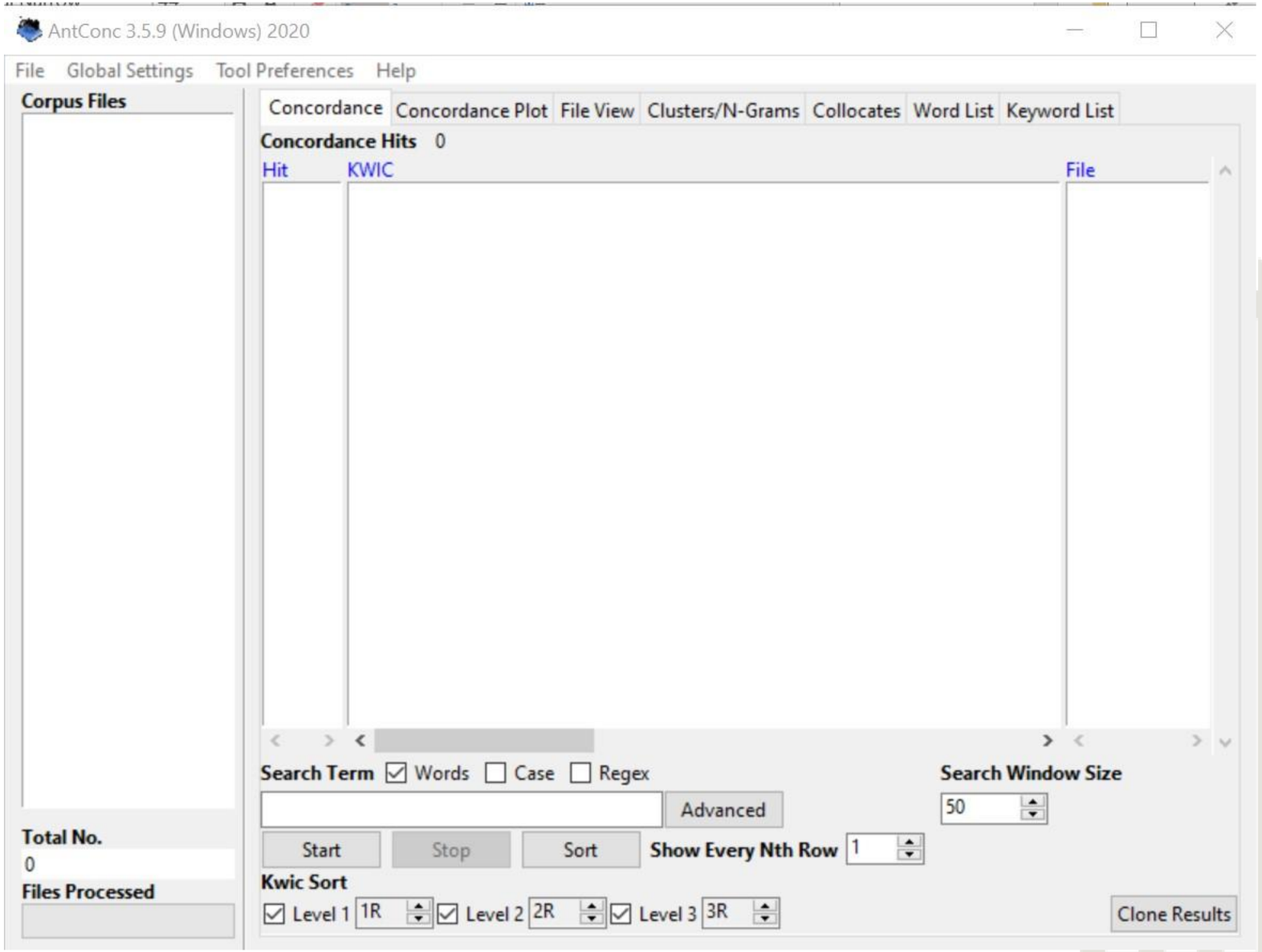
Die Methoden, mit denen wir uns beschäftigen

- Konkordanz
  - Traditionell alphabetisch geordnete Listen der wichtigsten Wörter und Phrasen in einem Werk (Register oder Index)
  - Heute i.d.R. elektronisch erstellte Trefferlisten, die die Ergebnisse der Suche nach einem bestimmten Suchbegriff in einem Korpus anzeigen, meist wird der sogenannte Kontext angeführt (vgl. KWIC)
- KWIC (Keyword in Context-Suche): zeigt den Suchbegriff und seinen, also die direkt davor oder danach stehenden Wörter
- Keyword-Plotanzeige: Anzeige der Stellen an denen ein bestimmter Suchbegriff im Korpus auftaucht
- Word List: Auflistung der im Korpus vorkommenden Wörter, nach Häufigkeit absteigend sortiert

