| stylo ah online - Handbuch                                       | 2  |
|--|----|
| Ursprung: stylo ah online - stylo ohne Installation              | 2  |
| Links zum Thema  | 2  |
| Implementierte Analyse-Pipeline(s)                               | 2  |
| Funktionsüberblick   | 2  |
| Benutzung  | 3  |
| GUI Konzept  | 3  |
| Überblick zur Benutzung  | 4  |
| 1. Konfiguration des Browsers                                    | 4  |
| 2. Konfigurieren der Analyseschritte                             | 4  |
| 2.1 Arbeitsschritt: Benennung                                    | 4  |
| 2.2 Arbeitsschritt: Configurations-Datei / Datenbank             | 5  |
| 2.3 Arbeitsschritt: Input / Replication                          | 5  |
| 2.4 Arbeitsschritt: Normalisierung (Normalization)               | 6  |
| 2.4.1 Word masking / stop words                                  | 7  |
| 2.4.2 Sign equalization  | 7  |
| 2.4.3 Markup / Format  | 8  |
| 2.4.4 Word level conversions                                     | 8  |
| 2.4.5 Combinations   | 8  |
| 2.4.6 Translitteration   | 8  |
| 2.5 Arbeitsschritt: Zerlegung (features / decomposition / token) | 8  |
| 2.6 Arbeitsschritt: Zählung (Selection / Counting)               | 9  |
| 2.6.1 Most frequent token / words (per text)                     | 10 |
| 2.6.2 Culling  | 10 |
| 2.7 Arbeitsschritt: Maßanwendung (Measure selection)             | 10 |
| 2.8 Arbeitsschritt: Gruppierung (Clustering)                     | 10 |
| 2.9 Arbeitsschritt: Export                                       | 11 |
| 3. Corpus-Auswahl  | 12 |
| 4. Erneute Berechnung und Serien                                 | 12 |

# stylo ah online - Handbuch

# Ursprung: stylo ah online - stylo ohne Installation

Im Rahmen unsere Lehrtätigkeit und unserer Forschungstätigkeit haben wir das R-Paket **stylo** eingesetzt. Wir haben dieses mittels R und Python weiterentwickelt. Vornehmlich vier Dinge sind an der R Implementierung schwierig:

- 1. Die Möglichkeit zur Selbstdokumentation der Experimente ist gering.
- 2. Die Interaktion zwischen Python und R ist problematisch.
- 3. Die Möglichkeit eines einfachen Einsatzes von Multiprocessing ist fraglich.
- 4. Die Software ist teilweise schwierig zu installieren.

Daher wollte wir die Funktionalität von **stylo** nachempfinden; ohne Installation und mit erweiterten Dokumentations- und Vergleichsmöglichkeiten ausstatten.

## Links zum Thema

- 1. stylo das Original in R geschrieben.
- 2. stylo ah die R / Python Kopie mit zusätzlichen Funktionen für klassische Texte.
- 3. stylo ah online die originale Kopie in JavaScript implementiert.

# Implementierte Analyse-Pipeline(s)

Die Implementierte Textanalyse verläuft in sieben Schnitten. Diese stellen die Abfolge bzw. die Pipeline der Verarbeitung dar:

- 1. Auswahl der Files (Corpus) (jedes File repräsentiert einen Text)
- 2. Normalisierung (Formatanpassung, Zeichenvereinheitlichung, Löschung von Struktur und Metadaten, Maskierung von Wortformen)
- 3. Zerlegung in Token (Wortformen, Silben, grame, Zeichen)
- 4. Zählung / Vektorbildung (01-Kodierung, abs. / rel. Häufigkeit, TF-IDF)
- 5. Auswahl aus den Vektoren (Beschränkung der Häufigkeitsliste, Häufigkeitsfenster, Culling)
- 6. Anwendung eines Maßes
- 7. Anwendung einer Cluster-Methode

# Funktionsüberblick

In stylo ah online sind folgende Funktionen verfügbar:

