

Historisches Korpus zum Schlagwort "Utraquismus"

- Über die Datensammlung und Dokumentation
- Umfang und Aufbau des Korpus
 - Konkretere Informationen zur Korpusaufbau und Korpusgröße in Bezug auf die Originalquellen
 - v 0.1
 - v 0.2
- Metadaten
- Orthographische und morphologische Normalisierung
 - Abweichtungstypen und Abweichungen

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	Veröffentlichungsdatum	ungefähre Größe	Link
v 0.1 / Korpus_Utraquis_v0.1.CSV	12.05.2021	156 Textausschnitte 31.176 Tokens	https://dioecloud.trans
v 0.2 / Korpus_Utraquis_v0.2.CSV	06.06.2021	156 Textausschnitte 31.176 Tokens	https://dioecloud.trans

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

Umfang und Aufbau des Korpus

Konkretere Informationen zur Korpusaufbau und Korpusgröße in Bezug auf die Originalquellen

In der folgenden Tabelle werden die bearbeiteten Textausschnitte in Relation zu einem Schlagwortkorpus, das das gesamte 19. und das frühe 20. Jahrhundert (bis 1918) grob abdeckt und nachvollziehbar macht, gesetzt. Insgesamt können (am 17. 05. 2021) mit der Suchabfrage **"utra??ist*" im ANNO bis inklusive 1918 6.338 Treffer** (= Ausgaben von Zeitungen und Zeitschriften, in denen die Suchabfrage mindestens ein Mal vorkommt) gefunden werden, von denen **6.179** auf Zeitungen und 159 auf Zeitschriften entfallen. Letztere wurden bei den Überlegungen zu einem möglichen Gesamtkorpus außer Acht gelassen.

Periode 1800–1829

Publikationsort	Link zu ANNO	Treffer	davon im Korpus
Brno	Link	15	13 (2 Duplikate)
Wien	Link	11	11
L'viv	Link	1	0 (weil Duplikat)

Periode 1850–1859

Publikationsort	Link zu ANNO	Treffer	davon im Korpus
-----------------	--------------	---------	-----------------

Wien	Link	29	8
Olomouc	Link	3	3
Salzburg	Link	3	3
Brno	Link	1	1
Opava	Link	1	1
Litoměřice	Link	1	0

Periode 1867–1868 (stellvertretend für die 1860er)

Jahr	Publikationsort	Link zu ANNO	Treffer	davon im Korpus
1867	Wien	Link	13	11
1867	Plzeň	Link	4	0
1867	Graz	Link	1	0
1867	Klagenfurt	Link	1	0
1867	Znojmo	Link	1	0
1867	Olomouc	Link	4	0
1867	Praha	Link	2	0
1868	Wien	Link	22	8
1868	Znojmo	Link	7	0
1868	Praha	Link	6	0
1868	Graz	Link	1	0
1868	Plzeň	Link	1	0

Jahr 1890

Publikationsort	Zeitung	Link zu ANNO	Treffer	davon im Korpus
Wien	Neue Freie Presse	Link	11	8
Wien	Die Presse	Link	8	8
Wien	Das Vaterland	Link	12	10
Wien	Wiener Montags-Journal	Link	1	0
Wien	Wiener Zeitung	Link	6	6
Wien	Deutsches Volksblatt	Link	4	4
Praha	alle	Link	13	0
Maribor	alle	Link	9	6
Salzburg	alle	Link	4	4
Litoměřice	alle	Link	8	0
Linz	alle	Link	7	0
Znojmo	alle	Link	3	0

Jahr 1910

Publikationsort	Zeitung	Link zu ANNO	Treffer	davon im Korpus
Wien	Neue Freie Presse	Link	13	10
Wien	Wiener Zeitung	Link	13	10
Wien	Neues Wiener Tagblatt	Link	6	6

Wien	Zollämter- und Finanzwachzeitung	Link	4	4
Wien	Deutsches Volksblatt	Link	5	5
Wien	Arbeiterzeitung	Link	3	3
Praha	alle	Link	30	0
Maribor	alle	Link	15	6
Klagenfurt	alle	Link	10	9
Graz	alle	Link	24	10
Czernowitz	alle	Link	13	0

Mitarbeit?

Sie haben Interesse, einzelne Ausgaben zu einem der Datenpakete zu ergänzen? Gerne sende ich Ihnen eine genaue Liste der noch fehlenden Ausgaben zu! Melden Sie sich bei:
 agnes.kim@univie.ac.at

v 0.1

Veröffentlichung

Datum: 12.05.2021

Link: <https://dioecloud.trans.univie.ac.at/index.php/s/36iS7xZ4f8dD6fK>

Größe des Korpus

Textausschnitte: 156

Tokens: 31,176

Wörter: 26,806

Überblick über den Korpusaufbau

Die Grafiken beziehen sich auf die Gesamtzahl der Token. Sie wurden mit SketchEngine erstellt.

