Nieländer / Jurish

GEORG ECKERT INSTITUT
Leibniz-Institut für internationale Schulbuchforschung

2021

D* für Anfänger:innen: Ein Tutorial

Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung

EDU|MERES

Maret Nieländer / Bryan Jurish

D* für Anfänger:innen: Ein Tutorial

Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung

urn:nbn:de:0220-2021-0088

Cite as:
Maret Nieländer und Bryan Jurish . *D* für Anfänger:innen: Ein Tutorial: Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung*. (2021). urn:nbn:de:0220-2021-0088
.

D* für Anfänger:innen: Ein Tutorial

Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung

Maret Nieländer, Bryan Jurish
Stand: 16.07.2021

GEORG ECKERT INSTITUT
Leibniz-Institut für internationale Schulbuchforschung

gei.Digital

Ziele und Motivation

Die Korpussuche und -analyse in der D*-Umgebung mit Werkzeugen wie DiaCollo bietet ihren Nutzer:innen eine breite Palette von Optionen für einfache und komplexe Anfragen in digitalen Textsammlungen.

Unser Tutorial soll Ihnen den Einstieg in die Materie erleichtern: vorgestellt werden die Benutzungsumgebung, die gebräuchlichsten Parameter der Werkzeuge und einige Beispielanfragen.

Diese Handreichung entstand 2020/21 im Rahmen des Projektes „DiaCollo für GEI-Digital“ am Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung (GEI) und nutzt Beispiele aus dem „GEI-Digital-2020“-Korpus mit mehr als 5000 überwiegend deutschsprachigen Schulbüchern, die zwischen 1648 und 1921 publiziert wurden.

Die digitalen Werkzeuge wurden am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) entwickelt und sind dort auch mit vielen weiteren Textsammlungen nutzbar.

Inhalt

Empfehlung für die Aussprache der Werkzeuge:

"dstar", "D*": /diːˈstɑɹ/ (in etwa: **dii**-star)
"DiaCollo": /diːˈakəloʊ/ (in etwa: dii-**ah**-ko-loh)

Vorab I: Kleines korpuslinguistisches Glossar

Zusammengestellt unter Nutzung des Glossars des „ForText“-Projektes (CC-BY-SA-3.0), verschiedenen Dokumentationen von Bryan Jurish und Wikipedia.

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z

CSV
CSV steht für „Comma Separated Values“ und ist ein Dateiformat in Tabellenform. In einer solchen Datei sind die einzelnen Werte durch Kommata getrennt; in Programmen wie Excel können sie als Tabelle angezeigt werden. Metadaten und KWIC von Suchabfrage- Ergebnissen mit DDC und DiaCollo können im u. a. CSV-Format exportiert werden.

D*
D*, bzw. dstar ist die Korpusmanagement-Umgebung des Zentrums Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW). Sie wird dort in den Projekten DWDS, DTA und ZDL eingesetzt. Für das Projekt „DiaCollo für GEI-Digital“ wurde eine D*-Instanz am Georg-Eckert-Institut für internationale Schulbuchforschung (GEI) eingerichtet.

DDC
DDC („DiaLing/DWDS Concordancer“) ist eine open-source Suchmaschine, die von verschiedenen Projekten der BBAW eingesetzt wird, so etwa in DWDS, DTA, und ZDL Projekten. Sie führt die von Nutzer:innen formulierten Suchanfragen (Queries*)* aus, indem sie die *Indizes* bestimmter Korpora durchsucht.

Default
Das oder der default (engl. für Standardeinstellung) bezeichnet die standardmäßig gesetzten Werte für bestimmte Parameter eines Tools oder Programms, auf die Nutzer:innen bei der ersten Verwendung treffen. Die per default festgelegten Werte der Parameter lassen sich in der Regel manuell umstellen.

DiaCollo
Digitales Werkzeug zur Untersuchung von Kollokationen über die Zeit (**dia**cronic **collo**cation analysis).

Index
Im (Datenbank-)Index werden die Token einer Textsammlung mit einer ID versehen und zusätzlich weitere, durch NLP/Preprocessing gewonnene Informationen gespeichert. Diese Datenstruktur ermöglicht ein effizientes Suchen mit Suchmaschinen wie DDC. Der von DDC in D* genutzte Index bezieht alle Types eines Korpus ein. Der für DiaCollo genutzte Index filtert hingegen mit einer PoS-Positivliste, so dass einige Token in den Suchen/Berechnungen nicht inkludiert werden (siehe stop words*)*.

Vorab I: Kleines korpuslinguistisches Glossar

Zusammengestellt unter Nutzung des Glossars des „ForText“-Projektes (CC-BY-SA-3.0), verschiedenen Dokumentationen von Bryan Jurish und Wikipedia.

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z

Kollokation
Statistisch auffälliges gemeinsames Vorkommen von Wörtern in einem vordefinierten Textabschnitt. Über Kollokationsabfragen (wie DiaCollo) kann z. B. herausgefunden werden, dass ein Wort X häufig in einem definierten Umkreis, z.B. 5 Wörter vor oder nach einem Stichwort Y vorkommt.

Korpus, das
Eine Textsammlung; typischerweise wird ein Korpus zur Beantwortung spezifischer Forschungsfragen oder zur repräsentativen oder vollständigen Abbildung einer Textsorte, eines Ouevres, einer Epoche o.ä. zusammengestellt.

KWIC
KWIC steht für „Keyword in Context“. Es handelt sich um ein Präsentationsformat, das ein ausgewähltes Wort eines Textes oder eines Korpus als Liste in seinen diversen Kontexten (= mit Umgebungswörtern) zeigt. Die Größe der Kontexte kann individuell festgelegt werden.

Lemmatisierung
Bestandteil des NLP/Preprocessing ist die lexikographische Reduktion der Flexionsformen eines im Korpus vorkommenden Wortes auf seine Grundform (Lemma). Z. B. werden Formen wie „sahen“, „sieh“, „gesehen“ dem Lemma „sehen“ zugeordnet.

NLP
NLP steht für „Natural Language Processing“ und wird im Deutschen auch als maschinelle Sprachverarbeitung bezeichnet. Gemeint sind die Bemühungen, Computern beizubringen, natürlichsprachliche Äußerungen korrekt zu verarbeiten und zu analysieren (z. B. durch Lemmatisierung, PoS-Tagging etc.).

OCR
OCR steht für „Optical Character Recognition“, also die automatische Texterkennung von gedruckten Texten: ein Computer „liest” einen gescannten Text und verwandelt diese Bilddatei in einen elektronischen Text. Dieses Verfahren ist kostengünstiger als manuelles Abtippen, allerdings für bestimmte Schriftarten, beschädigte Vorlagen u. ä. bislang noch recht fehleranfällig. Die Volltexte im GEI-Digital-2020 Korpus wurden durch OCR erstellt.

Vorab I: Kleines korpuslinguistisches Glossar

Zusammengestellt unter Nutzung des Glossars des „ForText“-Projektes (CC-BY-SA-3.0), verschiedenen Dokumentationen von Bryan Jurish und Wikipedia.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

PoS
PoS steht für „Part of Speech“, d. h. Wortart. Ein PoS-Tagging ist die automatische Erfassung und Kennzeichnung von Wortarten im Rahmen von NLP/Preprocessing. Die so gewonnenen Zusatzinformationen stehen dann im Index eines Korpus für Queries zur Verfügung, z.B. für die Eingrenzung von Suchen nach Nomen (pos-tag: NN), Adjektiven (ADJ) usw.

Query
Query (engl. für Anfrage/Abfrage): Eine computergestützte Abfrage zur Analyse eines Textes/Korpus. Dies kann ein gesuchtes Stichwort sein, oder eine komplexe Suche z. B. mithilfe Regulärer Ausdrücke oder unter Benutzung der Query Language (Abfragesprache) die von der genutzten Suchmaschine unterstützt wird
(z.B. für die DDC Suchmaschine die DDC query language).

Reguläre Ausdrücke
In Programmier- und Abfragesprachen (Queries) verwendete, nach vordefinierten Konventionen erstellte Zeichenketten (auch Regular Expressions oder RegEx genannt), die bestimmte Operationen wie z. B. Verkettung, Disjunktion, Wiederholung usw. prägnant darstellen.

Stop words/ Stoppwortliste
Als Stoppwörter bezeichnet man diejenigen Wörter, die bei einer digitalen Textanalyse jeweils unberücksichtigt bleiben sollen. Oft sind das Funktionswörter, die aufgrund ihrer grammatisch bedingten Häufigkeit die Auswertungsergebnisse verzerren würden. Sie können durch Filterungen oder mittels Stoppwortlisten entfernt oder ausgeblendet werden.

Type/Token
In der Linguistik verwendete Begriffe aus der Zeichentheorie/ formalen Logik. Während Type zusammenfassend jeden in einem Text oder einer Textsammlung vorkommenden Wort-Typ bezeichnet (z.B. einen Suchbegriff wie „Schule”), bezeichnet Token jedes einzelne Vorkommnis dieses Typs (d. h. bspw. mehr als 77.000 Treffer für „Schule“ im GEI-Digital-2020 Korpus).

Vorab II: Das GEI-Digital-2020 Korpus

gei.digital
Die digitale Schulbuch-Bibliothek

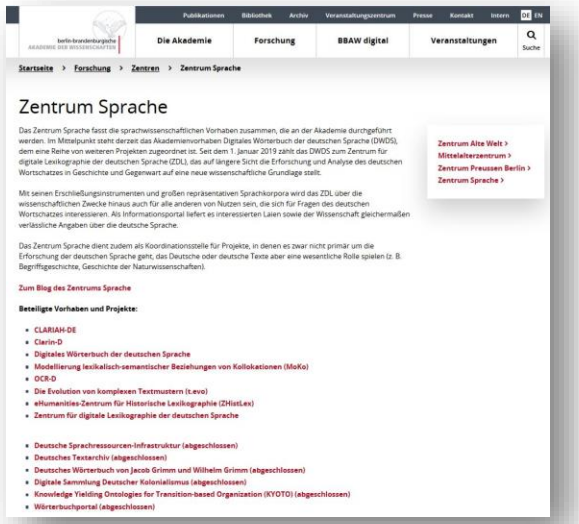
+

DTA

=>

gei.Digital

Historische Schulbücher aus der digitalen Schulbuchbibliothek GEI-Digital des Georg-Eckert-Instituts – Leibniz-Institut für internationale Schulbuchforschung

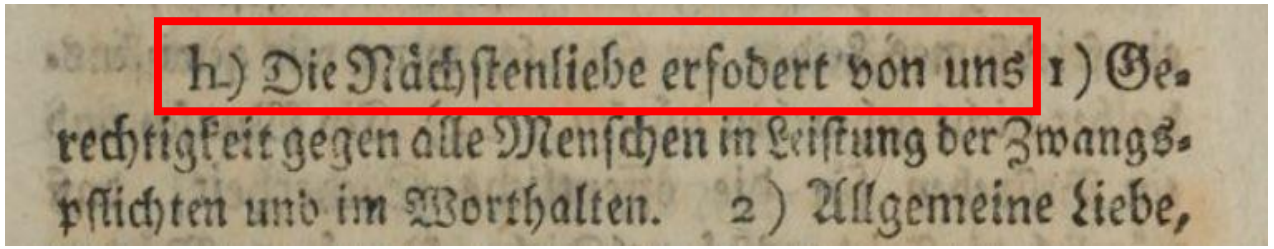


NLP- und Analyse-Werkzeuge des Zentrums Sprache
Berlin-Brandenburgische Akademie der Wissenschaften

Das GEI-Digital-2020 Korpus umfasst alle Werke, die Ende Dezember 2020 auf GEI-Digital mit automatisch generiertem digitalen Volltext zur Verfügung standen.

Das GEI-Digital-2020 Korpus: NLP – maschinelle Sprachverarbeitung

Die in diesen Folien gezeigten Beispiele stammen aus dem Korpus „GEI-Digital-2020“ des Projektes „DiaCollo für GEI-Digital“. Es wurde aus den Metadaten und automatisch generierten Volltexten von 5036 zwischen 1648 und 1921 publizierten überwiegend deutschsprachen Schulbüchern aus der digitalen Schulbuchbibliothek GEI-Digital erstellt. Die Texte wurden dafür nach TEI konvertiert und verschiedenen NLP-Verfahren unterzogen, darunter Tokenisierung, Normalisierung, Wortartenerkennung und Lemmatisierung.


h.) Die Nächstenliebe erfodert von uns 1) Ge-
rechtigkeit gegen alle Menschen in Leistung der Zwangs-
pflichten und im Worthalten. 2) Allgemeine Liebe,

Satz Nr.	Token Nr.	Token	POS-tag	Lemma
1	1	h	NN	h
1	2	)	$(	)
1	3	Die	ART	d
1	4	Nächstenliebe	NN	Nächstenliebe
1	5	erfo[r]dert	VVFIN	erfordern
1	6	von	APPR	von
1	7	uns	PPER	wir

Tokenisierung:
- ✓ Jedes Token erhält eine adressierbare ID
- ✓ Absolute und relative Häufigkeiten können ermittelt werden

