Anlautanalysator

1. Was ist Anlautanalysator?

Anlautanalysator ist ein **Korpustool**, mit dem in einem Textkorpus (der Quelldatei) vorkommende Wörter hinsichtlich ihres Anlauts analysiert werden. Unter **Anlaut** verstehen wir den Beginn eines Wortes bis inklusive zum ersten Vokal, also z. B.: *fra* in *fragen*, *мно* in *много*. Die Analyse besteht hauptsächlich darin, alle Wörter der Quelldatei einem bestimmten Anlaut zuzuordnen (z.B. werden *fragen, Frage, Fracht, fragil, frank,* … dem Anlaut *fra* zugeordnet).

2. Wie funktioniert Anlautanalysator?

Zunächst wird die Quelldatei in kontinuierliche Folgen aus alphanumerischen Zeichen zerlegt. In einem nächsten Schritt werden jene Zeichenfolgen, die kein Wort darstellen (z. B. Ziffernfolgen) sowie Zeichenfolgen, die Zeichen enthalten, die nicht zu der ausgewählten Sprache gehören, ausgesiebt, so dass nur noch Wörter der gewählten Sprache übrigbleiben. Diese Wörter werden sodann hinsichtlich ihres Anlauts analysiert und das Ergebnis wird in zwei **Ausgabedateien** geschrieben, eine **Text-Datei** und eine **HTML-Datei**. Die Text-Ausgabedatei enthält pro Zeile 5 durch Tabulatoren getrennte Spalten, die folgende Daten enthalten:

Spalte 1: den *Anlaut.*

Spalte 2: die Zahl der *Types* (unterschiedliche Wörter), die zu diesem Anlaut gehören.

Spalte 3: die Zahl der *Tokens* (Wörter insgesamt), die zu diesem Anlaut gehören.

Spalte 4: die *Type-Token-Ratio* (TTR): Quotient aus Types und Tokens. Je kleiner dieser Wert ist, umso weniger verschiedene Wörter gehören zu dem Anlaut.

Spalte 5: die konkreten *zu dem Anlaut gehörenden Types* mit jeweiliger *Frequenz.*

3. Ein Beispiel

Die Funktionsweise von Anlautanalysator sei an folgendem Minitext (aus der russischen Wikipedia) illustriert. (Um statistisch relevante Aussagen machen zu können, muss natürlich ein viel größeres Korpus analysiert werden.)

```
Солнце состоит из водорода (≈73 % от массы и ≈92 % от объёма), гелия (≈25 % от
массы и ≈7 % от объёма[9]) и других элементов с меньшей концентрацией: железа,
никеля, кислорода, азота, кремния, серы, магния, углерода, неона, кальция и
хрома[10].
```

Die Text-Ausgabedatei sieht folgendermaßen aus:

```
Spalte 1   Spalte 2   Spalte 3   Spalte 4   Spalte 5
со         2          2          1,00       состоит 1, солнце 1
о          2          6          0,33       от 4, объёма 2
ма         2          3          0,67       массы 2, магния 1
и          2          5          0,40       и 4, из 1
э          1          1          1,00       элементов 1
хро        1          1          1,00       хрома 1
у          1          1          1,00       углерода 1
се         1          1          1,00       серы 1
ни         1          1          1,00       никеля 1
не         1          1          1,00       неона 1
ме         1          1          1,00       меньшей 1
кре        1          1          1,00       кремния 1
ко         1          1          1,00       концентрацией 1
ки         1          1          1,00       кислорода 1
```

```
ка          1          1          1,00       кальция 1
же          1          1          1,00       железа 1
дру         1          1          1,00       других 1
ге          1          1          1,00       гелия 1
во          1          1          1,00       водорода 1
а           1          1          1,00       азота 1
```

Zeile 2 (zum Anlaut *o*) ist folgendermaßen zu interpretieren: Es gibt im Quelltext insgesamt sechs Wörter, die mit *o* beginnen (Tokens, Spalte 3), davon sind zwei unterschiedlich (Types, Spalte 2), die TTR beträgt 0,33 (Spalte 4). Konkret handelt es sich um *от* (vier Vorkommen) und *объёма* (zwei Vorkommen) (Spalte 5).