#### Normalisierung:

<u>Normalform</u> Wortmaskierung (Stopworte) <u>Vereinheitichen bestimmter Zeichen</u> UV-Angleich <u>JI-</u> <u>Angleich</u> Markup-Löschen <u>Interpunktion löschen</u> Zeilenumbrüche löschen <u>Elision auflösen</u> Alpha privativum behandeln <u>Entfernen der Nummerierung</u> Worttrennungen zusammenführen <u>Vereinheitlichung des lota subscriptum und lota adscriptum</u> Veränderung des Abschluss-Sigmas <u>Diakritische Zeichen löschen Ligaturen auflösen Kleinschreibung</u> Klammern entfernen

Normalisierung

#### Zerlegung (Token):

Zusätzlich ohne Konsonanten Zusätzlich ohne Vokale Zusätzlich lediglich kleine Wörter Zusätzlich lediglich große Wörter Zerlegung in Wortformen Zerlegung in Buchstaben n-grame Zerlegung in Buchstaben n-grame der Worformen Zerlegung in Wort n-grame Zerlegung in n-grame mit Lücken Zerlegung in Pseudo-Silben Zerlegung in Kopf Körper und Coda Zerlegungen in alle Permutationen von Kopf Körper und Coda

Zerlegung

#### Zählung:

absolute Häufigkeit relative Häufigkeit 0-1-Codierung TF-IDF Beschränkung/Spannen der Frequenzlisten (nach Rang, min-max-Angabe) Culling

Zählung

#### Maße:

euclidean, chebyshev, minkowski, manhatten, canberra, soerensen, gower, soergel, lorentzian, intersection, wavehedges, motyka, ruzicka, tanimoto, innerproduct, harmonicmean, cosine, kumar hasse brook, dice, fidelity, bhattacarya 1, bhattacarya 2, hellinger, jensen, jensen shannon, topsoee, kullback divergence, jeffreys, kullback leibler, squared euclidean, pearson chi squared, neyman chi squared, squared chi squared, divergence, clarck, additive symmetric chi squared, eder simple, burrows delta, argamon linear delta, eders delta, argamons quadratic delta, wasserstein 1d Maße

#### **Gruppierung:**

hierarchische Clusterung, multidimensional scaling (MDS), tSNR (t-distributed stochastic neigbor embedding)\_\_\_\_\_

Gruppierung

## Benutzung

## **GUI Konzept**

Jeder Verarbeitungsschritt wird in einem Konfigurationsschritt eingestellt. Jeder Konfigurationsschritt besitzt ein zusätzliches Kommentarfeld und eine der Anzeige der Zwischenergebnissen (rechte Spalte). Die GUI empfindet die Schritte der Textanalyse nach. Einzige Ausnahme bildet die Auswahl des Corpus, die ist zwar an oberer Stelle angezeigt, aber von der Benutzung her ist das der letzte Schritt! Jedes GUI-Element ist mit einer Beschriftung versehen. Sofern die Funktion einer weiteren Erklärung bedarf, dann ist zusätzlich eine kursive Erklärung angefügt.

## Überblick zur Benutzung

- 1. Konfiguration des Browsers
- 2. Konfigurieren der Analyseschritte
- 3. Corpus-Auswahl
- 4. Erneute Berechnung und Serien

## 1. Konfiguration des Browsers

Speicherort: Legen sie einen Ordner an in welchem die Ergebnisse von stylo-ah-online gespeichert werden können. Konfigurieren Sie ihren Browser so, dass er die Dowloads in diesem Ordner ablegt. die Anleitung für Firefox:

https://support.mozilla.org/de/kb/suchen-und-verwalten-heruntergeladener-dateien#w\_ziel-ordner-der -heruntergeladenen-dateien-andern

Datenbanken: Teilen sie dem Browser mit, dass er Datenbanken anlegen und nicht löschen soll. Dazu stellen sie sicher, dass die Chronik angelegt wird: https://support.mozilla.org/de/kb/firefox-chronik-zeigt-ihre-besuchten-webseiten