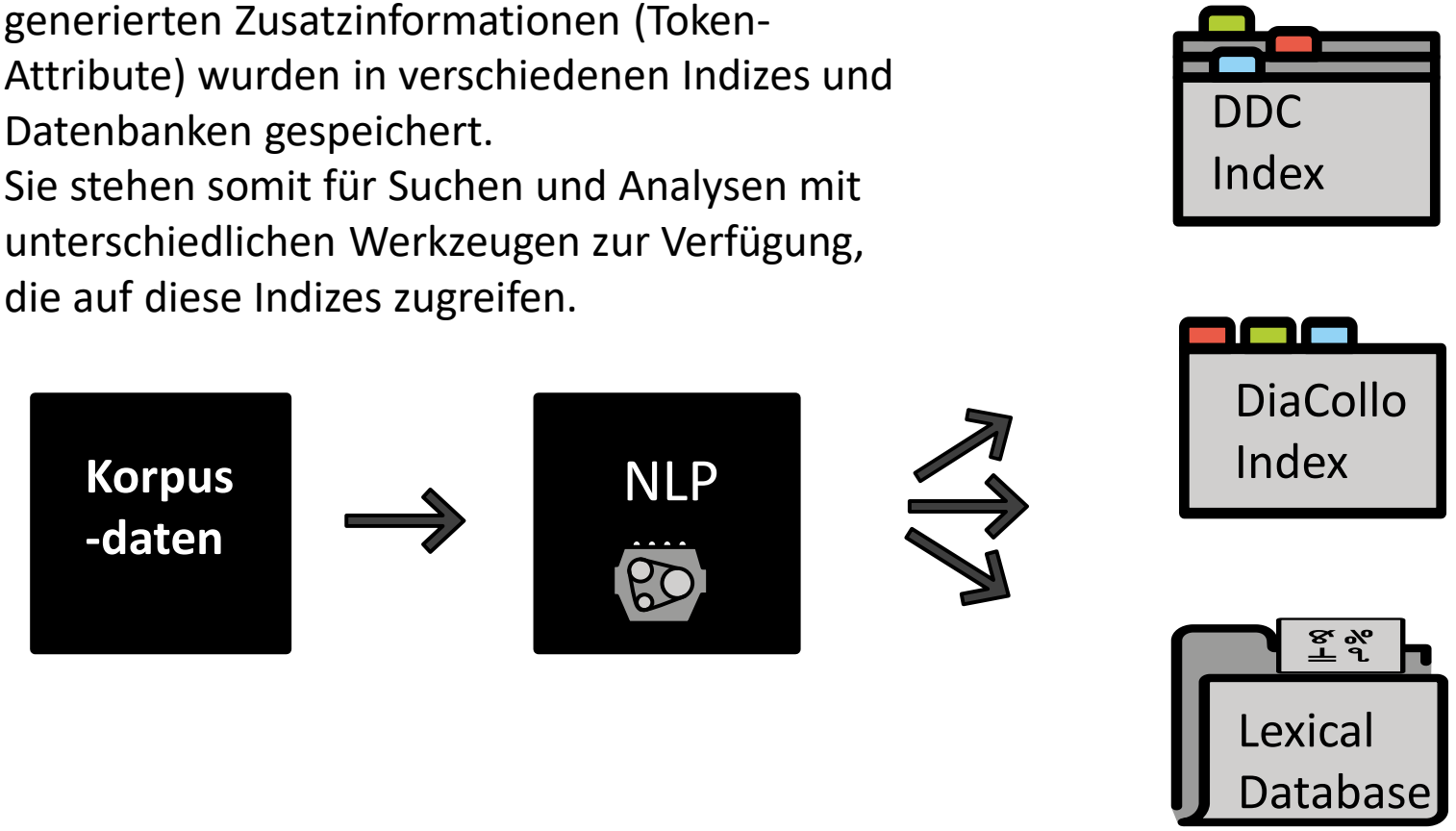
Normalisierung:
- ✓ Weitere Vorverarbeitungsschritte können mit normalisiertem Text durchgeführt werden
- ✓ Suche nach „normalisierter“ Form findet auch historische Formen (z. B. „Tatsachen“ > „Thatsachen“)

Wortartenerkennung (PoS-Tagging):
- ✓ Voraussetzung für weitere Vorverarbeitungsschritte
- ✓ Suchen lassen sich eingrenzen auf bestimmte Wortarten

Lemmatisierung:
- ✓ Suche nach Lemma findet auch verwandte Formen (z. B. „sehen“ findet „gesehen“, „sah“ usw.)

Das GEI-Digital-2020 Korpus: Indexierung

Die Daten, Metadaten und die mit NLP generierten Zusatzinformationen (Token-Attribute) wurden in verschiedenen Indizes und Datenbanken gespeichert.
Sie stehen somit für Suchen und Analysen mit unterschiedlichen Werkzeugen zur Verfügung, die auf diese Indizes zugreifen.



Das GEI-Digital-2020 Korpus: Besonderheiten

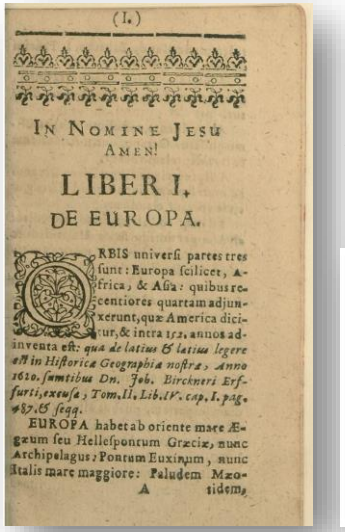
Bei der Nutzung der Analysewerkzeuge und Interpretation der Ergebnisse sind einige Besonderheiten des GEI-Digital-2020 Korpus zu beachten, z. B.

bezüglich der Quellen:

- Disclaimer: Die Texte und Bilder dienen zum Zwecke der Forschung, ihre Inhalte spiegeln nicht die Meinungen der Institutionen und Personen, die an ihrer Zurverfügungstellung beteiligt sind.
- Die Datengrundlage einzelner Publikationszeiträume ist unterschiedlich groß.
- Die Werke wurden für verschiedene Schulfächer, Schulformen (z.B. auch Lehrerbildungsanstalten), Schulstufen etc. geschrieben.
- Einige Werke ähneln sich inhaltlich z. T. stark (z.B. durch Regional-Ausgaben, durch bearbeitete spätere Auflagen usw.)
- Einige Texte enthalten fremdsprachige Abschnitte/Einschübe, und das Korpus vereinzelt fremdsprachige Werke.

bezüglich der Kuration der Daten:

- Die Volltexte weisen aufgrund der rein automatisch durchgeführten Texterkennung eine gewisse Fehlerrate auf.
- In der Konsequenz sind auch die mittels NLP generierten Zusatzinformationen teilweise fehlerhaft/unvollständig (siehe Beispiele unten).
- Die hier vorgestellten Werkzeuge wurden für die Analyse möglichst originalgetreuer Texte konzipiert und konfiguriert.


(I.)

IN NOMINE JESU AMEN!

LIBER I. DE EUROPA.

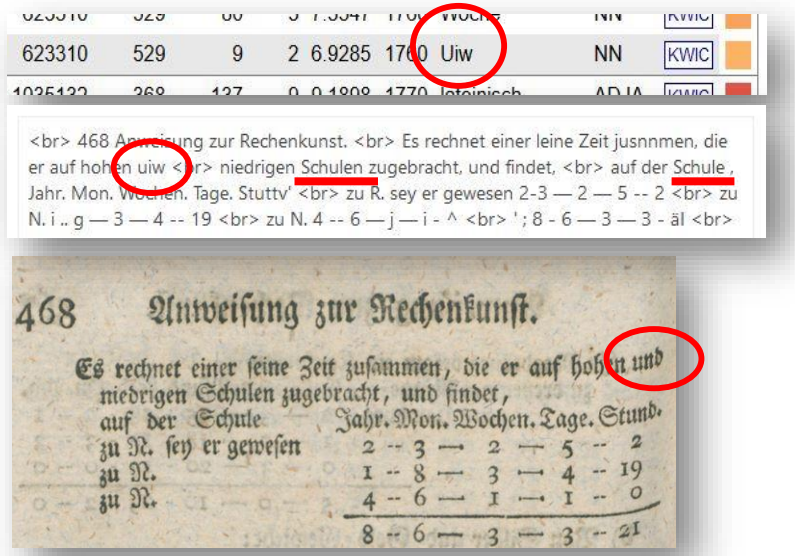
RBIS univerſi partes tres ſunt: Europa ſcilicet, Africa, & Aſia: quibus recentiores quartam adjunxerunt, quæ America dicitur, & intra 1,2. annos adinventa eſt: qua de latius & latius legere eſt in Hiſtorica Geographia noſtra, Anno 1620. ſumtibus Dn. Joh. Birckneri Erfurti, exeuſa; Tom. II. Lib. IV. cap. I. pag. 937. & ſeqq.

EUROPA habet ab oriente mare Ægæum ſeu Helleſpontum Græcix, nunc Archipelagus; Pontum Euxinum, nunc Italis mare maggiore: Paludem Mæo- A tidem,

Beispiel 1: Die früheste mit den digitalen Werkzeugen auffindbare Nennung von „Europa“ im Korpus stammt aus dem *Mercurius Cosmicus* von 1648. Die Nennungen von Afri[k]a, Asien und Ameri[k]a auf derselben Seite werden aufgrund mangelhafter OCR-Ergebnisse/lateinischer Schreibweise nicht gefunden/erkannt.

 (Ю
 o O_O O O O O o oo o
 I ^ Nomine Jesu
 Am en!
 LIBER L
 DE EUROPA.
 RBIS univcrii partes tres
 funt: Europa fcilicet, A,-
 frica> &AG'a: quibus rc-
 centiores quartam adjun-
 xerunt,quae America dici¬
 tur,& intra i/j, anno* ad-
 inventa eft; qua ¿e latitu & Utua legere
 в Я in Hift or i c л Geograf bi л noftra , y.inno
 iCio.fumtibu* Dn. Job. Birckneri Erf-
 furti,ex tu fa , Tomoli» Lib.IK, cap<, I. pag,
 feqq.
 EUROPA habet ab oriente ma-re ЛЕ-
 gzum feu Hellefpontum Grzcix; mmc
 Archipelagus : Pontum Euxinam , nunc
 Stalis шare maggiore: Paludem M*o«
 A tidem#

http://gei-digital.gei.de/viewer/resolver?urn=urn:nbn:de:0220-gd-12794530


623310 529 9 2 6.9285 1760 Uiw NN KWIC

 468 Anweiſung zur Rechenkunſt.
 Es rechnet einer leine Zeit jusnnmen, die er auf hohen uiw
 niedrigen Schulen zugebracht, und findet,
 auf der Schule , Jahr. Mon. Wochen. Tage. Stuttv'
 zu R. sey er gewesen 2-3 — 2 — 5 -- 2
 zu N. i .. g — 3 — 4 -- 19
 zu N. 4 -- 6 — j — i - ^
 ' ; 8 - 6 — 3 — 3 - äl

468 Anweiſung zur Rechenkunſt.

Es rechnet einer ſeine Zeit zuſammen, die er auf hohen und niedrigen Schulen zugebracht, und findet,

	Jahr.	Mon.	Wochen.	Tage.	Stund.
auf der Schule zu N. ſey er geweſen	2	3	2	5	2
zu N.	1	8	3	4	19
zu N.	4	6	1	1	0
	8	6	3	3	21

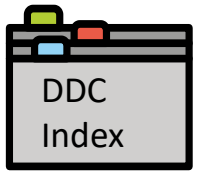
Beispiel 2: DiaCollo interpretiert den OCR-Fehler „uiw“ (statt „und“) als seltenes, aber häufig mit dem gesuchten Stichwort (hier: „Schule“) vorkommendes Wort.

http://gei-digital.gei.de/viewer/resolver?urn=urn:nbn:de:0220-gd-11608143

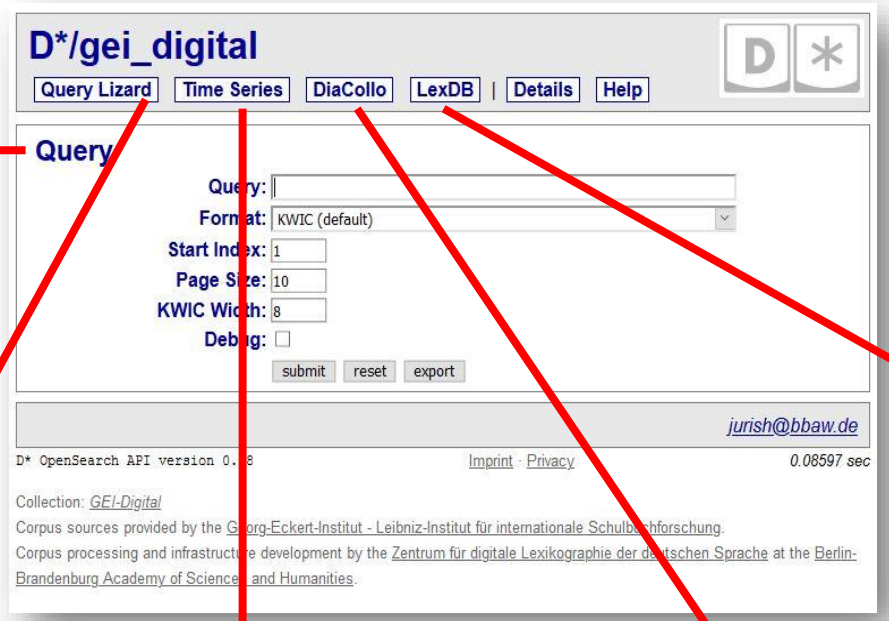
Teil 1:

Die D*-Startseite und die von dort erreichbaren Werkzeuge

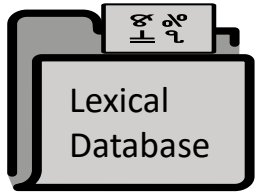
Die D*-Startseite ist zugleich die Startseite für DDC-Korpusabfragen („Query“) und verlinkt auf weitere Werkzeuge für die Korpusanalyse:



Korpusabfragen

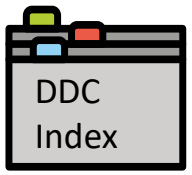


LexDB:
Lexikalische Datenbank

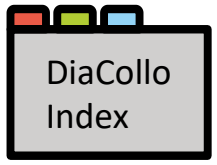


Query Lizard:
Interaktive Ausdehnung/Erweiterung („expansion“) für Suchanfragen

Time Series:
Häufigkeiten im zeitlichen Verlauf



DiaCollo:
Diachrone Kollokationsanalyse



Die Werkzeuge nutzen verschiedene Formen und Teilmengen der mit NLP-Verfahren aufbereiteten und angereicherten Korpusdaten:

D* („Dstar“)	Query	nutzt den DDC Index (=der in D* mit DDC durchsuchbare Hauptindex; Tokens mit einem Type-Index pro indizierten Token-Attribut)
	Query Lizard	nutzt DDC-eigene und ggf. integrierte Tools (wie z. B. GermaNet) für den DDC Index
	Time Series	nutzt eine SQL Datenbank
	DiaCollo	nutzt den nativen DiaCollo Index, nutzt Unigramme, nutzt eine Term Dokument Matrix (TDF index)
	LexDB	nutzt eine SQL Datenbank (komplexe Types)

Auswahl ist mit dem PROFILE-Parameter in DiaCollo möglich

Kombinierbare Werkzeuge

Die Werkzeuge nutzen verschiedene Formen und Teilmengen der mit NLP-Verfahren aufbereiteten und angereicherten Korpusdaten („one tool = one job“). Sie sind jedoch eng miteinander verbunden:

- Eine Ausdehnung/Erweiterung („expansion“) eines Suchbegriffs aus dem Query Lizard kann direkt als Query von DDC ausgeführt werden.