Die Tools: AntConc

- Toolkit zur Textanalyse
- Programm zum Download: http://www.laurenceanthony.net/software/antconc/
- Kann nach dem Download direkt ausgeführt werden
- Tools sind in Reitern angeordnet, zwischen diesen kann hin und her gewechselt werden
- Entwickelt und gepflegt von Laurence Anthony, Professor für Linguistik und Leiter des Center for English Language Education in Science and Engineering an der Waseda University in Japan
- Laurence Anthony bietet auf seiner Website auch weitere Tools an
- Die Tools werden nach wie vor gepflegt und dazu werden eine Reihe von Tutorials angeboten

Die Tools: AntConc



Die Tools: Voyant

- Browserbasiertes Tool
- Unter https://voyant-tools.org/ zu erreichen (Aktuell hat die Seite Probleme!)
- Mirror unter: https://voyant-tools.huma-num.fr/
- Open Source
- Entwickelt von Stéfan Sinclair & Geoffrey Rockwell, beide im Bereich Digital Humanities tätig
- Wird nach wie vor gepflegt

Die Tools: Voyant

UNIVERSITÄTS BIBLIOTHEK FRANKFURT AM MAIN

Praxislabor Digitale Geisteswissenschaften

Einführung in die Korpusanalyse (AntConc/Voyant) Hands-on-Übung

Helene Schlicht
h.schlicht@ub.uni-frankfurt.de

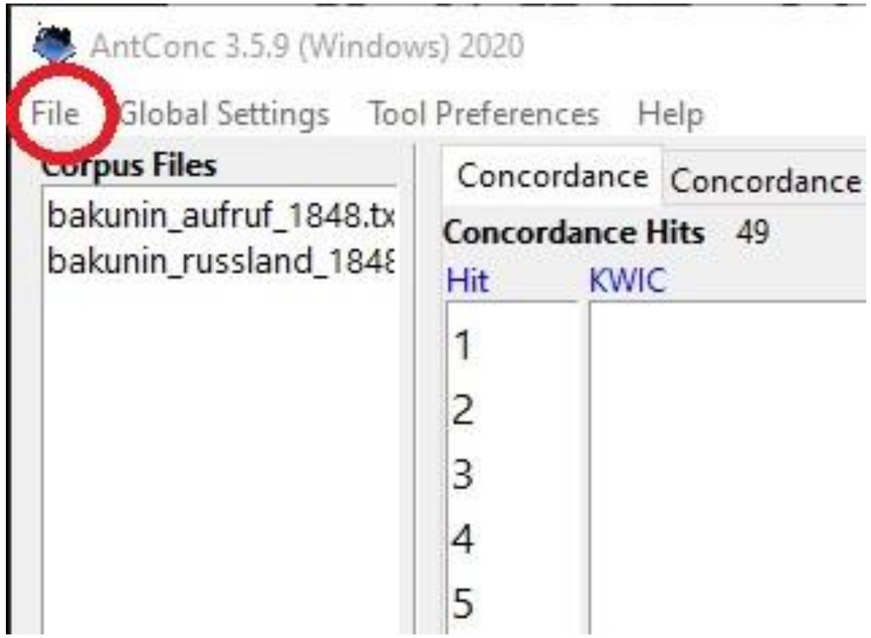
Technik-Setup

- AntConc heruntergeladen?
- Voyant-Homepage geöffnet?
- Korpus heruntergeladen?

Einen Korpus laden: AntConc

Einen Korpus laden

- AntConc starten
- Links oben auf File klicken
- Entweder Open File(s) oder Open Directory auswählen
- Zum entsprechenden Ordner navigieren und .txt auswählen



Einen Korpus laden: Voyant

- Auf Upload klicken
- Gewünschte Datei(en) auswählen
- Voyant führt nun bereits erste Operationen auf dem Korpus durch
- Sie können auch direkt Text oder eine oder mehrere URL in den Suchschlitz kopieren

Word List: AntConc

- Navigieren Sie zum Reiter Word List
- Klicken Sie auf Start ohne einen Suchbegriff einzugeben
- AntConc generiert eine Liste der am häufigsten vorkommenden Wörter in absteigender Reihenfolge