Web-Konsole: stylo-ah-online ist in JavaScript geschrieben. Die Web-Konsole bietet eine Darstellung von Meldungen zum Programmablauf und Hinweise auf Fehler. Nach einem Fehler können sie die Seite neu laden und den Ablauf erneut starten. Es bietet sich an die Web-Konsole anzuzeigen. Für Firefox aktiviert man diese so:

https://firefox-source-docs.mozilla.org/devtools-user/tools\_toolbox/index.html

## 2. Konfigurieren der Analyseschritte

### 2.1 Arbeitsschritt: Benennung

#### Naming

| default Sub<br>unfinished V        | ject (Name the subject of the file.) State (Select/give a state, if this is not given by file ending.) |  |
|------------------------------------|--|--|
| unfinished                         | State (Select/give a state, if this is not given by file ending.)                                      |  |
| ID /                               |  |  |
| ID (                               | Give the ID of the file.)  |  |
| 2023-07-14 Date                    | Date (The actual date, erase to reset.)  |  |
| Vers                               | sion (Choose a version of the file.)   |  |
| default Auth                       | Author name (Fill in the name of the author.)  |  |
| File                               | ending (Provide a file ending.)  |  |
| experiment_default_unfinished_2023 | 3-07-14_default.styloahonline  |  |

Das Benennungsmodel beinhaltet zusammengesetzte Namen für die Experimente. Dabei werden die einzelnen Teile der Benennung durch Unterstriche abgetrennt. Leerzeichen in Einträgen werden durch Bindestriche ersetzt. Die einzelnen Einträge der Benennung erlauben verschiedene Organisationsmuster für Experimente. Das Datum wird beim ersten Programmlauf automatisch ausgefüllt. Der einmal erstellte Name charakterisiert den Programmablauf und die Einstellungen genau. Änderungen in der Benennung führt auf ein neues Experiment, welches auch separat gespeichert wird. Die Benennung geht in die Config-Datei ein und ein erneuter Aufruf einer Config-Datei stellt alles unter dieser Bezeichnung wieder her. Auf diese Weise können mehrer Konfigurationen unterschieden werden.

## 2.2 Arbeitsschritt: Configurations-Datei / Datenbank

#### Configuration

#### Config for text analysis

 Save cont file
 (This will run the analysis again, if you made changes to the settings below.)

 Datei auswählen
 Keine ausgewählt
 (Choose a existing stylo-online configuration file to set the configuration for stylo-online.)

#### GEN config for SERIAL text analysis

 Gen all config token
 (This will generate config files for each token version (1-3 gram), but leaves the other configuration unchanged.)

 Gen all config counting
 (This will generate config files for each counting method, but leaves the other configuration unchanged.)

 Gen all config measures
 (This will generate config files for each measure, but leaves the other configuration unchanged.)

#### Config for stylo-ah-online display

Display size of results (Checked: Just show a sample of the results (1000 token/signs). Not checked: Results will be shown in full length.)

#### Delete

Delete configuration (This will delete the configuration.) Delete data base (This will delete the stored files and the results of the analysis.) Reset stylo-ah-online (This will reset stylo ah online to start an new analysis.)

Im Abschnitt "Configuration" können alle Einstellungen vorgenommen werden, die sich auf die Config-Datei beziehen. stylo-ah-online erlaubt die Speicherung der Konfiguration und genauso ihren erneuten Aufruf durch Eingabe einer Config-Datei. Für die Erstellung von Serien von Experimenten erlaubt es stylo-ah-online mehrer Config-Dateien zuerstellen und als Serie wieder zu öffnen. Die Ergebnisse werden für jede so gespeicherte Einstellung unter einem extra Namen abgespeichert. Für umfangreichere Corpora ist es nötig die Darstellung von Zwischenergebnissen zu beschränken. Das kann an dieser Stelle mit dem Haken bei "Display size of results" erzielt werden. stylo-ah-online verfügt über eine Auto-save-Funktion, außerdem werden die eingegebenen Daten gespeichert, um diese bei einem erneuten Analysedruchlauf aus der Browser eigenen Datenbank zu holen. Man kann hier die Einstellungen und die Datenbanken oder beides zurücksetzen.