Man beachte, dass

1. das Wort „с“ nicht in der Ausgabedatei erscheint, da es keinen Vokal enthält.
2. das Wort „солнце“ mit kleinem Anfangsbuchstaben erscheint. Die Umwandlung aller Wörter in Kleinbuchstaben ist die Standardeinstellung, diese kann aber verändert werden, s. unten.

4. Programmoberfläche und Bedienung

Nach dem Start erscheint das Programmfenster:

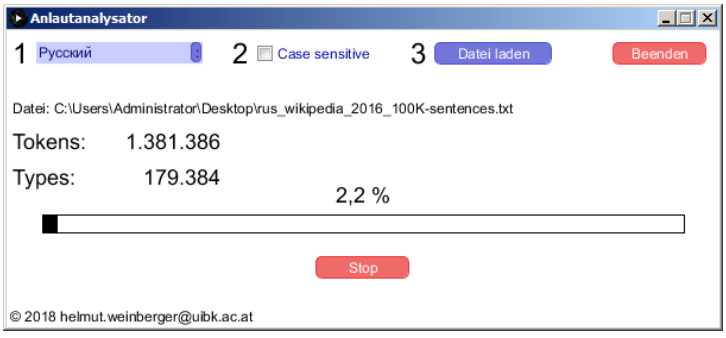
Um einen Text zu analysieren, wählen Sie zunächst unter **1** die **Sprache** des Quelltextes. Es stehen **alle slawischen Sprachen** sowie **Deutsch** und **Türkisch** zur Auswahl.

Unter **2** können Sie einstellen, ob **Groß-/Kleinschreibung** berücksichtigt werden soll. Standardmäßig ist dies nicht der Fall.

Unter **3** laden Sie schließlich die zu analysierende **Quelldatei**.

> Die Quelldatei muss eine Datei im **txt-Format** (reine Textdatei, „plain text“) sein. Beachten Sie, dass die Software kein Error-Checking macht. D. h., wenn Sie versuchen, eine nicht txt-Datei (z.B. eine Word- oder pdf-Datei) zu analysieren oder eine falsche Sprache einstellen, führt dies zu inkorrekten Ergebnissen. Solche Dateien müssen vor der Analyse in Textdateien konvertiert werden.

Nach dem Laden der Quelldatei beginnt die Analyse automatisch.



Während die Analyse läuft, wird der Dateiname angezeigt, darunter zwei statistische Werte: Die Gesamtzahl der zu analysierenden Wörter (Tokens) in der Quelldatei sowie die Zahl der verschiedenen Wörter (Types) in der Quelldatei. Ein Fortschrittsbalken zeigt den Verlauf an. Durch Drücken des Stop-Buttons kann die Analyse abgebrochen werden.

Nach Abschluss der Analyse werden die Ausgabedateien automatisch abgespeichert und die für die Analyse benötigte Zeit angezeigt.

Die **Ausgabedateien** werden in demselben Verzeichnis gespeichert, in dem sich auch die Quelldatei befindet. Beachten Sie daher, dass Sie Schreibrechte für dieses Verzeichnis benötigen.

Die Ausgabedateien erhalten die Namen

`<Name der Quelldatei>_processed.txt` und
`<Name der Quelldatei>_processed.html`.

Die **Text-Ausgabedatei** kann bspw. in Microsoft Excel importiert und weiterverarbeitet werden.

5. Die HTML-Ausgabedatei

Die HTML-Ausgabedatei enthält dieselben Daten wie die Text-Ausgabedatei, allerdings in übersichtlichem Layout sowie mit Sortier- und Suchfunktionen ausgestattet. Der unter 3 betrachtete Minitext sieht in der HTML-Ausgabedatei folgendermaßen aus (links ohne und rechts mit angezeigten Types):

со – Types: 2 • Tokens: 2 • TTR: 1,00
о – Types: 2 • Tokens: 6 • TTR: 0,33
ма – Types: 2 • Tokens: 3 • TTR: 0,67
и – Types: 2 • Tokens: 5 • TTR: 0,40
э – Types: 1 • Tokens: 1 • TTR: 1,00
хро – Types: 1 • Tokens: 1 • TTR: 1,00
у – Types: 1 • Tokens: 1 • TTR: 1,00