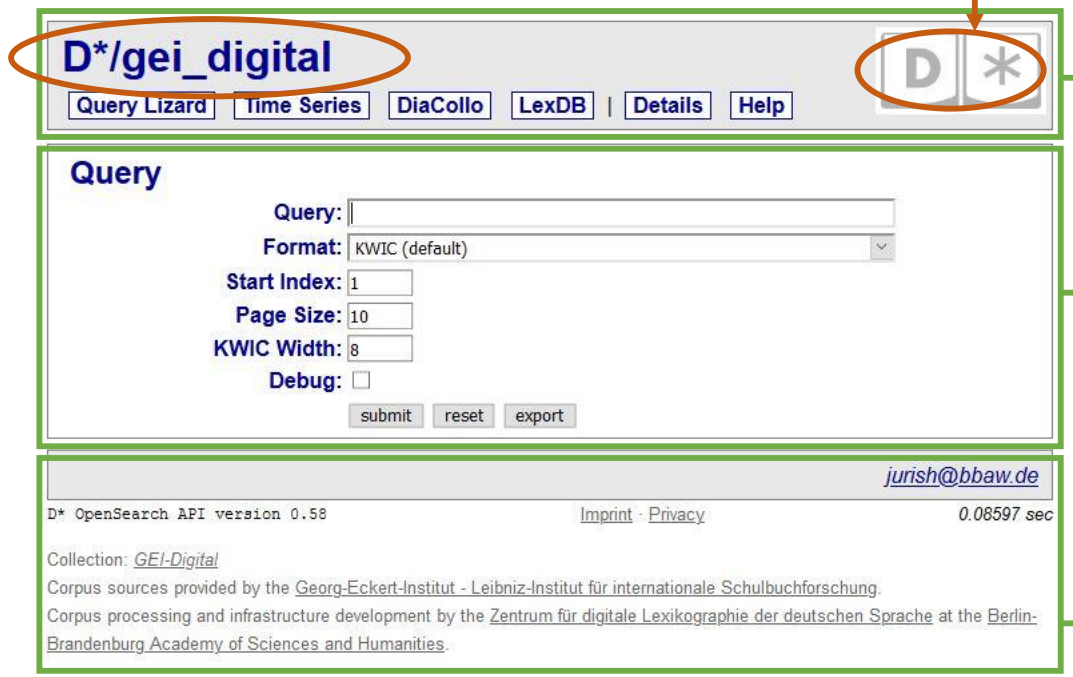
- Eine Ergebnismenge in DiaCollo (aus dem DiaCollo Index) kann zur näheren Ansicht der Stichworte im Kontext (KWIC) als Abfrage an die DDC Query (im DDC Index) „weitergereicht“ werden.

- Manche Berechnungen (z.B. Frequenzanalysen) kann man mit verschiedenen Werkzeugen (d.h. auch in verschiedenen Indizes!) durchführen.

Die Startseite für Korpusabfragen in der D*-Korpusmanagement-Umgebung

Hier wird angezeigt wo Sie sich befinden und welches Korpus ausgewählt wurde. In diesem Fall die [(Startseite der) D* Umgebung mit dem „GEI-Digital 2020“-Korpus]

Ein Klick auf das D* - Logo führt jederzeit zurück auf diese Startseite



Kopfzeile („header“) mit Links zu den verschiedenen Werkzeugen dieser Umgebung

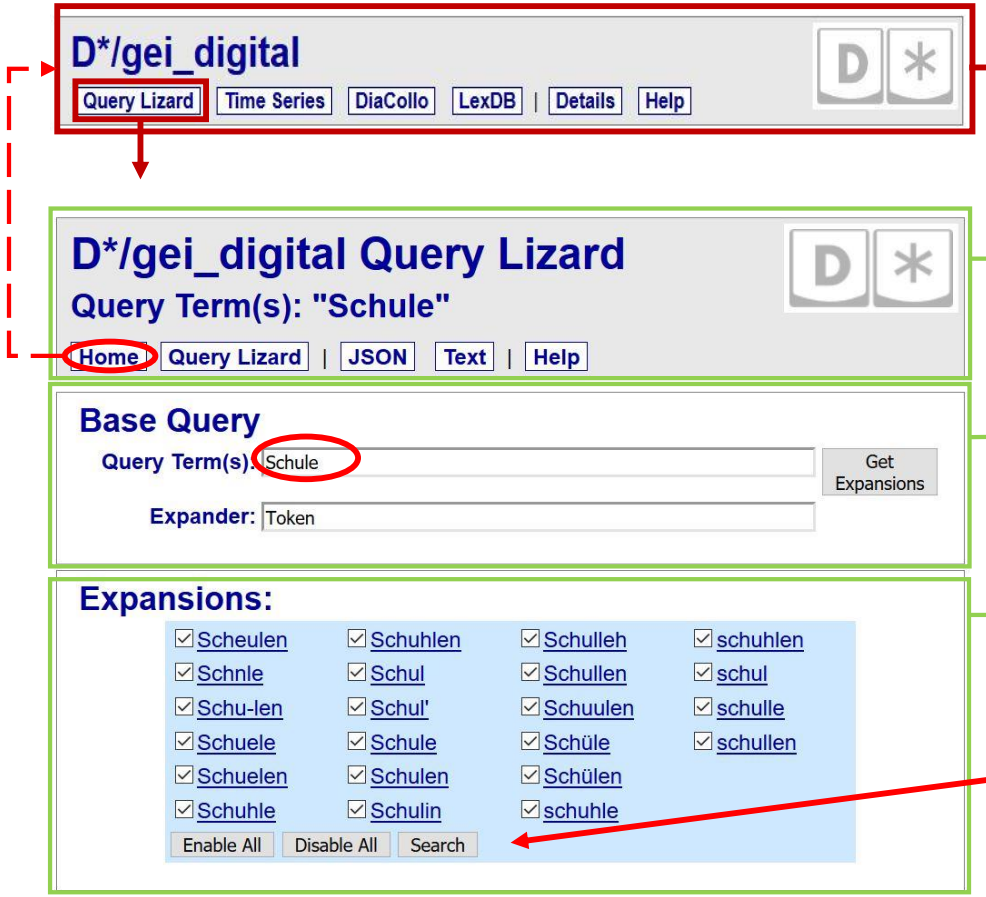
Eingabemaske des jeweils ausgewählten Werkzeugs, in diesem Fall für (DDC-) **Queries** für Suchen in der ausgewählten Textsammlung. Die Werte einiger Parameter sind standardmäßig voreingestellt („default“)

Fußzeile („footer“) mit Impressum, Herkunftsangaben zu Korpus und Werkzeugen u.a.

Mehr zu Korpusabfragen mit **Query** in [Teil 2] dieses Foliensatzes

Im Folgenden werden die in der Kopfzeile verlinkten Werkzeuge vorgestellt.

Query Lizard: Möglichkeit zur interaktiven Erweiterung oder Einschränkung von Suchabfragen



Linkauswahl der Korpusabfrage-Startseite

Kopfzeile mit Linkleiste innerhalb des **Query Lizard**

Eingabemaske; hier mit Beispielanfrage: „Schule“, standardmäßig mit dem Expander „Token“ erweitert. Für Anfragen kann auch die DDC Abfragesprache genutzt werden (vgl. Folie 34 ff)

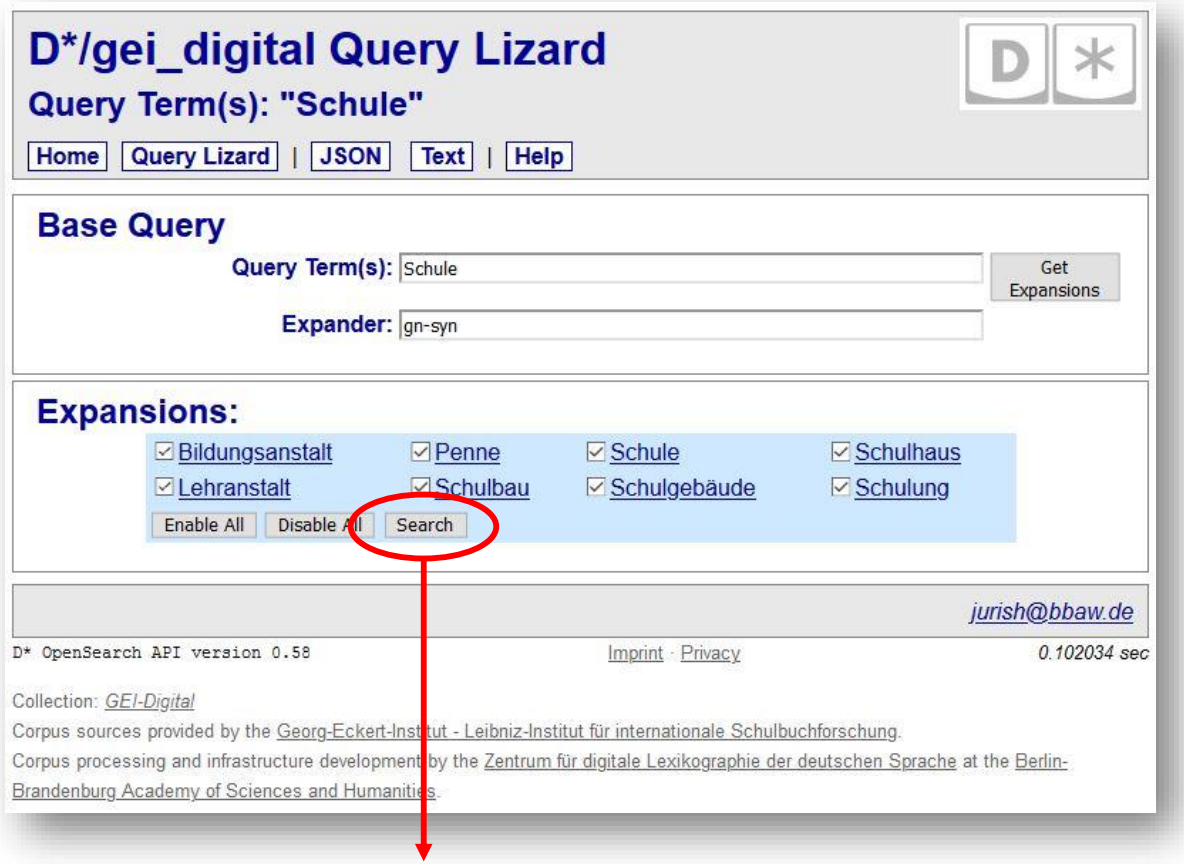
Ergebnisanzeige der Expansion.
Nach den hier gezeigten (bzw. dann ausgewählten) Lemmata wird gesucht, wenn man die Expansion in einer Korpussuche anwendet

- entweder durch Klick auf „Search“ direkt im Query Lizard, oder
- durch Eingabe von Suchbegriff und Expandierung als Suchanfrage in der Startseite der DDC-Korpusabfrage („Query“)

Für eine Liste und Beispiele dieser Expandierungsmöglichkeiten („expansions“) vgl. https://www.dwds.de/d/korpussuche#expansion
Siehe auch: http://diacollo.gei.de/gei-digital-2020/details.perl#expand. Einige Beispiele finden Sie auch auf den folgenden Folien.

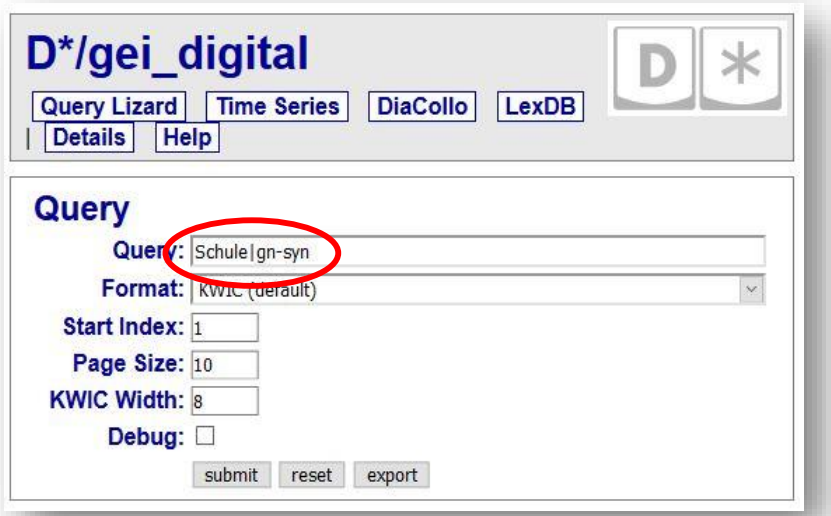
Query Lizard: Nutzung von GermaNet

GermaNet ist ein lexikalisch-semantisches Netz, das deutsche Substantive, Verben und Adjektive semantisch miteinander in Beziehung setzt, z.B. als Synonyme, Über- oder Unterbegriffe. Wenn ein Korpus mit GermaNet annotiert wurde, kann man diese verbundenen Begriffe mit Expandierer-Werkzeugen wie dem Query Lizard finden und nutzen.



Im Beispiel links werden durch gn-syn **Synonyme** des Stichworts „Schule“ angezeigt. Die Expansionen können für die Suche im Korpus übernommen, aber auch „abgewählt“ werden.

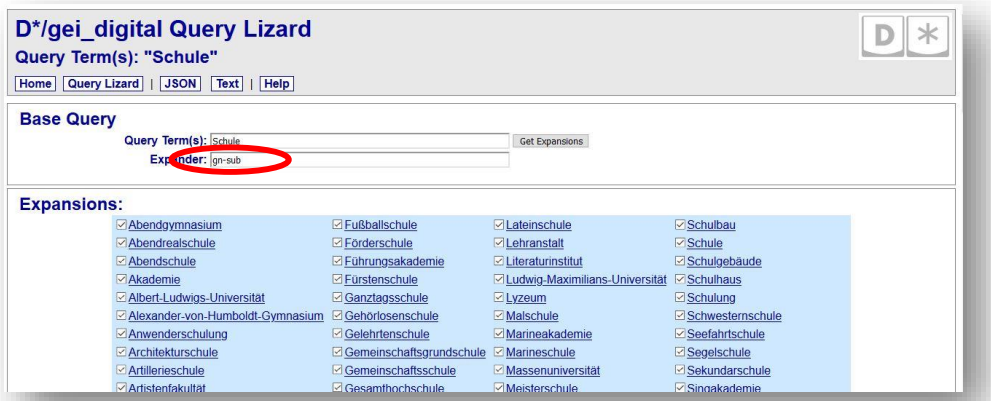
Nutzung dieser Expansionen für eine DDC-Suche:

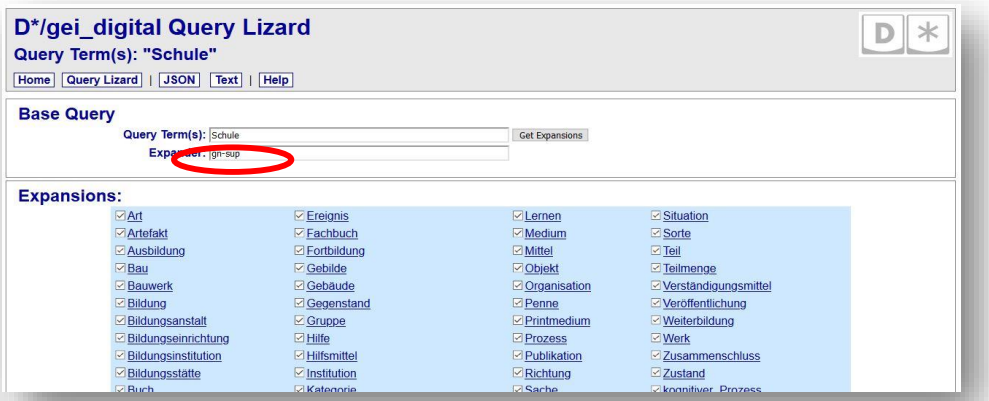


Auf der nächsten Folie finden Sie weitere Beispiele.