#### 2.3 Arbeitsschritt: Input / Replication

| Dateien auswählen Keine ausgewählt<br>applyed (start analysis) Data in the | (Just choose the CORPUS FILES, than the selection below will be<br>data base will be overwritten ) |
|--|--|
| Re-run anayisis (This will RUN the a                                       | nalysis AGAIN, if you made changes to the settings below. Data is                                  |
| taken from the data base.)   |  |
| Dateien auswählen Keine ausgewählt   | (Choose MULTIPLE config files to perfom multiple analysis on one                                   |
| corpus.)   |  |
| Note   |  |
|  |  |
|  |  |

(Log entry for the input section.)

Im Abschnitt "input / Replication" geht es um den Aufruf der Textdateien, die analysiert werden sollen. Diese Handlung löst die erste Analyse aus und speichert Ergebnisse in der Datenbank. Mit "rerun" kann man eine veränderte Konfiguration auf die vorhandenen Daten anwenden. Sollten Serienexperimente vorbereitet wurden sein, dann kann man die dazugehörigen Config-Files hier eingeben und so die Serienverarbeitung auslösen. Dazu werden die Daten aus der Datenbank verwendet. Der Arbeitsschritt der Dateieingabe muss der Letze sein, nachdem alle anderen Konfigurationen vorgenommen wurden.

#### 2.4 Arbeitsschritt: Normalisierung (Normalization)

## NORMALIZATION

(Please check <u>http://ecomparatio.net/~khk/NORM-DECOMP-DIST/textnorm.html</u> to see some examples of how the selection would work.)

## Word masking / stop words

Use Word masking (Give back the string without stop words.)

Show stop word list (Check this to apply stop word removal.)

Datei auswählen Keine ausgewählt (Choose a existing stop word file (CSV format, divider: ;;).)

## Sign equalization

- Disambiguate diacritica
- Disambiguate dashes
- Text output latin u-v (repaces all u with v)
- Text output latin j-i (repaces all j with i)
- □ lota sub to ad (takes greek utf8 string and repleces iota subscriptum with iota ad scriptum)
- Text output tailing sigma uniform (equalize tailing sigma)
- Text output without diacritics (replaces diacritics)
- □ Text output without some signs (delete some to the programmer unknown signs: †, \*, <, #, §, \$)
- Text output without ligature (takes a string, return string with ligatures turned to single letters)
- Text output equal case (input a string and get it back with all small case letters)
- Text output no brackets (input string and get it back with no brackets)

## Markup / Format

- Without markup (input a string and get it back with markup (html / xml) removed)
- Delete punctuation (takes string and returns the string without punctuation)
- Without newline (input string and get it back with linebreaks removed)

## Word level conversions

- Elision expansion (elusion it will be expanded)
- □ Alpha privativum / copulativum (takes utf8 greek and splits the alpha privativum and copulativum from wordforms)
- Text output without numbering (takes string, return string without the edition numbering i.e. [2])
- Text output no hypenation (removes hyphenation)

## Combinations

(Select one of the combined normalization functions (none of the single steps is used).)

## Translitteration

(Select one of the transliterations.)

Noto

## 2.4.1 Word masking / stop words

Setzt man den Haken in diesem abschnitt, dann werden die Wortformen auf der Stop-Wortliste aus den Strings entfernt. Den Vorgang nennt man ebenfalls Maskierung von Wortformen. Mann kann sich mit dem Button die aktuelle Stop-Wortliste anzeigen lassen. Eine andere Stop-Wortliste kann man durch die Eingabe einer Datei, die jedes Stop-Wort durch ";;;" vom nächsten getrennt enthält, nutzen.

## 2.4.2 Sign equalization

In diesem Abschnitt der Einstellungen geht es um die Behandlung einzelner Zeichen und

Zeichengruppen. Die ersten beiden Haken dienen der Vereinheitlichung von diakritischen Zeichen und Strichen. Dadurch wird die unterschiedliche Kodierung visuell gleicher Zeichen bewirkt. Die weiteren Normalisierungen erklären sich (Ersetzung aller u und v durch v, Ersetzung aller i und j durch i, Vereinheitlichung von Schreibweisen Iota und Sigma, Auflösung von Ligaturen, die Lösung einiger Zeichen).