Zeitliche Dimension: Korpusgröße nach Jahrzehnten

Jahrzehnt	Tokenzahl
1800	382
1820	1.115
1850	2.343
1860	3.419
1890	10.024

1900	2.650
1910	11.241

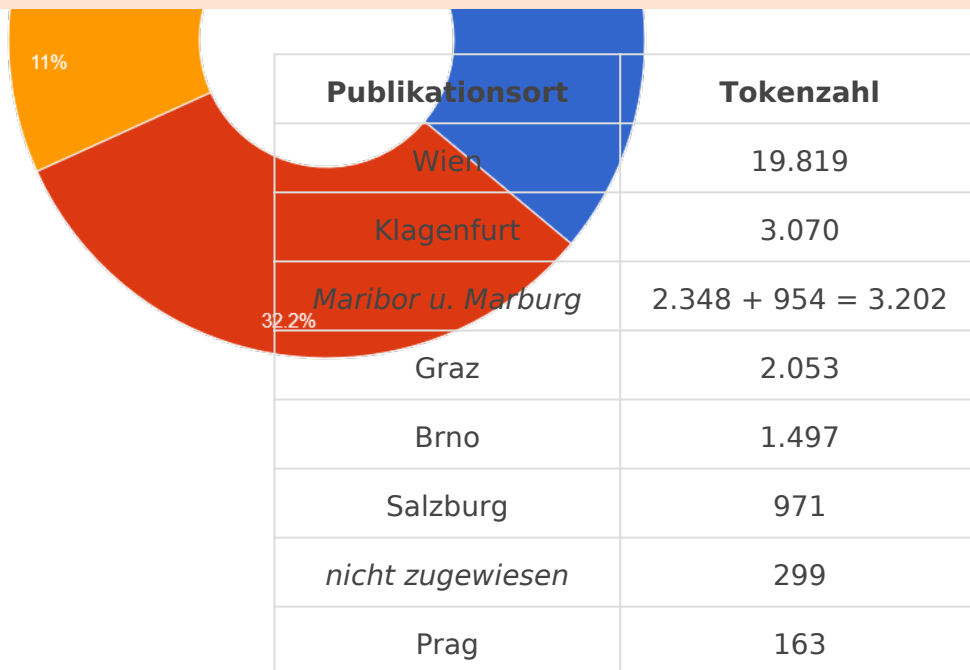
text - period

Räumliche Dimension: Korpusgröße nach Publikationsorten

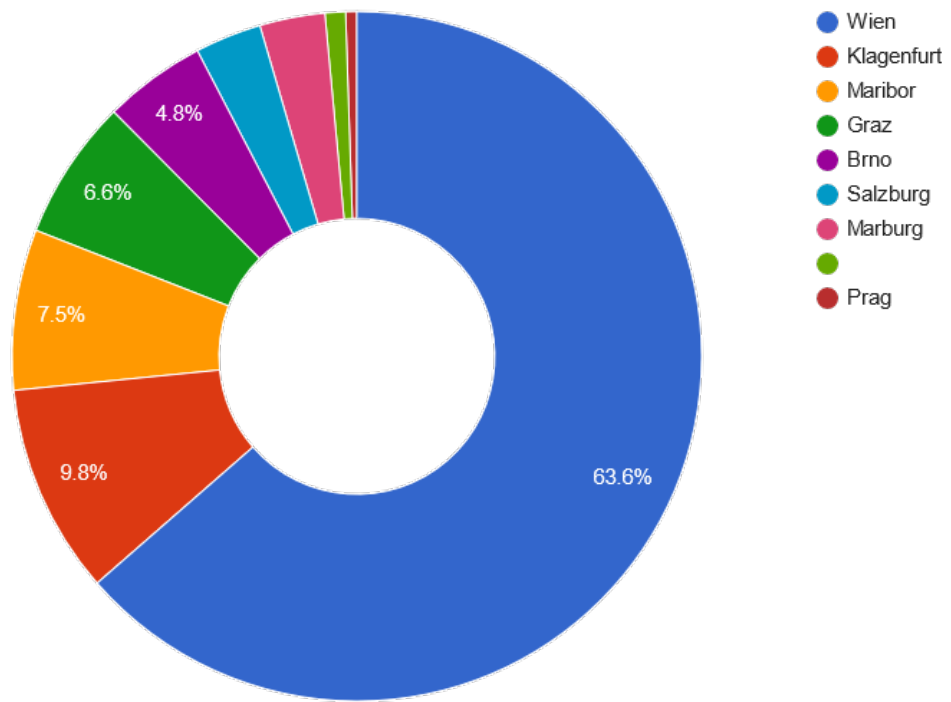


Achtung: In v 0.1 bestehen noch Probleme in Bezug auf die Zuweisung bzw.

