

CLARIN-Tools und Topologische Muster

05/02/2024 22:42:06

[FAQ Article Print](#)

Category:	FAQ Clarin::Tools	Votes:	0
State:	public (all)	Result:	0.00 %
Language:	de	Last update:	11:34:23 - 07/25/2016 (Europe/Berlin)

Keywords

Syntax

Symptom (public)

Existiert eine Software in der Clarin-Umgebung, mit deren Hilfe ich für einen spezifischen Text die wiederholte Verwendung eines bestimmten Satzbauplans bzw. topologischen Musters nachweisen kann?

Problem (public)

Solution (public)

Um einen Satzbauplan wie "Subjekt - Prädikat - direktes Objekt" in einem Text nachzuweisen, muss der Text mit grammatischen Funktionen ausgezeichnet ("annotiert") sein. Die Abfrage von topologischen Mustern verlangt eine entsprechende Annotation von topologischen Feldern.

Wenn Sie einen eigenen Text analysieren wollen, können Sie ihn in der WebLicht-Toolchain (<https://weblicht.sfs.uni-tuebingen.de>) mit den entsprechenden Annotationen anreichern. Genauer dazu weiter unten.

1. Suche auf vorhanden Korpora.

Im CLARIN-D-Kontext gibt es auch Korpora, die mit dieser Art von Annotation bereits manuell ausgezeichnet sind:

- Grammatische Funktionen: [1]Tiger , [2]TüBa-D/Z

- Topologische Muster: [3]TüBa-D/Z

TüBa-D/Z ist über den Webservice Tundra abfragbar. Sie können dort durch ihre Suchergebnisse browsen. Tundra ist angelehnt an das Suchtool TIGERSearch, das Sie sich plattformunabhängig lokal installieren können. TIGERSearch bietet zusätzlich zum Ergebnisbrowsen einfache Frequenzaufstellungen und Download-Optionen.

- [4]Tundra

- [5]TIGERSearch (lokale Installation nötig)

- [6]TüBa-D/Z

- [7]Tiger Korpus: (Download über "License")

1.1 Beispielanfrage für topologische Muster Eine Beispielanfrage auf TüBa-D/Z (für Tundra und TIGERSearch), die Verbzweitsätze findet:

"Finde ein Vorfeld (VF), das unmittelbar einer linken Satzklammer (LK) vorangeht (und beide topologische Knoten werden unmittelbar vom selben Satz-knoten (SIMPX) dominiert)"

#1:[cat="VF"].#2:[cat="LK"] & #0:[cat="SIMPX"] > #1 & #0 > #2

1.2 Beispielanfrage für Satzbauplan Satzbaupläne (grammatische Funktionen) sind auf TüBa-D/Z theoretisch abfragbar, erfordern aber etwas komplexere Suchanfragen wegen der "intervenierenden" Knoten für topologische Felder. Im Tigerkorpus sind Satzbaupläne unmittelbar abfragbar.

Eine Beispielanfrage für Satzbaupläne auf Tiger (momentan nur über TIGERSearch):

"Alle Sätze mit dem Satzbauplan Subjekt - Prädikat - Akkusativobjekt"

#0:[cat="S"]& #0 > SB #1:[] & #0 > HD #2:[] & #0 > OA #3:[] & #1 . #2 & #2 . #3 & arity (#0,3) (= Gesucht ist ein S-Knoten, der unmittelbar drei andere Knoten dominiert, und zwar mit den Funktionen Subjekt (SB), Kopf (HD) und Akkusativobjekt (OA), wobei folgende unmittelbare Präzedenz zwischen den Töchtern besteht: SB vor HD vor OA. Die Stelligkeit (arity) von S ist auf drei beschränkt.)

2. Suche auf eigenen Texten (mit Vorverarbeitung)

Wenn Sie einen eigenen Text analysieren wollen, können Sie ihn mit der WebLicht-Toolchain (<https://weblicht.sfs.uni-tuebingen.de>) anreichern und anschließend in Weblicht selbst sichten und in Tundra oder TIGERSearch gezielt abfragen.

2.1. Syntaktische Vorverarbeitung

hierzu müssen Sie Ihren Text in WebLicht hochladen. WebLicht akzeptiert eine ganze Reihe von Input-Formaten (txt, pdf, doc, verschiedene xml-Formate).

2.1.1 Annotation von topologischen Feldern (und Konstituenten)

WebLicht-Einstellung: production Toolchain: - Text laden - Sfs To TCF
Converter - IMS Tokenizer - IMS TreeTagger - Sfs Berkley Parser => Ausgabe
ähnlich wie TüBa-D/Z (ohne grammatisch Funktionen)

2.1.2 Annotation von Konstituenten und Funktionen

WebLicht-Einstellung: production Toolchain: - Text laden - Sfs To TCF
Converter - IMS Tokenizer - IMS TreeTagger - IMS Constituent Parser => Ausgabe
ähnlich wie das Tigerkorpus (ABER: Konstituenten und Funktion als ein
kombiniertes Label, z.B. NP-SB)

2.2 Sichtung der annotierten Daten

Die automatische Analysen ähneln den Analysen in den manuell annotierten
Korpora (Tiger, TüBa-D/Z), sind aber bis zu einem gewissen Grad
fehlerbehaftet. Als Grundlage für eine Datensichtung sind sie aber sicher
hilfreich.

2.2.1 Browsen der Ergebnisse in WebLicht

Eine einfache (grafische) Sichtung ohne Suchfunktion bietet WebLicht über die
"Visualize Results"-Funktion.

2.2.2 (Online-)Abfrage der Ergebnisse in Tundra

Hierzu müssen Sie das Ergebnis des Parsens über die Download -Funktion in
WebLicht herunterladen und anschließend in Tundra hochladen
(<http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra>). =>
gezielte Suchanfragen möglich (siehe oben)

2.2.3 Abfrage der Ergebnisse in TIGERSearch

TIGERSearch muss lokal auf Ihrem Rechner installiert sein
(<http://www.wolfganglezius.de/doku.php?id=cl:tigersearch> (Ein neuer
Webaufruf am IMS Stuttgart ist in Arbeit)). Um Ihre Daten in TIGERSearch
einlesen zu können müssen Sie in WebLicht einen weiteren
Vorverarbeitungsschritt an das Ende der oben genannten Toolchains ergänzen: -
Sfs Convert to Negra Dann Download des Konvertierungsergebnisses und Upload in
TIGERSearch (über TIGERRegistry: Negra-Filter). => gezielte Suchanfrage,
Frequenzauflistungen und Download der Suchergebnisse möglich

[1] <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html>

[2] <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html>

[3] <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html>

[4] <http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra>

[5] <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html>

[6] <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html>

[7] <http://www.ims.uni-stuttgart.de/forschung/ressourcen/corpora/tiger.html>