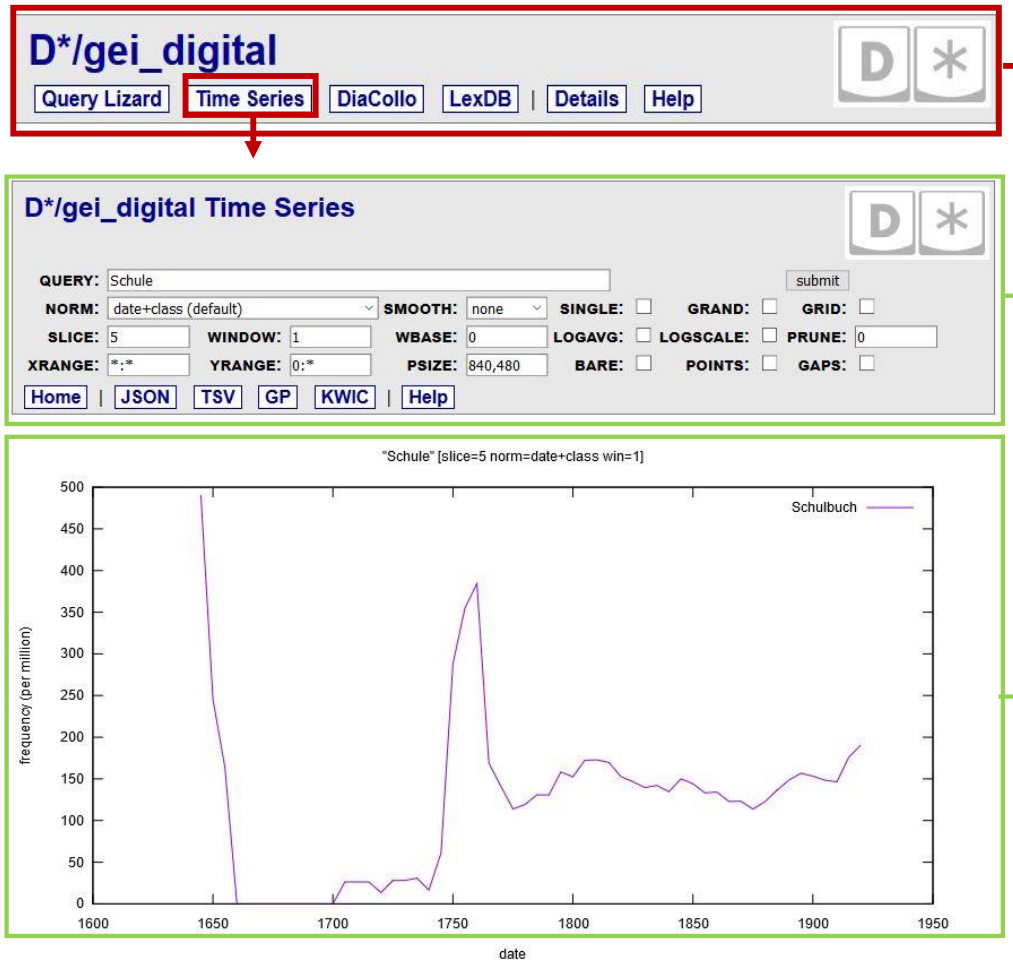
Query Lizard: Beispiele für die Nutzung von GermaNet-Expansionen im GEI-Digital-2020 Korpus:

gn-syn	Synonyme	Beispiel „Schule“
gn-syn1	Synonyme, direkte Unter- und Oberbegriffe	Beispiel „Schule“
gn-syn2	Synonyme, direkte und deren Unter- und Oberbegriffe	Beispiel „Schule“
gn-sub	Synonyme und Unterbegriffe	Beispiel „Schule“
gn-sub1	Synonyme und direkte Unterbegriffe	Beispiel „Schule“
gn-sub2	Synonyme, alle direkten und deren Unterbegriffe	Beispiel „Schule“
gn-sup	Synonyme und Oberbegriffe	Beispiel „Schule“
gn-sup1	Synonyme und direkte Oberbegriffe	Beispiel „Schule“
gn-sup2	Synonyme, alle direkten und deren Oberbegriffe	Beispiel „Schule“





Time Series: Für Frequenzanalysen im zeitlichen Verlauf



Linkauswahl der Korpusabfrage-Startseite

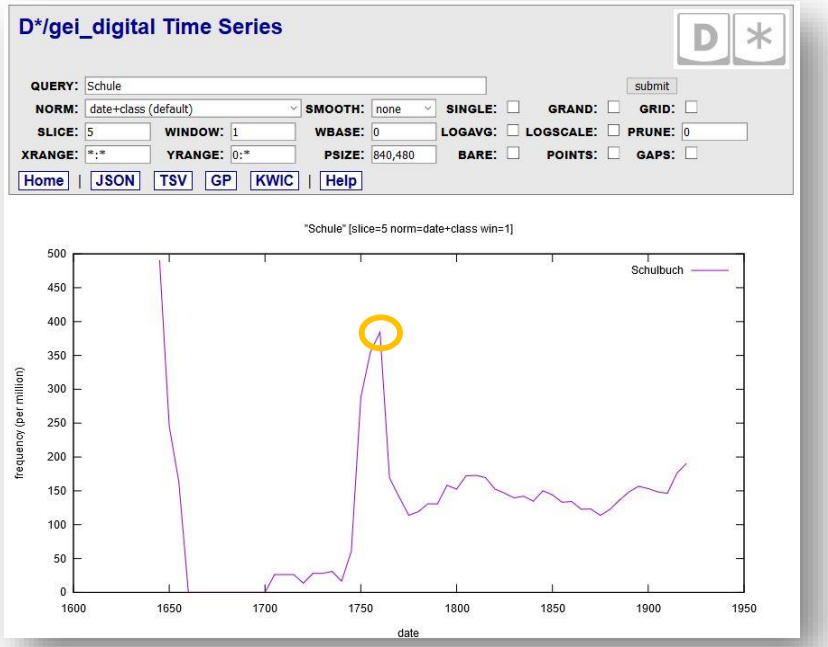
Eingabemaske und Links innerhalb des **Time Series**-Werkzeugs (hier mit den Standardeinstellungen der Parameter bzw. ihrer Werte und der Beispielabfrage „Schule“).
Anpassen lassen sich die Werte von Parametern wie z.B. der gesamte untersuchte Zeitbereich (**XRANGE**), darin untersuchte Zeit-Intervalle (**SLICE**), Fenstergröße für gleitende Mittelwertglättung (**WINDOW**), Konfidenzintervall für Ausreißererkennung (**PRUNE**) und graphische Glättungsmethode (**SMOOTH**).

Verlaufskurve als Ergebnisanzeige der Anfrage.
Die Darstellung ist in dieser voreingestellten Sicht für eine leichtere Interpretierbarkeit „geglättet“: Frequenzwerte werden relativ zur Gesamtanzahl der Token (f pro Million) angezeigt und per Default-Voreinstellung nicht jahresgenau, sondern in Abschnitten von jeweils 5 Jahren (**SLICE:** 5) gemessen und mit einem gleitenden Durchschnitt über die unmittelbare benachbarten Abschnitte (**WINDOW:** 1) geglättet dargestellt.

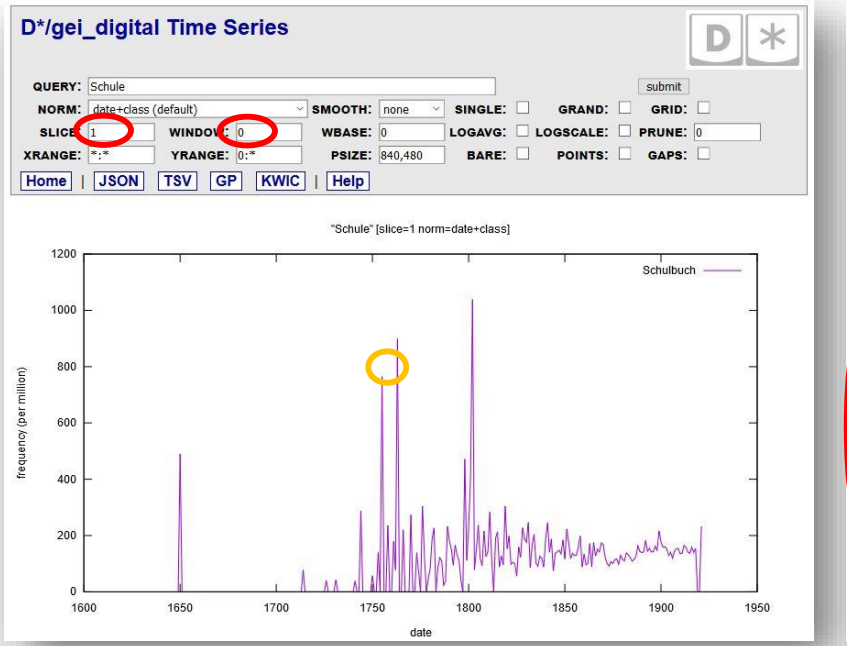
Die Time Series nutzt (ebenso wie LexDB, aber anders als die DDC-Query und DiaCollo) eine SQL Datenbank der Korpusdaten.

Time Series: Hinweise zur Interpretation – Datengrundlage, Glättung und Ausreißer

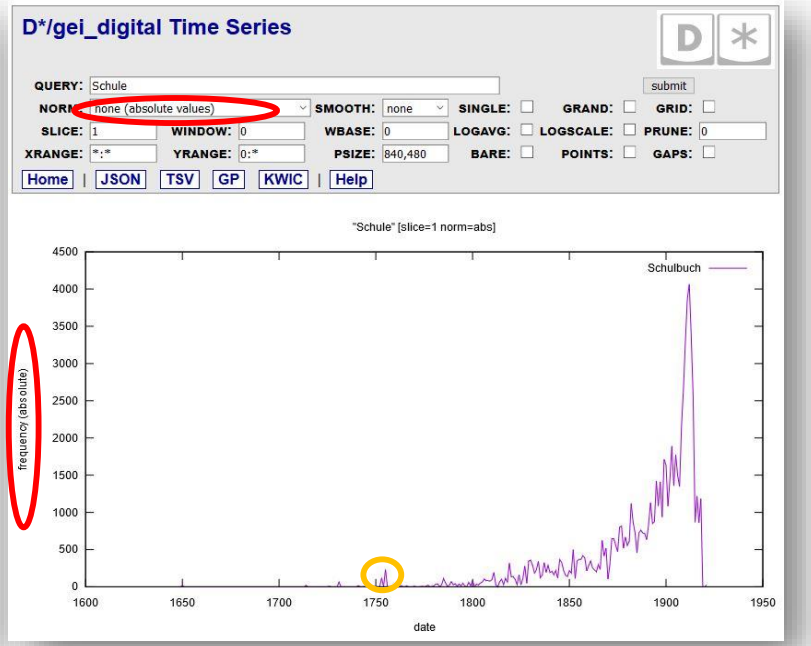
Die Ergebnisse der Time Series sind umso aussagekräftiger (für das jeweilige Korpus), je umfangreicher und zeitlich ausgewogener zusammengesetzt das Korpus ist. Im GEI-Digital-2020 Korpus sind – auch aufgrund der historisch veränderlichen Schulbuchproduktion – deutlich weniger Daten für das 17. und 18. Jhd. als für das späte 19. und frühe 20. Jhd. vorhanden. Generell können bei kleiner(er) Datengrundlage besonders bei mittel- und niedrigfrequenten Wörtern Ausreißer in Einzelwerken dazu führen, dass die Wortverläufe verzerrt dargestellt werden.



Verlaufskurve für das Stichwort „Schule“ mit Standardglättung, also optimiert für die leichte Erkennbarkeit von Trends in großen, ausgewogenen Korpora. Verzeichnet ist z. B. ein „Peak“ der Vorkommen ca. 1750-70.



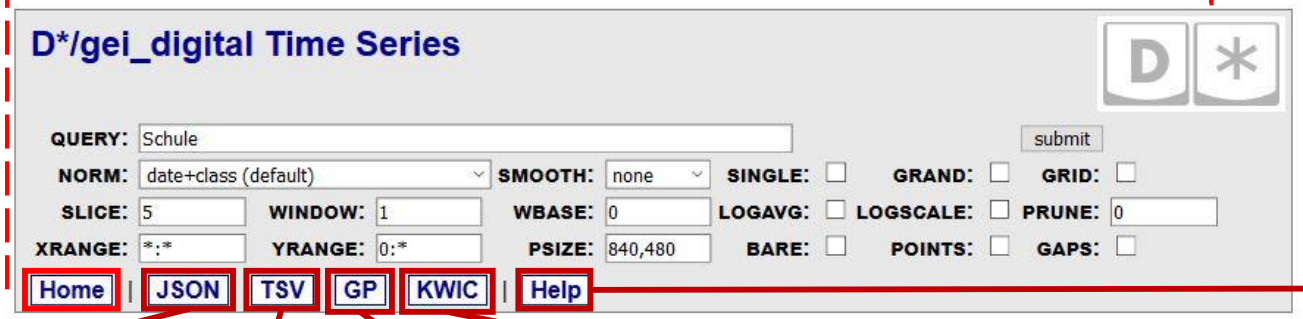
Ergebnisanzeige derselben Anfrage mit veränderten Parameter-Werten (hier: **SLICE**: 1, **WINDOW**: 0 für eine Anzeige **ohne Glättung**): Der „Peak“ um 1750-70 ist verursacht von zwei Ausreißer-Jahren.