Vereinheitlichung der Publikationsorte, die in der Tabelle durch Kursivsatz hervorgehoben werden!



text - place



v 0.2

Veröffentlichung

Datum: 06.06.2021

Link: <https://dioecloud.trans.univie.ac.at/index.php/s/P7p2jMcoFMgWFbx>

Änderungsnotizen

- Vereinheitlichung der Periodenzuordnung (`period`)
- Vereinheitlichung der Publikationsorte (`place`)
- Vereinheitlichung der Annotation der orthographischen und morphologischen Normalisierungen (`edit type`)

Größe des Korpus

Textausschnitte: 156

Tokens: 31,176

Wörter: 26,806

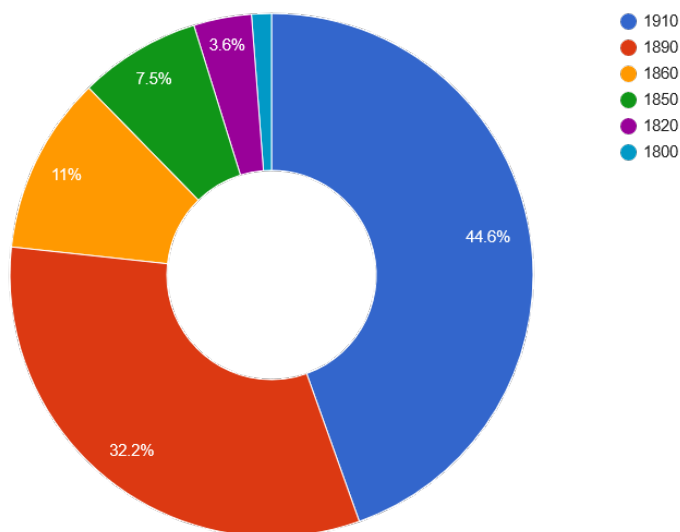
Überblick über den Korpusaufbau

Die Grafiken beziehen sich auf die Gesamtzahl der Token. Sie wurden mit

Zeitliche Dimension: Korpusgröße nach Jahrzehnten

Jahrzehnt	Tokenzahl
1800	382
1820	1.115
1850	2.343
1860	3.419
1890	10.024
1910	13.888

text - period



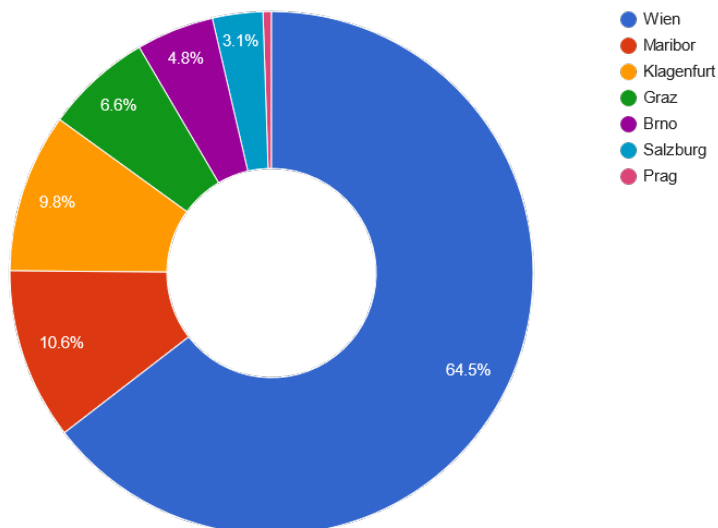
Räumliche Dimension: Korpusgröße nach Publikationsorten

Achtung: In v 0.1 bestehen noch Probleme in Bezug auf die Zuweisung bzw.

Vereinheitlichung der Publikationsorte, die in der Tabelle durch Kursivsatz hervorgehoben werden!

Publikationsort	Tokenzahl
Wien	20.119
Maribor	3.302
Klagenfurt	3.070
Graz	2.053
Brno	1.497
Salzburg	971
Prag	163

text - place



Metadaten

Orthographische und morphologische Normalisierung

Abweichungstypen und Abweichungen

Abkürzung	Erklärung	Beispiele
A	<a>-Schreibung	<i>Kommissär</i>
ABK	Abkürzung	<i>Abg.</i>
AE	<Ä>-Schreibung	<i>Aenderung</i>
DAT	Dativ-e	
EI	<ei>-Schreibung	<i>ey</i>
ERR	Abweichungen, die als Fehler (morphologisch oder orthographisch) eingestuft werden	<i>Agitarion</i>
IE	<i> statt <ie> in {ieren}/{ierung}	<i>finanziren</i>
K	<k>-Schreibung	<i>Collission</i>
KOMP	Schreibung von Komposita	
LEX	lexikalische Abweichungen	
M	<m>-Schreibung	<i>gesammt</i>
MOR	morphologische Abweichungen	
OE	<Ö>-Schreibung	<i>Oesterreich</i>
ORTHKON	andere Abweichungen in der Schreibung von Konsonanten	

ORTHVOK	andere Abweichungen in der Schreibung von Vokalen	
T	<th>-Schreibung	<i>Muth</i>
TSCH	<tsch>-Schreibung	<i>czechisch</i>
UE	<UE>-Schreibung	<i>Uebung</i>
VOKL	Vokallänge	<i>Abendmahl, Schooß</i>
W	<w>-Schreibung	<i>slovenisch</i>
Z	<z>-Schreibung	<i>Civilisation</i>