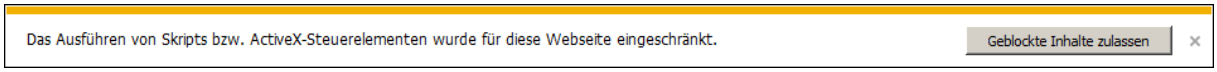
со – Types: 2 • Tokens: 2 • TTR: 1,00
состоит 1, солнце 1
о – Types: 2 • Tokens: 6 • TTR: 0,33
от 4, объёма 2
ма – Types: 2 • Tokens: 3 • TTR: 0,67
массы 2, магния 1
и – Types: 2 • Tokens: 5 • TTR: 0,40
и 4, из 1
э – Types: 1 • Tokens: 1 • TTR: 1,00
элементов 1
хро – Types: 1 • Tokens: 1 • TTR: 1,00
хрома 1

Über die Menüleiste

kann nach Anlaut, Types, Tokens und TTR jeweils auf- und absteigend sortiert werden. Die zu den jeweiligen Anlauten gehörenden Types können en bloc ein- und ausgeblendet werden, auch eine Filterfunktion steht zur Verfügung, mit der Subsets von Anlauten angezeigt werden können. Die Filterfunktion erwartet einen Regulären Ausdruck (Regular Expression). Um z.B. nur jene Anlaute anzuzeigen, die mit „b“ beginnen, geben Sie in das Filterfeld `^b` ein (`^` kennzeichnet den Zeilenbeginn) und klicken anschließend auf *Filter+.*

Wenn Sie die Types eines **einzelnen** Anlauts ein-/ausblenden möchten, klicken Sie einfach auf den entsprechenden Anlaut.

Damit die Menüleiste funktioniert, müssen die Sicherheitseinstellungen Ihres Browsers die Ausführung von Javascript-Code erlauben. Die HTML-Ausgabedatei wurde unter *Firefox, Opera, Chrome, Safari (Mac)* und *Internet Explorer* getestet; bei allen Browsern lief die Datei auf Anhieb, nur beim Internet Explorer erschien folgende Meldung:



Hier muss der Button „Geblockte Inhalte zulassen“ geklickt werden, damit die Menüleiste funktioniert.

6. Benchmarks

Manipulationen an großen Korpora können sehr zeitintensiv sein, die für eine Anlautanalyse benötigte Zeit hängt sehr von der Leistung der verwendeten Hardware ab.

Die folgende Tabelle gibt die zur Anlautanalyse unterschiedlich mächtiger Korpora auf zwei verschiedenen Systemen benötigte Zeit an. Bei den Systemen handelt es sich um:

A - Lenovo ThinkPad E531 mit Intel i3-3120M CPU @ 2,5 GHz und 4 GB RAM

B - Lenovo ThinkCentre mit Intel i5-4670S CPU @ 3,1 GHz und 16 GB RAM

Die Korpora stammen von der Leipzig Korpora Collection (http://corpora.uni-leipzig.de).

Korpus	Tokens	Types	A	B
2016 RU Wikipedia 10.000 Sätze	138.993	41.399	16s	7s
2016 RU Wikipedia 30.000 Sätze	414.125	84.992	1m 13s	27s
2016 RU Wikipedia 100.000 Sätze	1.381.386	179.384	6m 59s	2m 08s
2016 RU Wikipedia 300.000 Sätze	4.131.482	335.717	41m 44s	10m 22s
2016 RU Wikipedia 1.000.000 Sätze	13.800.984	644.571	3h 03m 00s	50m 13s

7. Systemvoraussetzungen

Anlautanalysator wurde mithilfe des Java-Frameworks Processing erstellt. Es benötigt daher **Java**, welches auf neueren Computern in der Regel vorinstalliert ist. Sollte Java auf Ihrem Computer nicht installiert sein, können Sie es über https://java.com/de/download/ gratis herunterladen. (Bei der Windows 64-Bit-Version ist Java von vorneherein dabei.)

Kontakt: helmut.weinberger@uibk.ac.at