TXM-Tutorial

Workshop Digitale Forschungsmethoden in der Romanistik

Forum Junge Romanistik 2016, Universität Würzburg

Christof Schöch
CLiGS - Universität Würzburg

Überblick

1. Einzelne Funktionen
2. Verschiedene Suchmodi
3. Erweiterte Suchen
4. Wildcards und Symbole
5. Strukturierung eines Korpus
6. Vergleichende Abfragen

Textsammlung als Beispiel

- 12 englische Romane von Arthur Conan Doyle
- Aus verschiedenen Jahrzehnten und verschiedenen Untergattungen

1. Einzelne Funktionen

Lexique

Liste aller Wortformen, die im Korpus vorkommen, alphabetisch oder nach Frequenz sortierbar. Nützlich, um besonders häufige und besonders seltene Wörter festzustellen.

- Sortierbar nach Alphabet oder Frequenz

Index

Abfrage der Frequenz einer bestimmten Wortform (oder anderer Suchbegriffe) im Korpus.

- Man kann einstellen, welche Information als "pivot" (das im Zentrum stehende Element) angezeigt werden soll: die Wortform, das Lemma, oder das POS-Tag; oder mehrere dieser Informationen.
- Es werden die Treffer und ihre Frequenz angezeigt; man kann von hier aus weitere Sichten ansteuern (siehe unten)

Concordance

Suche nach dem Kontext einer bestimmten Wortform (oder anderer Suchbegriffe).

- Hier wird das Suchwort im Kontext angezeigt, also mit einigen Wörtern, die davor stehen und die folgen. Ganz links steht, in welchem Text der Treffer jeweils gefunden wurde.
- Auch hier kann man einstellen, welche Informationen als "pivot" angezeigt werden sollen. Die Sortierung kann variiert werden.

Cooccurrences

Suche nach den Wortformen, die oft gemeinsam mit einer bestimmten Wortform vorkommen. Man kann nach verschiedenen Kriterien sortieren:

- `fréquence`: absolute Häufigkeit der kookurrierenden Wörter; abhängig von der Textlänge.
- `cofréquence`: absolute Häufigkeit, mit der das Suchwort und das gefundene Wort gemeinsam vorkommen.
- `indice`: Maß für den Grad an Spezifik, den eine Kookurrenz von zwei Wörtern hat; dies ist die am stärksten aussagekräftige Angabe.
- `distance`: durchschnittlicher Abstand zwischen dem Suchwort und dem gefundenen Wort im Textverlauf.

2. Verschiedene Suchmodi (Index)

Wortformen

Suche nach graphischen Wortformen.

- `"worked"` - einfachste Variante; findet die Anzahl des Wortes "worked"
- `[word="worked"]` - ebenso, aber diese Formulierung ist besser erweiterbar (siehe unten)

Lemmata

Suche nach allen Worten, die einer bestimmten Grundform zugehören.

- `[enlemma="mouse"]` – findet alle Stellen, an denen das Lemma "mouse" in seinen verschiedenen Formen vorkommt: "mouse" und "mice".

Part-of-Speech

Suche nach allen Wortformen, die einer bestimmten grammatikalischen Kategorie (POS) zugehören.

- `[enpos="VV"]` - Suche nach Verben. Findet “was”, “is”, “had”, etc.
- `[enpos="JJ"]` – findet beliebige Adjektive vorkommt: "other", "great", "little".
- (Das Inventar der POS-Tags ist ebenfalls sprachabhängig)

3. Erweiterte Suchen

Folge mehrerer Suchbegriffe

Ein Suchbegriff, gefolgt von einem weiteren Suchbegriff. Solche Abfragen können mit dem Query Editor ("assistant de requêtes") erstellt werden; dort auf "mot supplémentaire" klicken.

- `[word="she"][word="did"]` – Findet die Anzahl der Stellen, an denen "she did" vorkommt. Man kann hier auch verschiedene Suchmodi kombinieren:
- `[enpos="JJ"][enlemma="house"]` - Findet Treffer, bei denen auf ein Adjektiv das Lemma "house" folgt. Bspw.: "great house", "empty house" und "old house".

Suchbegriffe in bestimmtem Abstand

Mehrere Suchbegriffe, mit Minimal- und Maximalabstand dazwischen:

- `[enpos="VV"][]{0,5}[enlemma="money"]` – Findet alle Kombinationen eines Verbs, gefolgt von null bis fünf anderen Worten, gefolgt von dem Lemma "money";
- Beispielsweise: "take his money" und "add to the money" sowie "do to raise the money".