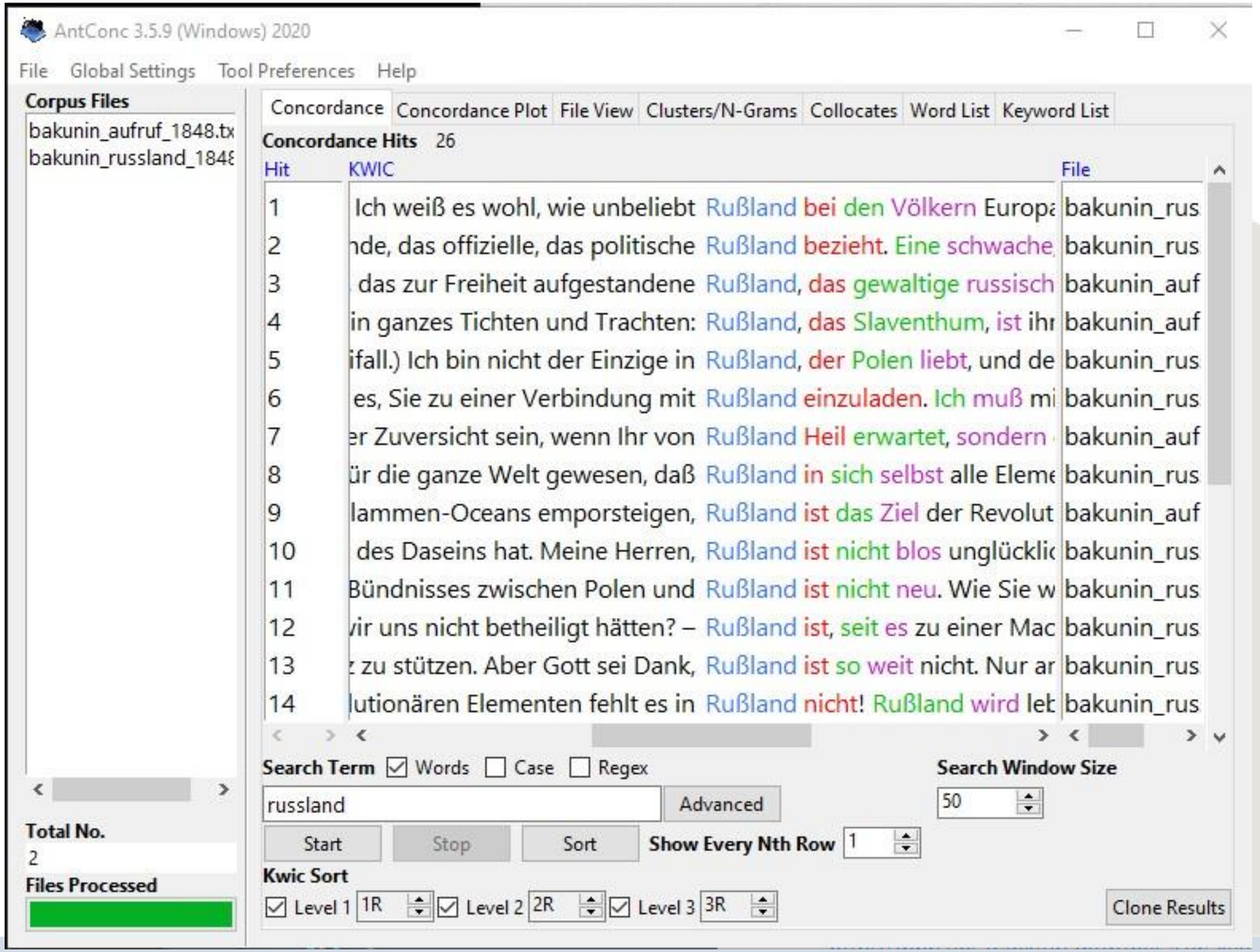
Word List: Voyant (Cirrus und Terms)

- Voyant erstellt automatisch eine Word List, wenn das Korpus geladen wird, diese wird als Wortwolke dargestellt
- man kann sich diese als Liste anzeigen lassen, wenn man zum Tool Terms wechselt

KWIC: AntConc (Concordance)

- Der Suchbegriff wird in den Suchschlitz eingetragen
- Im linken Fenster findet sich Nummerierung und Anzahl der Ergebnisse
- In der Mitte findet sich der Suchbegriff in blau mit dem ihn umgebenden Kontext
- Im rechten Fenster wird angezeigt, in welchem Dokument der Suchbegriff gefunden wurde
- Über dem Suchschlitz lässt sich festlegen, ob nach Wörtern gesucht wird, ob diese case sensitive gesucht werden oder ob eine Suche mittels Regulärer Ausdrücke stattfinden soll
- Unter dem Suchschlitz kann man bestimmen, welche Begriffe rund um den Suchbegriff gehighlighted werden
- Level 1 ist das Level, nach dem die Begriffe gerankt werden