#### 2.4.3 Markup / Format

Vornehmlich geht es um die Lösung zusätzlicher Formatierungsangaben und der Metadaten/Struktur. Sollte die Eingabe in wohlgeformtem XML geschehen, dann löscht der erste Haken dieses unter Verwendung eines XML Parsers. Sollte dem nicht so sein, dann werden die XML Tags mittels eines regulären Ausdrucks gelöscht. Hier können zudem Interpunktion und Zeilenumbrüche gelöscht werden.

#### 2.4.4 Word level conversions

Unter die Veränderungen auf Wortform ebene zählt alles, was die Wortform als logische, organisatorische Einheit berücksichtigt. Die betrifft Wortformtrennungen, Nummerierungen, Elisionen und Alpha privativum.

#### 2.4.5 Combinations

Hier können Kombinationen von Normalisierungsschritten gewählt werden. Andere Einstellungen werden dann ignoriert.

#### 2.4.6 Translitteration

Sollte es nötig sein Texte ihrem Zeichenbestand nach zu vereinheitlichen, dann kann diese durch die Transliteration (Griechisch / Latein) geschehen.

### 2.5 Arbeitsschritt: Zerlegung (features / decomposition / token)

## FEATURES / DECOMPOSITION / TOKEN

(The word level decomposition and the gram decomposition will be combined. Check <u>http://ecomparatio.net/~khk/NORM-DECOMP-DIST/zerl.html</u> for some examples to see how decomposition will work.)

#### Word level decomposition

- O None
- Without consonants (string without consonants)
- Without vowels (string without vowels)
- Small words (string with just small words (stopwords))
- Big words (string with just big words (not stopwords))

#### **General N-Gram decomposition**

| Word level (groups of words) | ✓ Gram-level  |  |  |
|------------------------------|---|--|--|
| 1                            | N (gram-size, set to one means for example word statistics) |  |  |
| 1                            | M (gap-size, for gap n-gram                                 |  |  |
| ✓ Padding (used for          | sign level of words)  |  |  |

#### Note

(Log ontry for the token costion )

### 2.6 Arbeitsschritt: Zählung (Selection / Counting)

## **SELECTION / COUNTING**

relative frequency 
 (select the counting methode)

#### Most frequent token / words (per text)

| 0   | Min value (po |
|-----|---------------|
| 100 | Max value (po |

value (position in frequency ordered list) value (position in frequency ordered list)

#### Culling (per corpus)

\_\_\_\_\_

Min value (per cent of presents of a token in all texts) Max value (per cent)

#### Text length normalization

□ compare fractions of texts (smallest text gives the length; none other methode is applied)

#### Note

(Log entry for the counting section.)

Im oberen Pulldown-Menü des Abschnitts kann man sich für die Zählweise entscheiden, die auf die Menge aller Token (Wortformen) angewendet werden soll. Die absolute Häufigkeit, ist die Anzahl des Auftretens eines Token. Die relative Häufigkeit ist die Auftretenszahl eines Token geteilt durch die Textlänge. Das schwächt den Einfluss der Textlänge ein. Allerdings überdeckt dies das Problem im Korpus sehr kurzer Texte. Die Kodierung des Auftretens in einem Text oder Nichtauftretens kann mittels 0 und 1 geschehen. Dann wird lediglich die Existenz von Token untersucht. Abschließend stehen zwei Zählungen bereit, die eine Beziehung zum Gesamtkorpus herstellen. Die TF-IDF stellt den Bezug zwischen relativer Häufigkeit in einem Text eines Token zur Häufigkeit in Texten vertreten zu sein her. Der Quotient aus relativer Häufigkeit pro Text und relativer Häufigkeit im Korpus mindert mildert die Funktion der relativen Häufigkeit ab und sollte Werte ergeben, die sich zwischen der absoluten Häufigkeit und der relativen Häufigkeit bewegen.