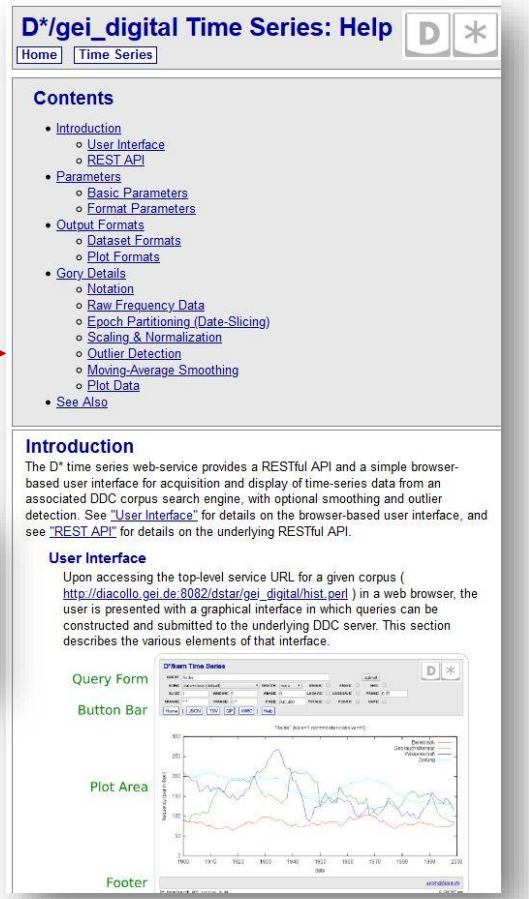


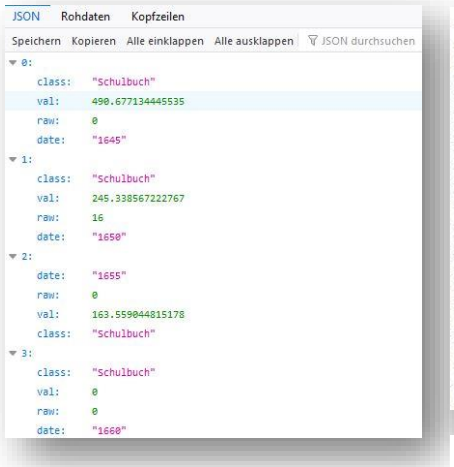
Ergebnisanzeige derselben Anfrage mit veränderten Parameter-Werten (hier: **NORM**: none für eine Anzeige der **absoluten** Häufigkeit, SLICE: 1, **WINDOW** 0 für eine Anzeige **ohne Glättung**): Die absolute Häufigkeit von „Schule“ im Korpus in den Jahren 1750-70 ist nicht sehr groß.

Time Series:
Ausgabeformate und Doku/Hilfe

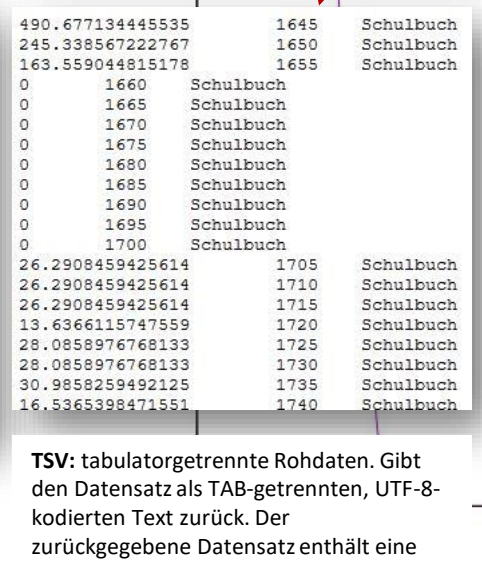
Link zurück zur Korpusabfrage-Startseite



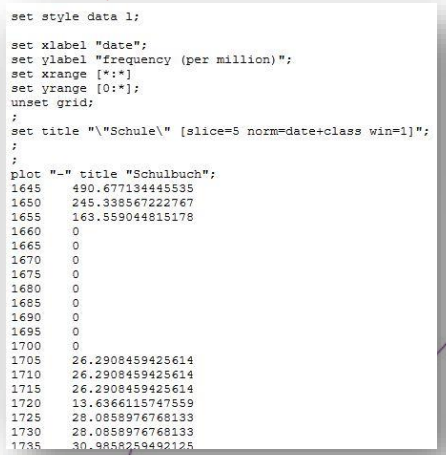




Jason: Gibt den Datensatz kodiert als flaches JSON-Array zurück.



TSV: tabulatorgetrennte Rohdaten. Gibt den Datensatz als TAB-getrennten, UTF-8-kodierten Text zurück. Der zurückgegebene Datensatz enthält eine Zeile für jeden Datenpunkt, und jede Zeile ist in drei TAB-getrennte Spalten unterteilt. Die erste Spalte ist der (geglättete) Häufigkeitswert (y), die zweite Spalte ist die Epochenbezeichnung (x) und die letzte Spalte ist die zugehörige Textklasse (z).



GP: Gnuplot Script:
Gibt ein eigenständiges Gnuplot-Skript mit Datenblock(s) für die geglätteten Daten zurück. Wird intern verwendet, um die Diagrammformate zu erzeugen.

KWIC: Keyword in Context. Link zurück zur Ergebnisseite einer Korpussuche nach dem Stichwort (hier „Schule“).

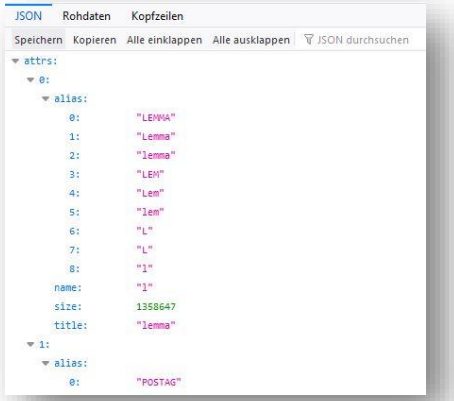
Help: Hilfe und Dokumentation zu diesem Werkzeug in englischer Sprache

Vgl. hierzu auch:
https://www.dwds.de/d/plot

DiaCollo: Interaktive Berechnung und Visualisierung von Kollokationen über die Zeit

Linkauswahl der Korpusabfrage-Startseite

Integrierte Eingabemaske und Links innerhalb von „DiaCollo“ (hier: mit den Standardeinstellungen der wählbaren Parameter, ohne Beispielsuchanfrage)



Details zum DiaCollo-Index

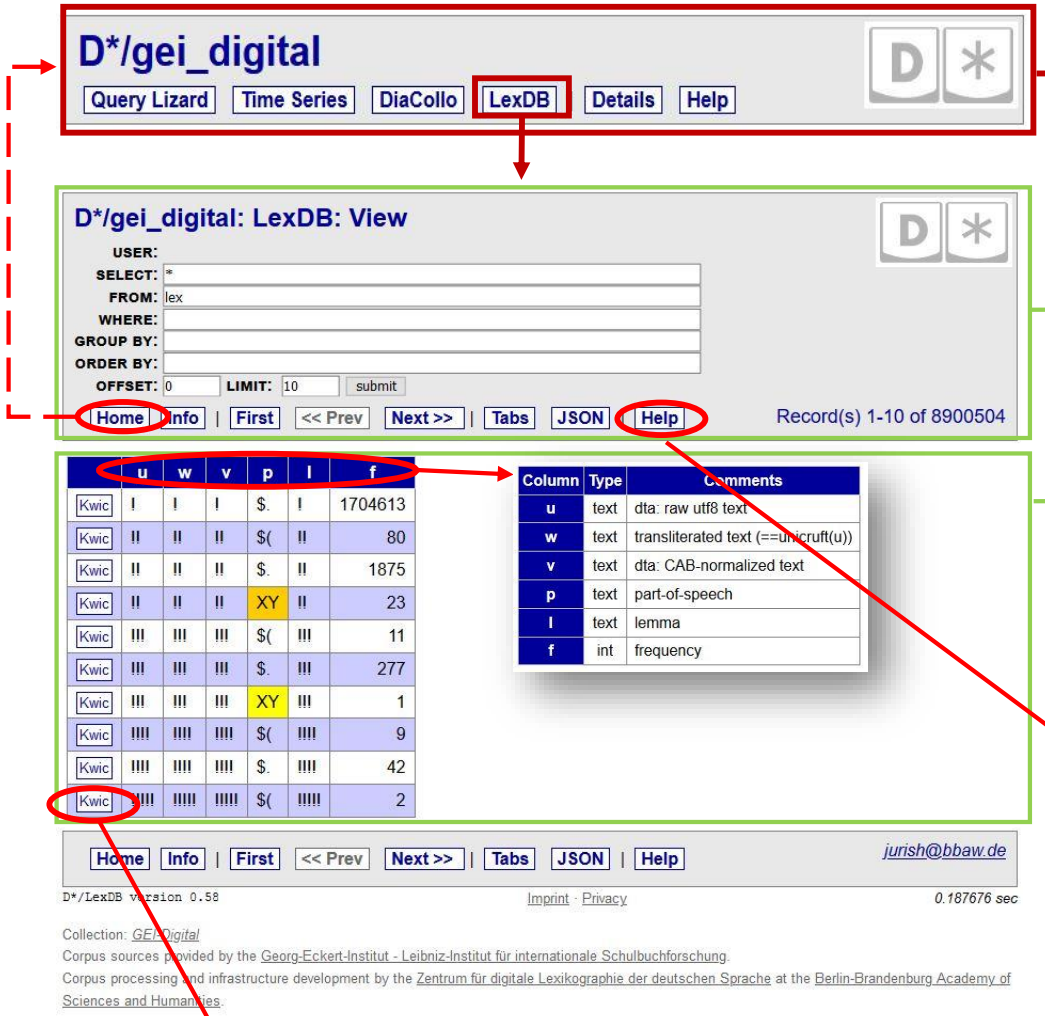
Erklärung und Definitionen aller Parameter

Deutschsprachiges Tutorial für das DiaCollo-Werkzeug

Mehr zu **DiaCollo** in Teil 3 dieses Foliensatzes

LexDB: Lexikalische Datenbank zu allen Attributen der Token sowie Frequenzen im jeweils untersuchten Korpus

vgl. https://www.dwds.de/d/korpussuche#lexdb und https://www.dwds.de/r/lexdb



Linkauswahl der Korpusabfrage-Startseite

Integrierte Eingabemaske und Links innerhalb der **LexDB**; hier mit Standardabfrage (*=alles) und Standardwerten der wählbaren Parameter

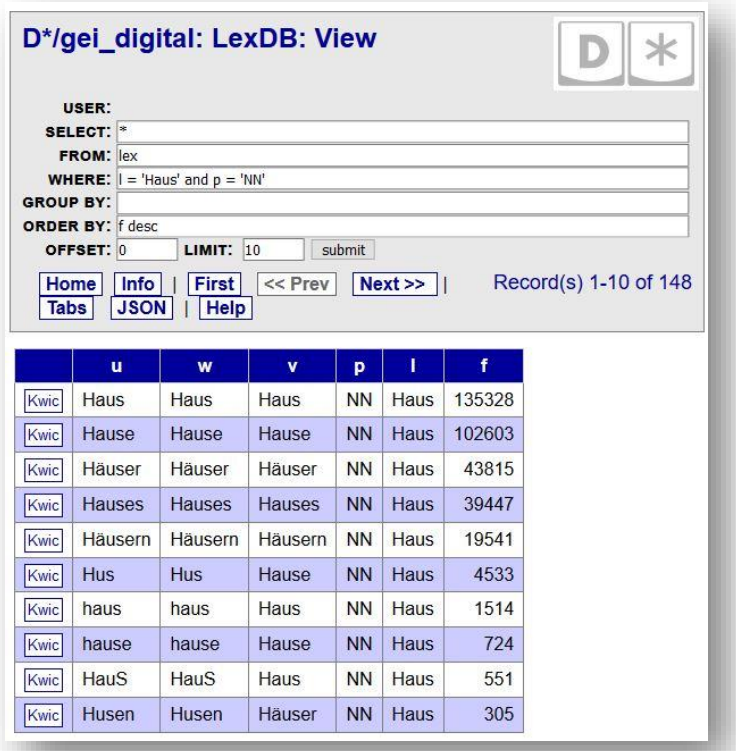
Ergebnisanzeige der Standard-Anfrage (hier die ersten 10 Ergebnisse von 8900504 im GEI-Digital-2020 Korpus, in diesem Fall zunächst eine Reihe Sonderzeichen/Nicht-Worte). Gezeigt werden das Rohdatum (Spalte **u**), der transliterierte Text (**w**), der normalisierte Text (**v**), Wortarten (**p**), Lemma (**l**) und Häufigkeit (**f**).

LexDB ist eine relationale SQLite-Datenbank; Hinweise zur Abfragesprache finden Sie unter **Help**

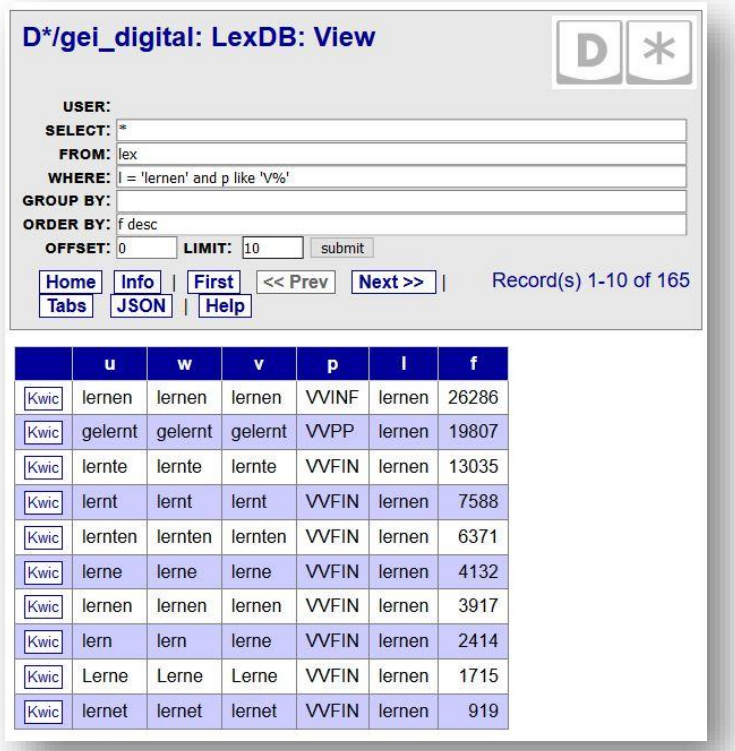
KWIC = Link zur „Stichwort-im Kontext“-Ergebnisanzeige einer DDC-Suche nach dem jeweiligen komplexen Type (= logische Konjunktion über alle Attributspalten der Reihe)

LexDB Grundlagen

In der LexDB sind alle Attribute zusammen indiziert, so dass eine Abfrage nach der Häufigkeit eines (komplexen) Types, wie z.B. „*alle Formen des Lemmas ‚Haus' als Nomen, absteigend nach Frequenz sortiert*" oder „*alle Formen des Lemmas ‚lernen' als Verb (nicht "lernend/ADJA" oder "Lernen/NN"), absteigend nach Frequenz sortiert*" schnell beantwortet werden kann. Die Datenbank beinhaltet sozusagen bereits alle Ergebnisse für alle möglichen Anfragen.



	u	w	v	p	l	f
Kwic	Haus	Haus	Haus	NN	Haus	135328
Kwic	Hause	Hause	Hause	NN	Haus	102603
Kwic	Häuser	Häuser	Häuser	NN	Haus	43815
Kwic	Hauses	Hauses	Hauses	NN	Haus	39447
Kwic	Häusern	Häusern	Häusern	NN	Haus	19541
Kwic	Hus	Hus	Hause	NN	Haus	4533
Kwic	haus	haus	Haus	NN	Haus	1514
Kwic	hause	hause	Hause	NN	Haus	724
Kwic	HauS	HauS	Haus	NN	Haus	551
Kwic	Husen	Husen	Häuser	NN	Haus	305



	u	w	v	p	l	f
Kwic	lernen	lernen	lernen	VVINF	lernen	26286
Kwic	gelernt	gelernt	gelernt	VVPP	lernen	19807
Kwic	lernte	lernte	lernte	VVFIN	lernen	13035
Kwic	lernt	lernt	lernt	VVFIN	lernen	7588
Kwic	lernten	lernten	lernten	VVFIN	lernen	6371
Kwic	lerne	lerne	lerne	VVFIN	lernen	4132
Kwic	lernen	lernen	lernen	VVFIN	lernen	3917
Kwic	lern	lern	lerne	VVFIN	lernen	2414
Kwic	Lerne	Lerne	Lerne	VVFIN	lernen	1715
Kwic	lernet	lernet	lernet	VVFIN	lernen	919

Mit DDC dauert eine solche Abfrage länger, weil alle Tokens der angefragten Stichworte und Attribute durchgezählt werden müssen, um die erfragte Token-Schnittmenge zu ermitteln.