KWIC: AntConc

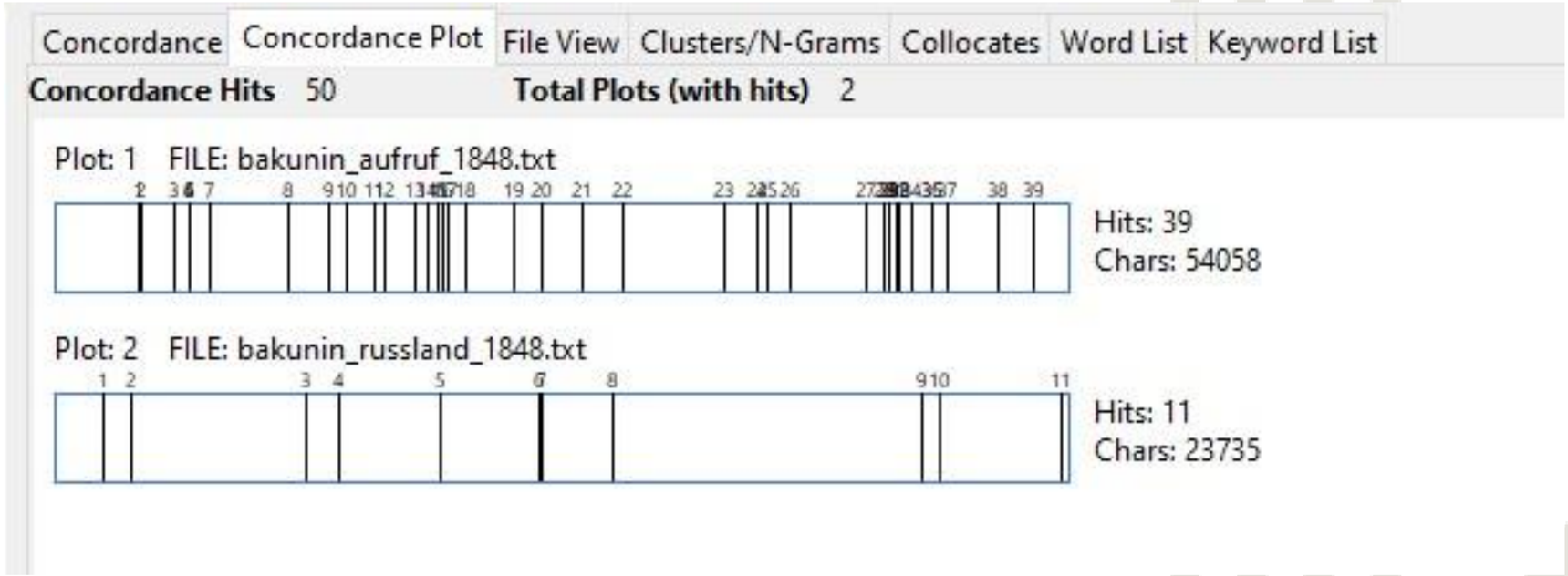


KWIC Voyant (Contexts)

- Voyant führt automatisch eine erste KWIC-Suche für den am häufigsten vorkommenden Begriff des Korpus durch
- Im Suchschlitz kann ein neuer Begriff eingegeben werden