Kombination von Kriterien für ein Element

Verknüpfung von Kriterien unterschiedlicher Art bezogen auf ein einziges Element, bspw. Information über die Graphie eines Wortes mit seiner grammatikalischen Kategorie:

- `[word="dog" & enpos="VV"]` – Findet alle Wörter, die "dog" lauten UND ein Verb sind (to dog = jemanden verfolgen), d.h. schließt alle diejenigen Fälle aus, in denen "dog" als Substantiv verwendet wird.

4. Wildcards und Symbole

Einige nützliche Wildcards:

- `.` Punkt = jedes beliebige Wort-Zeichen (aber keine Leerzeichen etc.)
- `*` Stern = das direkt vorangehende Zeichen soll null mal oder beliebig oft auftauchen
- `+` Plus = das direkt vorangehende Zeichen soll mindestens einmal oder beliebig oft auftauchen
- `?` Fragezeichen = das direkt vorangehende Zeichen soll entweder null mal oder einmal auftauchen

Einige nützliche Symbole:

- `{x}` Numerische Angabe: Gibt an, dass das vorangehende Zeichen oder Element genau x-mal vorkommen soll.
- `{x,y}` Numerischer Bereich: Gibt an, dass das vorangehende Zeichen oder Element zwischen x-mal und y-mal vorkommen soll.
- `%c` Zusatz nach den Anführungszeichen: die Groß- und Kleinschreibung wird ignoriert.
- `%d` Zusatz nach den Anführungszeichen: die Akzente und andere diakritischen Zeichen werden ignoriert. Kann auch als `%cd` kombiniert werden.

Beispielabfragen:

- `[word="heaven"%c]` – Ignoriert Groß- und Kleinschreibung; findet: "heaven" (physisch) und "Heaven" (metaphysisch / am Satzanfang).
- `[word="the"][enpos="JJ"]{2}[enpos="NN"]` - Wildcard auf Ebene der Einheiten; Sequenzen, in denen erst das Wort "the" kommt, dann genau zwei Adjektive, dann ein Substantiv; findet bspw. "the deep blue sky" oder "the angry old man".

Suche mit Alternativen

Mehrere Suchbegriffe alternativ suchen (mit dem "|"-Zeichen)

- `[word="city|country"]` - Alternativen bezogen auf Einzelwörter; findet jeweils die Stellen, an denen "city" oder "country" vorkommt.
- `[enpos="DT|PP"]` - Alternativen bezogen auf die Wortart; Suche nach Artikeln oder Personalpronomina; findet u.a. "the", "a", "it", "you", etc.

5. Strukturierung eines Korpus

Unter-Korpus definieren

Eine Teilmenge des Gesamtkorpus.

- Rechts-Klick auf `DOYLE`, dort `Sous-Corpus` auswählen.
- Dem Unterkorpus einen Namen geben, bspw. "historical".
- Im Reiter `simple` bleiben und die Einstellungen anpassen (Structure: `text` – Propriété: `subgenre`).
- Aus der sich öffnenden Liste der Genres "historical" auswählen, OK klicken.
- Für andere Kriterien das Vorgehen entsprechend anpassen oder die feiner zu steuernden Verfahren `assisté` oder `avancé` wählen.

Korpus partitionieren

Das Korpus intern strukturieren.

- Rechts-Klick auf `DOYLE`, dort `Partition` auswählen.
- Dem geteilten Korpus einen Namen geben, bspw. "Jahrzehnte".
- Im Reiter `simple` bleiben und die Einstellungen passend einrichten: Structure: `text` – Propriété: `decades`). OK klicken.
- Hier wird Gesamtkorpus automatisch nach den relevanten Klassen aufgeteilt.

6. Vergleichende Abfragen

Progression

- Rechtsklick auf einen Treffer (in Index oder Lexique)
- Auf `Envoyer vers progression` klicken
- Type de graphe: `cumulatif` auswählen (zeigt die Trefferzahlen kumulativ, d.h. als ansteigende Kurve an; die Steigung der Kurve zeigt dann die lokale Frequenz an.)
- Échelle des bandeaux: "0.5" einstellen (oder zwischen 0.2 und 1.0 variieren)
- Unité strucurelle: `text` auswählen
- Propriété: `subgenre` oder `title` auswählen; OK klicken!

Spécificités

- Auf eine Partition klicken, dann Rechtsklick
- Auf `Spécificités` klicken
- Propriété definieren (bspw. `word`)
- `Focus de partie`: leer lassen für Vergleich aller Teile; oder einen Teil auswählen. OK klicken.

Christof Schöch, 2016

http://www.christof-schoech.de