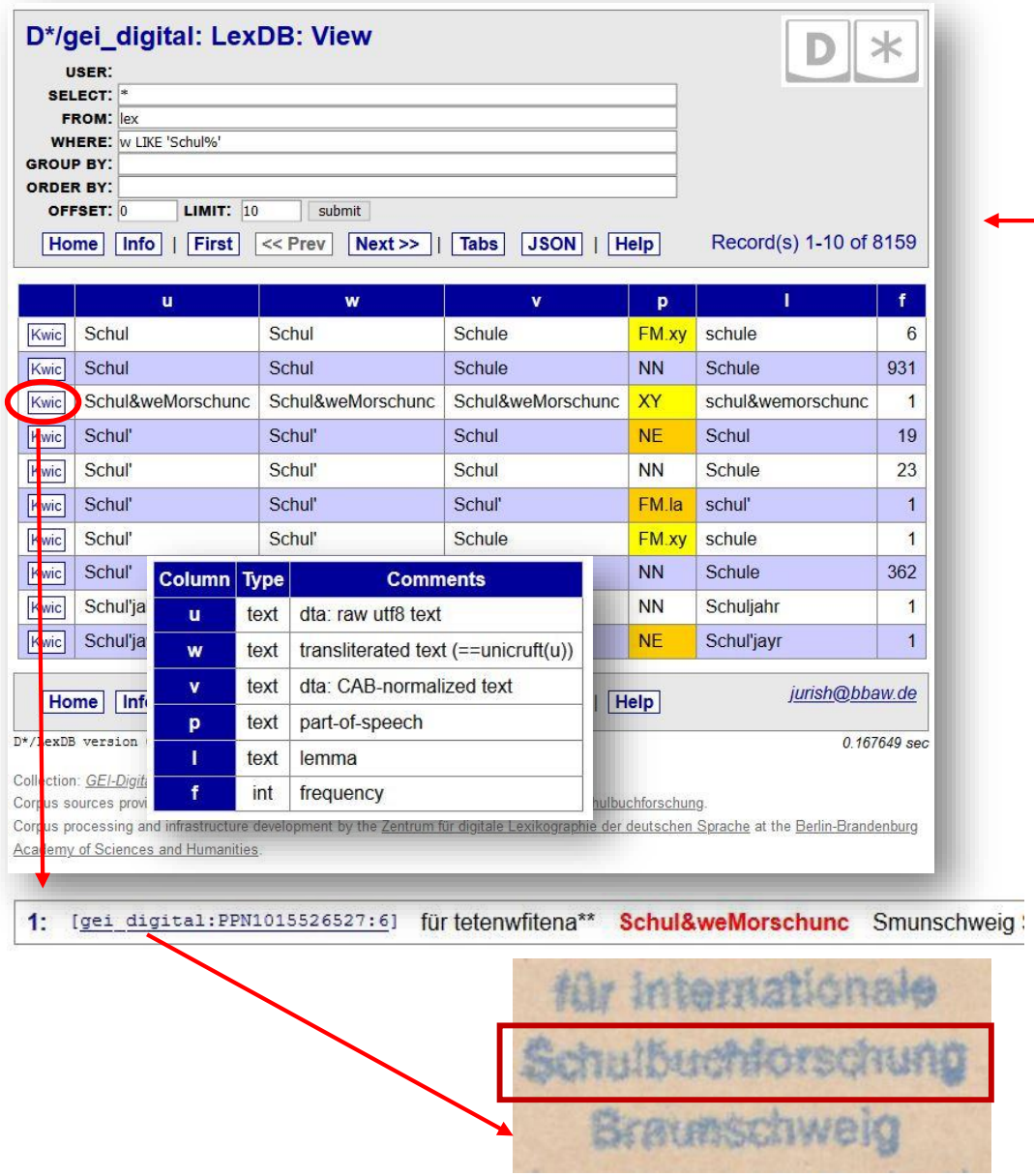
Column	Type	Comments
u	text	dta: raw utf8 text
w	text	transliterated text (==unicruft(u))
v	text	dta: CAB-normalized text
p	text	part-of-speech
l	text	lemma
f	int	frequency

LexDB ist nützlich, um das Vokabular eines Korpus' in den Blick zu nehmen:

- Welche Lemmata (l) werden einer bestimmten vorkommenden Wortform (u) zugewiesen?
- Welche PoS-Tags (p) werden einem bestimmten Lemma (l) oder einer Oberflächenform (u) zugewiesen?
- Wie häufig kommt ein bestimmter Begriff in einem Korpus vor (f)?

Informationen über Wortverbindungen, oder über Publikationsjahr und andere Metadaten sind in der LexDB nicht enthalten.

Beispielabfrage in der LexDB: Wortformen



Beispiel: "Liste *Alle Types dieses Korpus', die mit ‚Schul-' anfangen und die dazugehörigen Attribute*": FROM:`lex` WHERE:`w LIKE 'Schul%'`

Dieselbe Anfrage als (rechenintensivere) DDC-Abfrage sähe so aus:
`count(Schul* #sep)#by[$u,$w,$v,$p,$l]`

D*/gei_digital Counts
Rows 1 - 10 of at most 11332
Home | Query Lizard | Previous | Next | Help count(Schul* #sep) #by[$u,$w,$v,$p,$l] submit

count	key1	key2	key3	key4	key5
6	Schul	Schul	Schule	FM.xy	schule
931	Schul	Schul	Schule	NN	Schule
1	Schul&weMorschunc	Schul&weMorschunc	Schul&weMorschunc	XY	schul&wemorschunc
19	Schul'	Schul'	Schul	NE	Schul
23	Schul'	Schul'	Schul	NN	Schule
1	Schul'	Schul'	Schul'	FM.la	schul'
1	Schul'	Schul'	Schule	FM.xy	schule
362	Schul'	Schul'	Schule	NN	Schule
1	Schul'jahr	Schul'jahr	Schuljahr	NN	Schuljahr
1	Schul'jayr	Schul'jayr	Schul'jayr	NE	Schul'jayr

Eine Abfrage wie im Beispiel links kann auch dazu dienen zu ermitteln, welche (möglicherweise falschen) Oberflächenformen von der automatischen Texterkennung (OCR) erkannt wurden.
In Zeile 3 wurde z. B. das einmal (f=1) im Korpus auftretende Token „Schul&weMorschunc" (u) als Wortart „Nichtwort" (p= XY) klassifiziert (vgl. hierzu das verwendete STTS-Tagset zur Bezeichnung der Wortarten; Guidelines PDF). Durch die Verlinkung zum Stichwort im Kontext (KWIC) und von dort zur Original(digital)quelle lässt sich nachvollziehen, dass es sich hierbei um einen Fehler der automatischen Texterkennung handelt.

Beispielabfrage in der LexDB: Häufigste Wörter im Korpus

D*/gei_digital: LexDB: View

USER:
SELECT: w,p,l,sum(f) as freq
FROM: lex
WHERE: P= 'NN'
GROUP BY: l
ORDER BY: freq desc
OFFSET: 0 LIMIT: 10 submit

	w	p	l	freq
Kwic	HLAND	NN	Land	676495
Kwic	HJahrh	NN	Jahr	604874
Kwic	Allerstadt	NN	Stadt	581568
Kwic	Coening	NN	König	578028
Kwic	CEiT	NN	Zeit	574451
Kwic	Allerdag	NN	Tag	424739
Kwic	G-Ott	NN	Gott	424700
Kwic	DEll	NN	Teil	419008
Kwic	Folck	NN	Volk	412245
Kwic	Allermannes	NN	Mann	410659

Die 10 häufigsten Nomen im Korpus

D*/gei_digital: LexDB: Viev

USER:
SELECT: p,l,sum(f) as freq
FROM: lex
WHERE: P= 'ADJA'
GROUP BY: l
ORDER BY: freq desc
OFFSET: 0 LIMIT: 10 submit
Home Info | First << Prev

	p	l	freq
Kwic	ADJA	groß	1081037
Kwic	ADJA	deutsch	625766
Kwic	ADJA	ander	554587
Kwic	ADJA	alt	463826
Kwic	ADJA	neu	426020
Kwic	ADJA	ganz	409702
Kwic	ADJA	klein	399312
Kwic	ADJA	hoch	386161
Kwic	ADJA	erst	357264
Kwic	ADJA	gut	329388

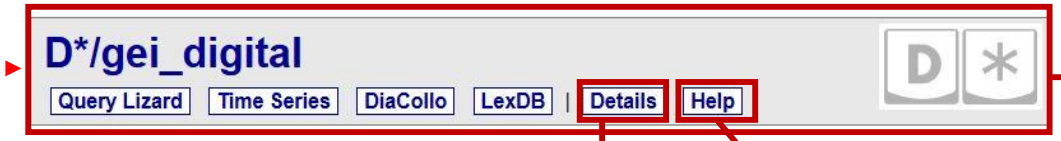
D*/gei_digital: LexDB: V

USER:
SELECT: p,l,sum(f) as freq
FROM: lex
WHERE: P= 'NE'
GROUP BY: l
ORDER BY: freq desc
OFFSET: 0 LIMIT: 10
Home Info | First << Prev

	p	l	freq
Kwic	NE	Friedrich	311455
Kwic	NE	Deutschland	290005
Kwic	NE	I.	250863
Kwic	NE	Frankreich	217528
Kwic	NE	Karl	202949
Kwic	NE	Wilhelm	173819
Kwic	NE	Ii.	171306
Kwic	NE	Heinrich	166146
Kwic	NE	F	165407
Kwic	NE	Italien	152067

Analog dazu: die häufigsten Lemmata für Adjektive und Eigennamen

D*: Details und Hilfe



Linkauswahl der Korpusabfrage-Startseite

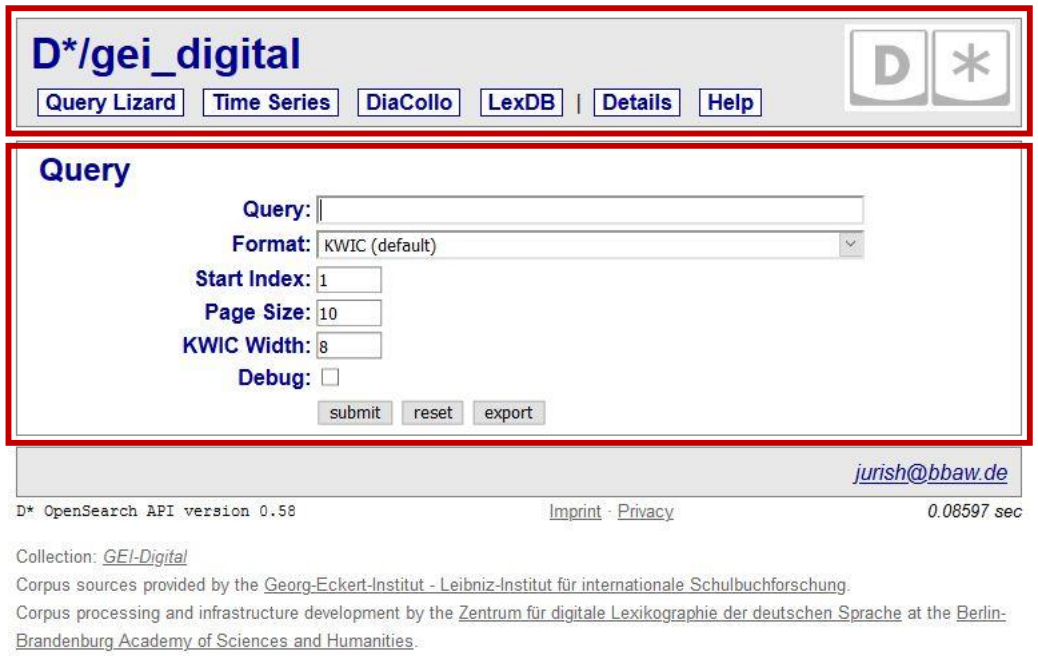
Details: Informationen zum Status des DDC Servers und zum jeweiligen Index (hier: „gei_digital“ des „GEI-Digital 2020“- Korpus)

Help: Hilfe für die Nutzung der DDC Abfragesprache. Für eine deutschsprachige Beschreibung und weitere Beispiele siehe auch: https://www.dwds.de/d/korpussuche sowie Teil 2 dieses Foliensatzes

Teil 2:

D*/Query – Parameter, Ergebnisansichten und Beispielanfragen

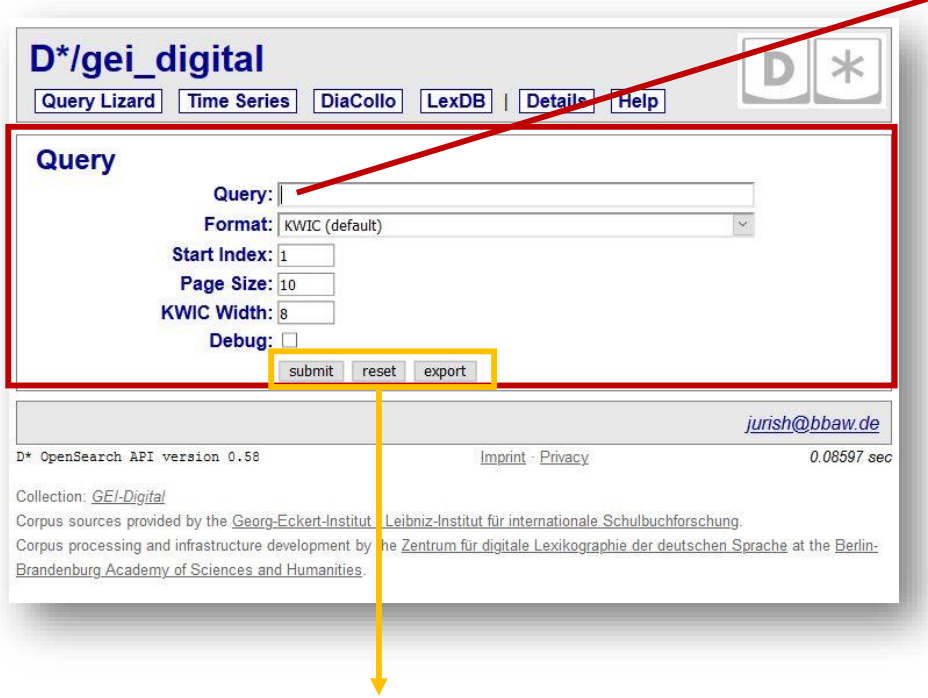
Zur Erinnerung:
Dies ist die Startseite für Korpusabfragen in der D*-Korpusmanagement-Umgebung:



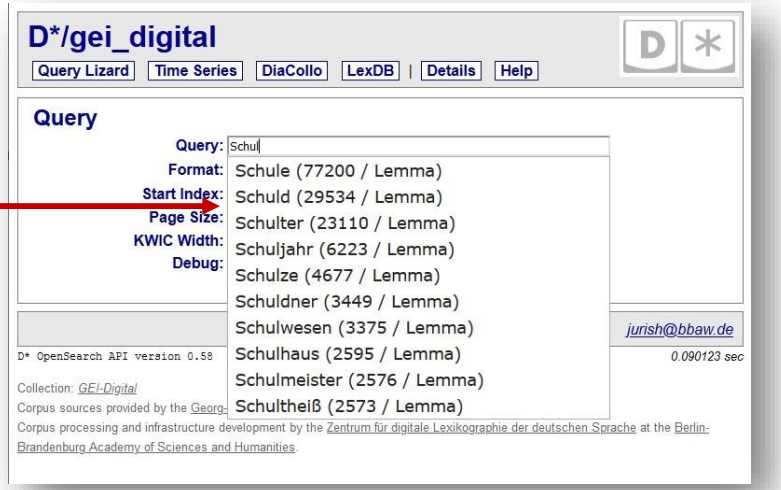
In Teil 1 dieses Foliensatzes wurden die in dieser Kopfzeile der Startseite verlinkten Werkzeuge vorstellt.

In diesem Teil des Foliensatzes werden nun einige der wählbaren Parameter der **Query** - Eingabemaske, verschiedene Ergebnisansichten und Beispiele für einfache und komplexe Suchanfragen vorgestellt.

Die Parameter der Eingabemaske



Query: Eingabe einer Suchanfrage (mit Vorschlagsfunktion aus der LexDB).
Die Anfragen müssen den Regeln der DDC-Abfragesyntax entsprechen, was auch bei einfachen Stichwörtern und Regulären Ausdrücken der Fall ist.



Format: Wahlmöglichkeiten für das Ausgabeformat der Ergebnisse bzw. die Anzeigen der Treffer; Standard-Voreinstellung ist „KWIC“ (= Keyword in Context, Stichwort im Kontext)

Submit: Anfrage abschicken

Reset: Eingabefelder leeren/auf Standard-Voreinstellungen zurücksetzen

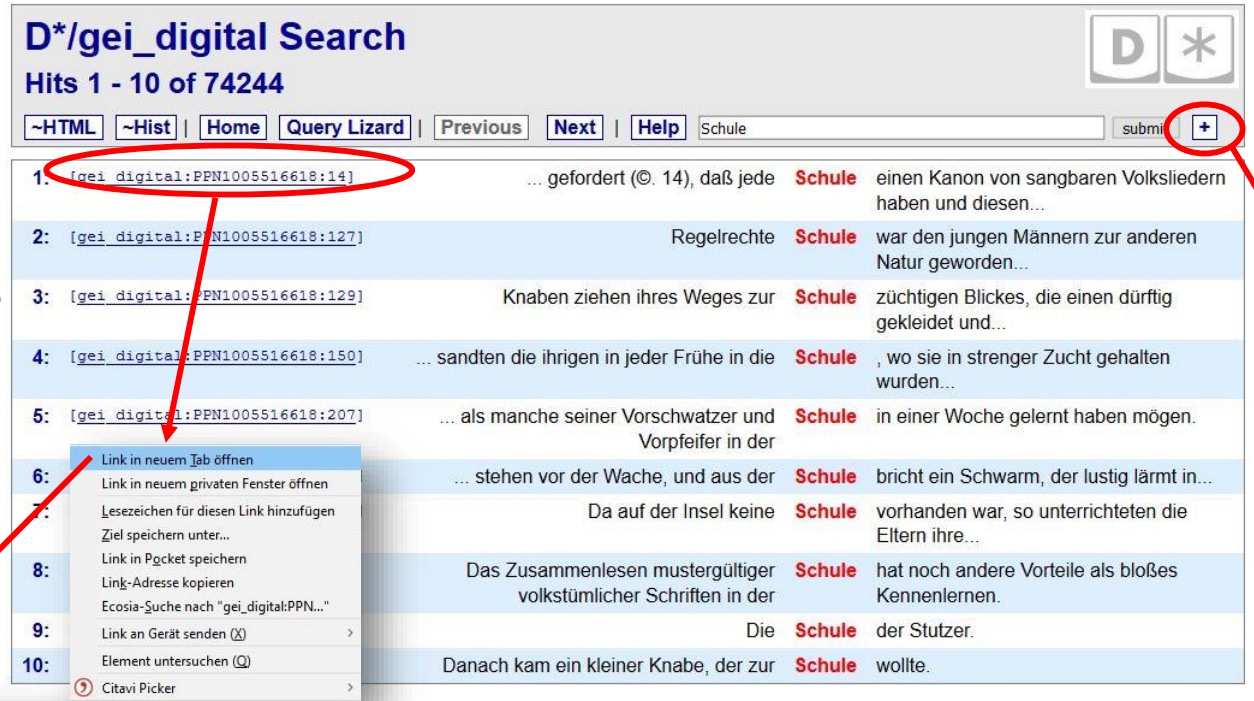
Export: Exportieren der Treffer in ein (unter **Format**) wählbares Dateiformat

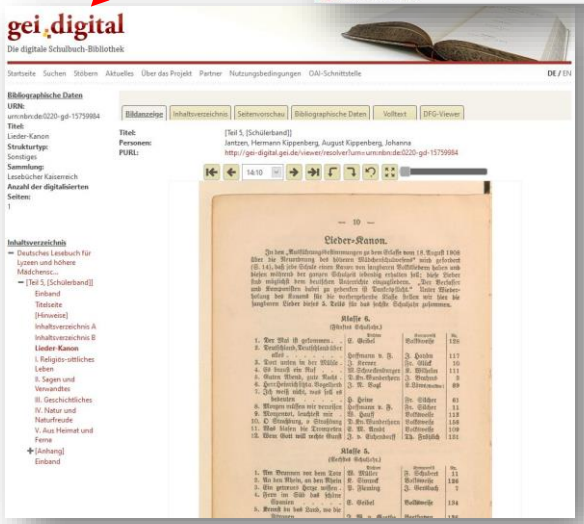
Page Size: Hier kann man angeben, wie viele Treffer auf einer Seite in der Ergebnisansicht angezeigt werden sollen; Standard-Voreinstellung ist 10

KWIC Width: Wahlmöglichkeiten für die Länge der anzuzeigenden Textumgebung rund um das gesuchte Stichwort in der Ergebnisansicht; Standard-Voreinstellung sind 8 Wörter vor und nach dem Treffer

Ergebnisansicht einer D*-Suchanfrage: Standard-Anzeigeformat „KWIC“ und Link zum Digitalisat

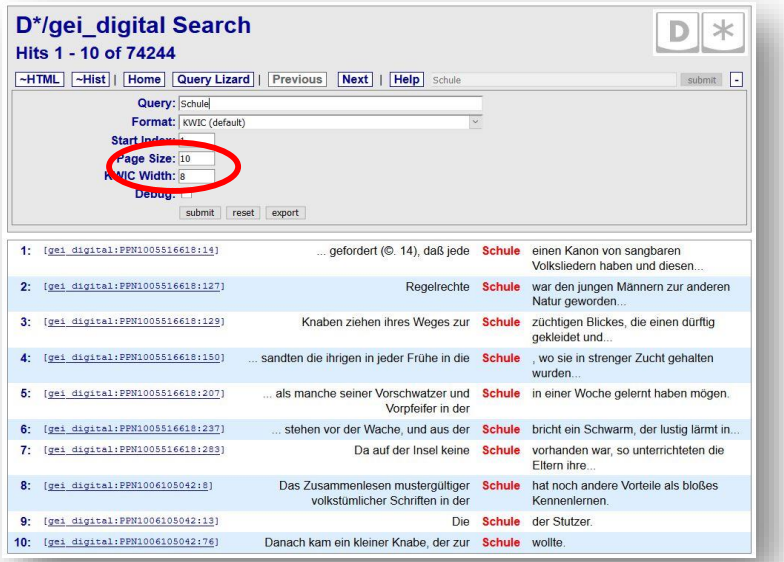
Link zur Originalquelle



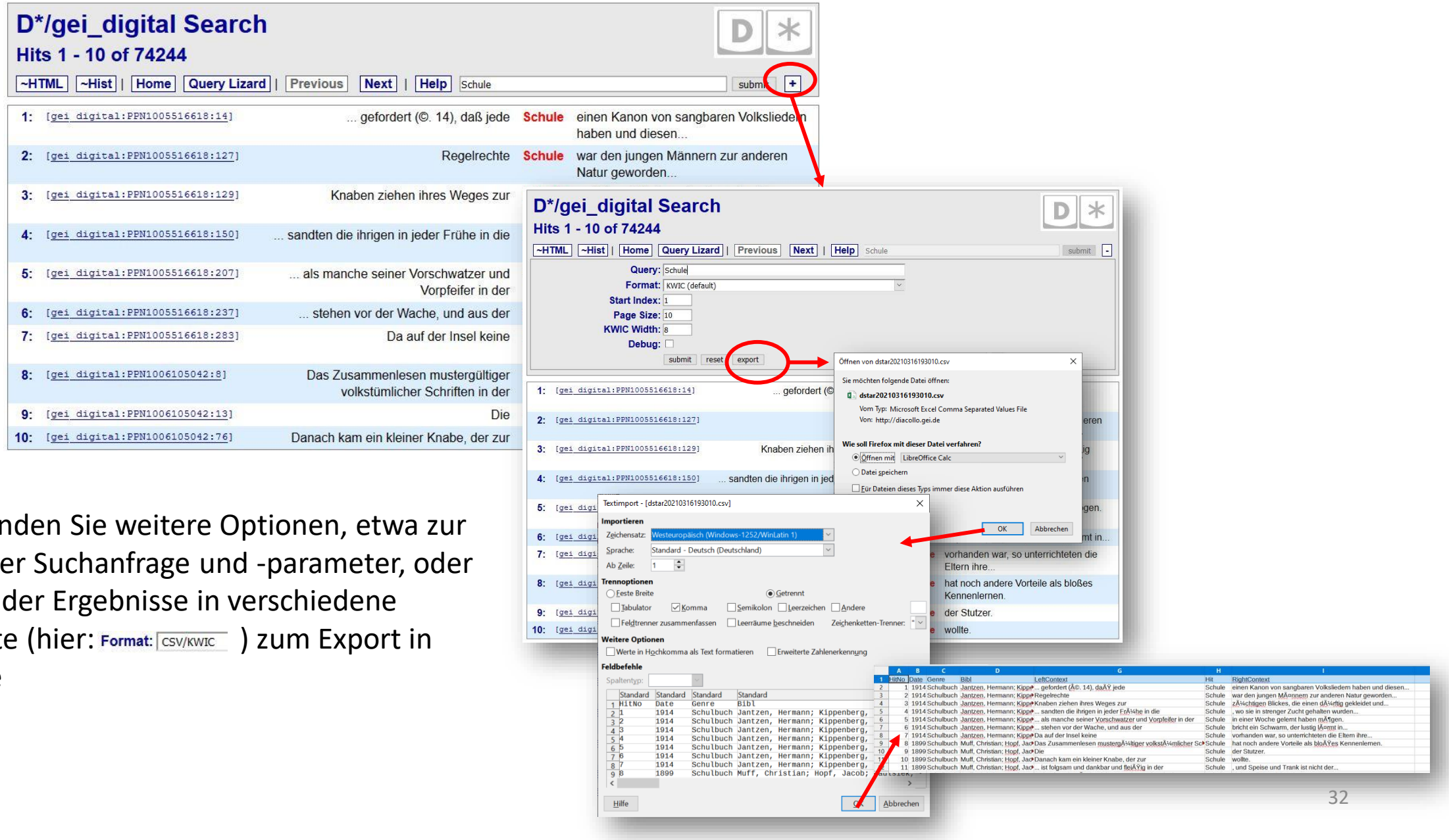


Das gesuchte Stichwort (hier „Schule“) im Kontext der Quelle.
Standardmäßig werden bis zu 8 Tokens vor und nach dem gefundenen Stichwort angezeigt

\+ Weitere Optionen: **KWIC Width** passt die Größe des maximal angezeigten Kontextes an; **Page Size** bestimmt die Anzahl der auf einer Seite angezeigten Treffer (standardmäßig 10)

