Über die Datensammlung und Dokumentation

Inhalte der Datensammlung

Das Korpus enthält Textausschnitte rund um das **Schlagwort "utraquistisch"/"Utraquismus"/"Utraquist" (Abfrage: "utra??is*"** aus historischen Zeitungen und Zeitschriften, die im ANNO (AustriaN Newspapers Online) der Österreichischen Nationalbibliothek verfügbar und durchsuchbar sind. Die dort verfügbaren OCR-Scans wurden manuell überprüft und korrigiert sowie in Bezug auf Orthographie und Morphologie normalisiert, um bestmögliche automatische Verarbeitung mit auf die Gegenwartssprache des Deutschen ausgelegten Taggern und Lemmatizern zu ermöglichen.

Es wurden keine kompletten Zeitungsausgaben oder Zeitungsartikel sondern nur Textausschnitte bearbeitet, die wie folgt definiert wurden:

> **ein maximaler Textausschnitt** = der Satz mit dem Schlagwort plus bis zu fünf Sätze vor und/oder nach diesem innerhalb eines Absatzes

Ein Textausschnitt (kurz: "text") enthält demnach maximal 11 Sätze innerhalb eines Absatzes.

Jeder Textausschnitt (XML-Element `text`) ist in Bezug auf die folgenden **Metadaten** (als Attribute) annotiert:

- Zeitung/Zeitschrift (`source`)
- eventuell: Beilage der Zeitung/Zeitschrift (`source2`)
- Veröffentlichungsort (`place`)
- Veröffentlichungsdatum (`date`)
- Periode (für einfachere Durchsuchbarkeit, als Zuordnung zu einem Jahrzehnt, `period`)
- Genrezuordnung (aktuell noch provisorisch, `genre`)
- Thema des Artikels (aktuell noch provisorisch, `topic`)
- Bezugsraum des Artikels (`space`)

Jede **orthographische/morphologische Normalisierung** (XML-Element `edit`) ist durch folgende Attribute näher spezifiziert:

- originale Variante (`original`)
- Klassifikation des Abweichungstyps (aktuell noch provisorisch, `type`)

Arbeit mit der Datensammlung

Das Korpus wird einerseits als **.xml-Dokument** zur Verfügung gestellt. Andererseits ist es auf **SketchEngine** verfügbar und kann dort auf Anfrage freigeschalten werden.

> **ACHTUNG**: Der Lemmatizer erkennt die verschiedenen Formen der Schlagwörter nicht als jeweils ein Lemma! Um alle Formen (aller Schlagwörter zu finden) muss mit der **CQL-Abfrage** `[lemma="utraquis.*"]|[lemma="Utraquis.*"]` gearbeitet werden.

Entstehung

Die Datensammlung ist das **Ergebnis einer Hausaufgabe** in einem von Agnes Kim am Institut für Germanistik der Universität Wien im Sommersemester 2021 angebotenen Proseminar zur "Historischen Soziolinguistik". Ziel der Aufgabe war, zwischen acht und zehn Texte aus einem vorgegebenen Paket rund um Suchergebnisse für die **Abfrage "utra??is*" aus dem ANNO,** also aus historischen Zeitschriften aus dem Bestand der Österreichischen Nationalbibliothek in ein XML-Dokument zu übertragen und so zu normalisieren, dass die Texte auch mit den auf die Gegenwartssprache trainierten, auf SketchEngine zur Verfügung stehenden Taggern für das Deutsche bearbeitet werden können. Die Daten wurden von den Studierenden eingegeben und von Agnes Kim überprüft und angeglichen.

Verfügbarkeit

Da es sich aktuell um eine erste Version dieser Datensammlung handelt, ist sie noch ausschließlich als **Download aus der dioecloud** verfügbar. Außerdem ist das Korpus auf SketchEngine verfügbar und kann auf Anfrage (per Mail an agnes.kim@univie.ac.at) freigeschalten werden!

Version/Dateiname	Veröffentlichungs-datum	ungefähre Größe	Link
v 0.1 / Korpus_Utraquis_v0.1.CSV	12.05.2021	156 Textausschnitte 31.176 Tokens	https://dioecloud.trans.univie.
v 0.2 / Korpus_Utraquis_v0.2.CSV	06.06.2021	156 Textausschnitte 31.176 Tokens	https://dioecloud.trans.univie.

Änderungsnotizen können den Seiten zum Korpusaufbau entnommen werden.

Nach weiteren Überprüfungen und eventuellen Ergänzungen ist geplant, das Korpus in einer ersten stabilen Version zu archivieren und auf SketchEngine öffentlich zur Verfügung zu stellen.

Zitation

Zitation der Datensammlung

> Kim, Agnes (JAHRESZAHL DES VERÖFFENTLICHUNGSDATUMS DER VERSION): Historisches Korpus zum Schlagwort "Utraquismus". Datensammlung. VERSIONSNUMMER ANGEBEN In: Dies. (Hg.): Datensammlungen aus der bzw. für die Forschung und Lehre. Online verfügbar unter: LINK ZUR VERSION.

Zitation der Dokumentation

> Kim, Agnes (2021): Historisches Korpus zum Schlagwort "Utraquismus". Dokumentation. In: Dies. (Hg.): Datensammlungen aus der bzw. für die Forschung und Lehre. Online verfügbar unter: https://wiki.dioe.at/books/historisches-korpus-zum-schlagwort-%22utraquismus%22.

Fragen? Fehler gefunden?

Sie haben einen Fehler gefunden, haben Ergänzungen oder Nachfragen? Bitte wenden Sie sich per Mail an mich: **agnes.kim@univie.ac.at**

Revision #2
Created Wed, May 12, 2021 8:27 AM by Agnes Kim
Updated Sun, Jun 6, 2021 2:36 PM by Agnes Kim