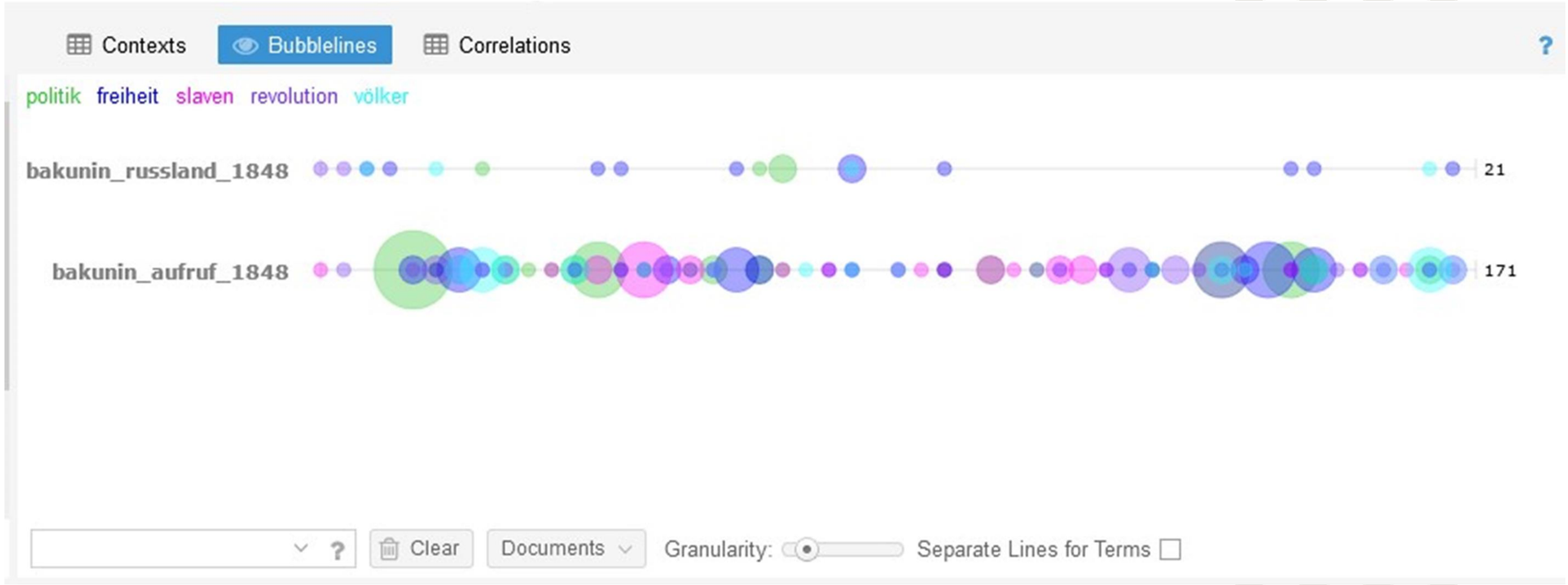
Keyword-Plotanzeige: AntConc (Concordance Plot

- Anzeige der Stellen an denen ein bestimmter Suchbegriff im Korpus auftaucht, bei AntConc visualisiert im „Barcode“-Format.
- Klickt man einen bestimmte Stelle an, öffnet sich diese im File View
- Links neben den Balken sieht man, wie viele Treffer der Suchbegriff im jeweiligen Dokument hat und darunter wie viele Zeichen das Dokument insgesamt enthält



Keyword-Plotanzeige: Voyant (Bubblelines)

- Voyant plottet automatisch die fünf am häufigsten vorkommenden Begriffe, diese erden in unterschiedlichen Farben dargestellt
- Im Suchschlitz lassen sich neue Begriffe hinzufügen
- Clear löscht alle Begriffe
- Begriffe lassen sich auch einzeln entfernen



Übung 1

Suchen Sie sich einen oder mehrere Texte im Deutschen Textarchiv (oder einem anderen Korpus) den oder die Sie untersuchen möchten und überlegen Sie eine Forschungsfrage von der Sie glauben, sie mit Voyant oder AntConc beantworten zu können.

1) Entfernen Sie unnötige Informationen aus dem Text.

2)Nutzen Sie die Word List oder Cirrus, um sich inspirieren zu lassen und neue Suchbegriffe zu entwickeln.

3) AntConc: laden Sie eine Stoppwortliste

4) AntConc: generieren Sie eine Word List: Überprüfen Sie, ob diese weitere nicht-inhaltstragende Begriffe enthält und ob Sie diese in die Stoppwortliste aufnehmen wollen.

5)Experimentieren Sie mit den verschiedenen Features und Tools auf Voyant und/oder AntConc, um zu sehen, was sie über den Text herausfinden können.

Übung 2

- Welche Begriffe werden im vorliegenden Korpus am häufigsten verwendet?
- In welchem Kontext kommen ausgewählte Wörter vor?
- Auf welche Weise ballen sich Begriffe in einer Textsammlung zusammen?
- Finden sich sprachliche Muster in allen Texten des Korpus wieder?

Welche Frage lässt sich mit welcher Methode beantworten?

Rekapitulation

Wie war Ihre Erfahrung mit den Tools? Entsprechen die Ergebnisse dem, was sie erwartet hätten? Haben Sie in Ihren Texten Ergebnisse gesehen, die Sie nicht erwartet hätten?

Welche der Tools fanden Sie am hilfreichsten und warum?

UNIVERSITÄTS
BIBLIOTHEK
FRANKFURT AM MAIN

Vielen Dank für Ihre Aufmerksamkeit!