#### 2.6.1 Most frequent token / words (per text)

Mit den beiden Zahlenangaben kann der Frequenzrang angegeben werden, der im Profil Berücksichtigung finden soll. Sollen lediglich die Werte der 100 häufigsten Token Betrachtung finden, dann muss man den niedrigsten Rang 0, als minimal zu berücksichtigenden Rang, angeben und den Wert 100, als höchsten zu berücksichtigenden Rang.

#### 2.6.2 Culling

Mit den beiden Werten kann man angeben in wie viel Prozent der Texte ein Token mindestens auftreten soll und maximal auftreten soll, um im Profil Berücksichtigung zu finden. Im Kontrast zu Most frequent words oder TF-IDF kann man dabei nicht absehen, wie die Profile beeinflusst werden. Es kann durchaus sein, dass damit null-besetzte Bereiche vermindert werden, allerdings kann auch das Gegenteil eintreten. Hier sollte man die Profile genau untersuchen.

### 2.7 Arbeitsschritt: Maßanwendung (Measure selection)

## MEASURE SELECTION

(Please check <u>http://ecomparatio.net/~khk/measuredisplay</u> to see a discription and comparison of the measures usable. See <u>http://ecomparatio.net/~khk/NORM-DECOMP-DIST/index.php</u> for some examples.)

 Cosine
 Measure selection

 1
 Measure order (the order of the measure, additional to minkowski, burrows delta, argamon linear delta, eders delta, argamons quadratic delta, wasserst 1d, gower)

## Note

(Log entry for the comparing section.)

## 2.8 Arbeitsschritt: Gruppierung (Clustering)

## CLUSTERING

strict hierarchical cluster analysis 
Cluster method
single linkage
Hierarchical cluster linkage method

### **Display options**

 430
 Offset pixels (set the pixel distance for the lables in the visualization; used in distance heatmap and cluster visualization)

 600
 Width of diagram (set the width (pixel) of the diagram, the space for the lables is not included)

 600
 Height of diagram (set the height in pixels of the visualization, space for labels not included)

#### Note

(Log entry for the cluster section.)

### 2.9 Arbeitsschritt: Export

### EXPORT

#### **Export Configuration / Presets**

- Export config as text file
- Export stop words as CSV file

#### Multi file export

- Export raw text input (as text file, renamed)
- Export normed string (as text file)
- Export decomposition (as text file)
- Export frequency of token (as CSV file)

#### Single file result export

- Export distance matrix (as text file; usable as gephi import)
- Export cluster analysis (as nodes and edges file; for example as gephi import)
- Export cluster visualization (as SVG)

#### Note

| (Log entry for the export section.)                                   | Λ.                         |
|---|----------------------------|
| Liniversity Trier / Ancient History Trier / eAOUA digital resources / | Volkswagen <b>Stiftung</b> |

im Abschnitt "Export" können, durch das Setzen von Haken, verschiedene Dateien exportiert werden. Diese werden bei jedem Programmdurchlauf geschrieben. Im Abschnitt wird nach drei Typen von Export unterschieden. Der Export grundsätzlicher Dateien, wie der Config-Datei, dem Export von Zwischenergebnissen (die dann auch für jeden eingegebenen Text existieren) und dem Export von Ergebnissen, die für das ganze Corpus gelten. Die Dateien werden im angewählten Download Ordner gespeichert.

## 3. Corpus-Auswahl

Im Vergleich zu stylo ist der Vorgang anders, wenn es um die Auswahl von Corpora geht: In stylo genügt es einen Ordner anzugeben, welcher den Ordner "corpus" enthält. Für die Eingabe in stylo-ahonline muss die gesamte Liste der Texte ausgewählt werden, die das Corpus bilden.

## 4. Erneute Berechnung und Serien

From: http://replicatio.science/dokuwiki/ - documentatio replicationis

Permanent link: http://replicatio.science/dokuwiki/doku.php/de/styloahonline/handbuch?rev=169961516

Last update: 2023-11-10