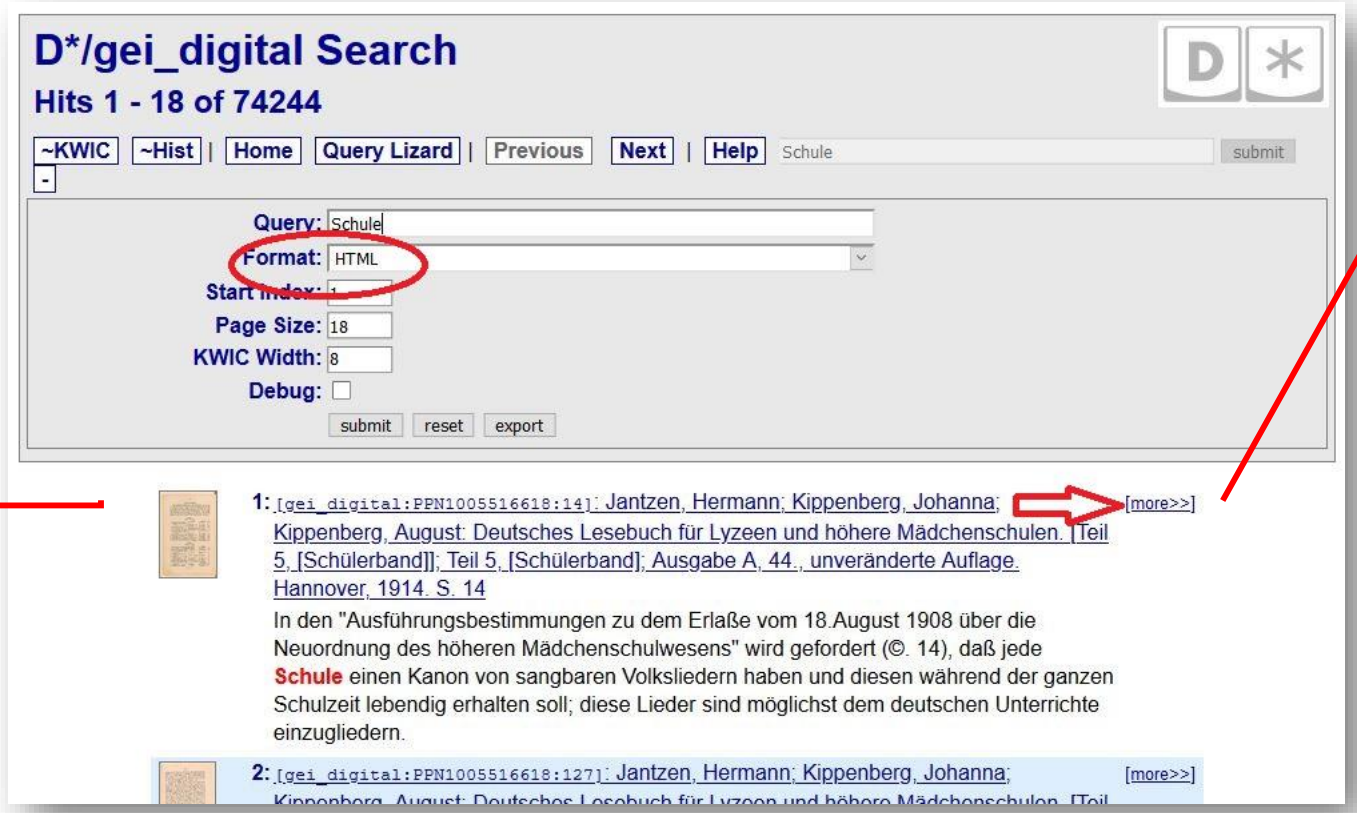


Ergebnisansicht im KWIC-Format und Exportfunktion

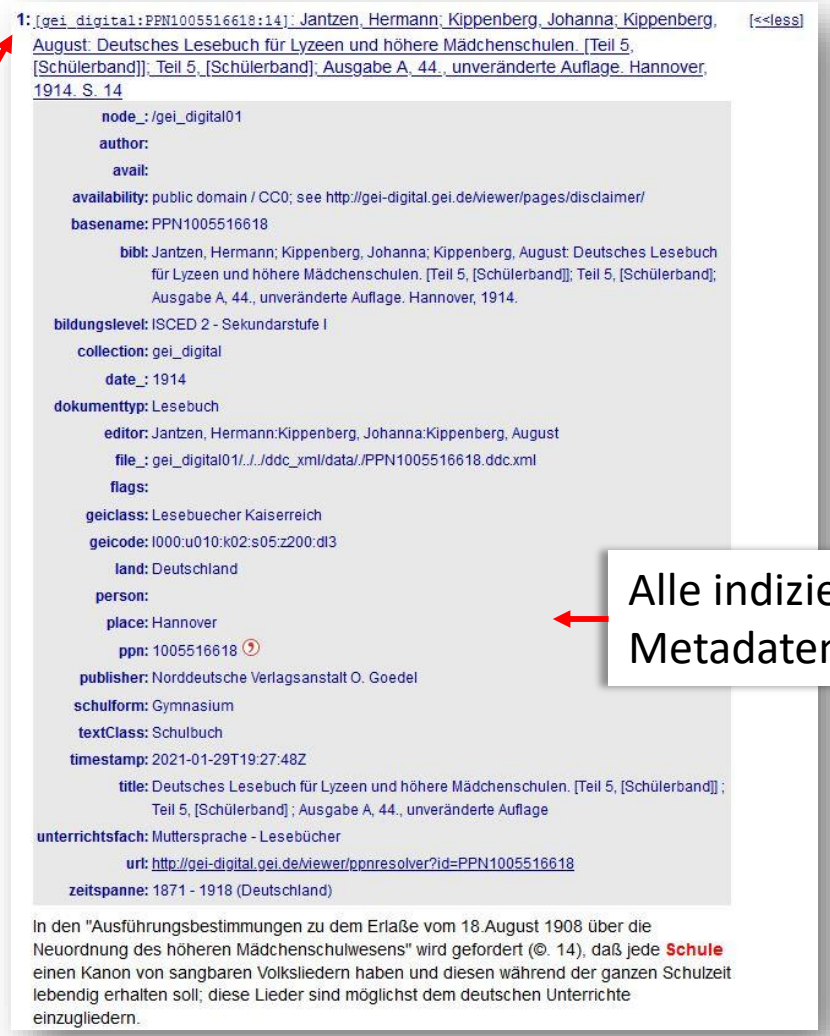


Unter [+] finden Sie weitere Optionen, etwa zur Änderung der Suchanfrage und -parameter, oder zum Export der Ergebnisse in verschiedene Dateiformate (hier: **Format:** CSV/KWIC) zum Export in eine Tabelle

Ergebnisansicht im HTML-Format

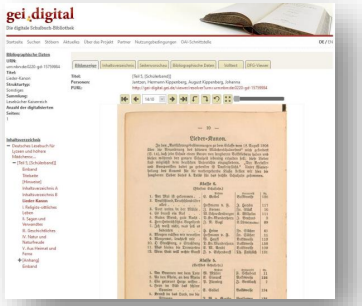


Ein Klick auf „more“ führt zur Detailansicht der zum Treffer verfügbaren Metadaten



Alle indizierten Metadaten

Ein Klick auf die Seitenvorschau oder die bibliographischen Angaben führt zum Digitalisat der Quelle



Mögliche Formate einer Suchabfrage

Query: Eingabefeld für DDC-(konforme) Suchanfragen

- **Stichworte**, z.B. Suchbegriffe wie das Lemma „Schule“
  Solche „Nur-Wort“-Suchanfragen werden (mittels der impliziten, standardmäßig erfolgenden Expansion |Lemma) automatisch lemmatisiert. D.h. eine Anfrage wie „lernte“ wird auf das Lemma „lernen“ zurückgeführt. In jedem Fall werden, wenn nicht explizit anders gewünscht (z.B. mit @lernte), alle Formen eines Lemmas (mittels des $Lemma Index-Attributs) bei der Bearbeitung einer Anfrage mit einbezogen. Treffer beinhalten in diesem Beispiel also auch Formen „gelernt“, „lernte“, lernt“ usw.
- Suchanfragen mit **Regulären Ausdrücken**,
  (vgl. https://www.dwds.de/d/korpussuche#re)
- **Suchanfragen, die Elemente der DDC Abfragesprache nutzen** (wie z.B. NEAR(Schule,Kind,5) oder Aggregierungsabfragen wie „count(* #in file) #by[geiclass]“)

Die folgenden Abfragen sind äquivalent:

Schule

$Lemma=/schule/

$Lemma=Schulen
$Lemma=Schulen|-
$Lemma=Schulen|Lemma
$Lemma=@{Schulen,Schule}

Die folgenden Folien zeigen eine Reihe von Beispielen für die verschiedenen Abfragemöglichkeiten

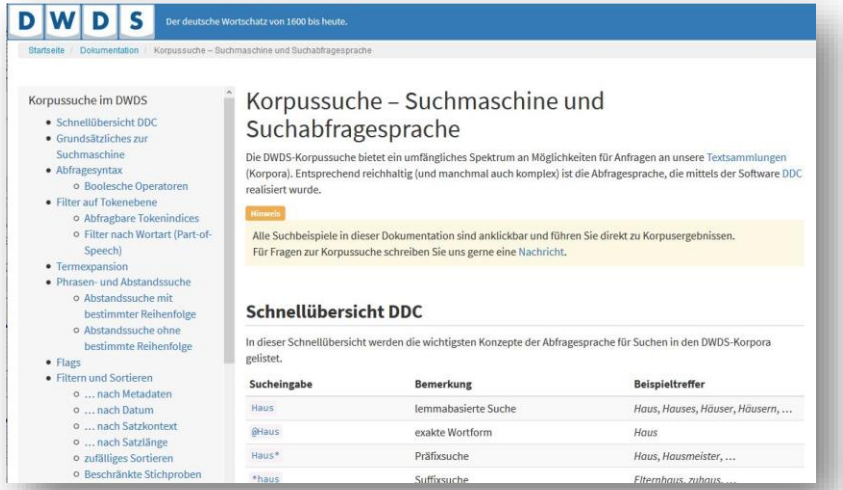
Beispiele und Tipps für die Formulierung von DDC-Suchanfragen

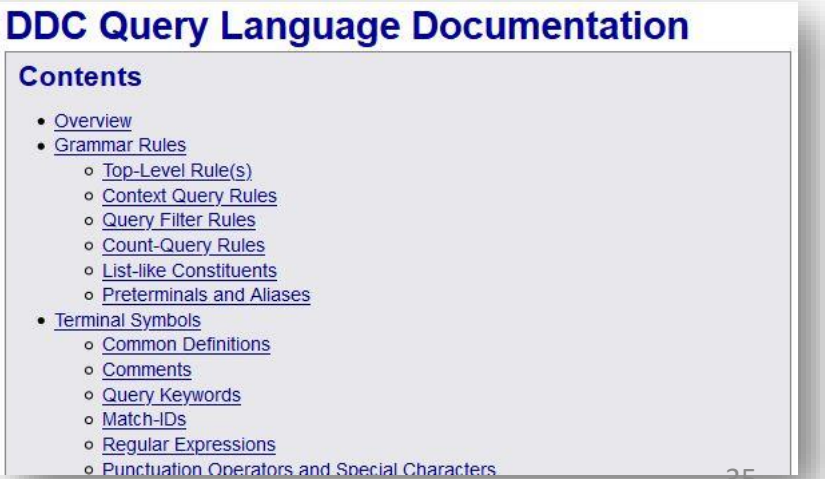
Die Suchmöglichkeiten in **D*** entsprechen im Wesentlichen der Korpussuche im *Digitalen Wörterbuch der Deutschen Sprache* (DWDS), die auf dieser Seite mit vielen Beispielen erläutert wird: https://www.dwds.de/d/korpussuche

(Zu den Unterschieden vgl. https://www.dwds.de/d/korpussuche#do und Folie 44).

Die dortigen Beispiele verlinken auf entsprechende Suchanfragen im DWDS-Kernkorpus. Diese Beispielabfragen können aber auch kopiert, und für die Suche in „GEI-Digital-2020“ Korpus genutzt (und natürlich angepasst) werden – eine gute Möglichkeit, um auf den Geschmack zu kommen. Auf den folgenden Folien finden Sie ein paar Beispiele.

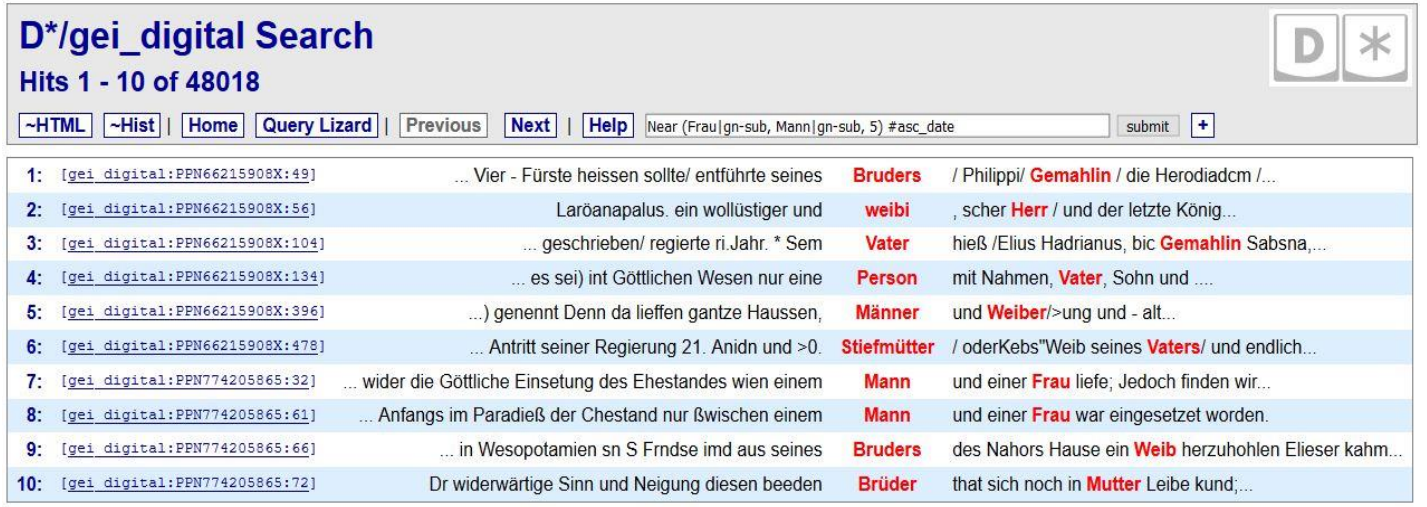
Zu einer englischsprachigen technischen Dokumentation der Abfragesprache in/für DDC gelangen Sie durch einen Klick auf die Schaltfläche „Help“ in D*, nämlich hierhin:
http://kaskade.dwds.de/~moocow/software/ddc/querydoc.html
Vgl. dort besonders:
http://kaskade.dwds.de/~moocow/software/ddc/querydoc.html#ex





Einige Beispiele für DDC-Abfragen

Gesucht werden hier alle im Abstand von maximal 5 Worten innerhalb eines Satzes gemeinsam vorkommenden Instanzen von „Frau” und „Mann”, bzw. deren Synonymen; die Ergebnisse sollen nach Publikationsdatum aufsteigend sortiert werden:
[Near (Frau|gn-sub, Mann|gn-sub, 5) #asc_date]



Suche nach Adjektiv + „Schüler“: ["$p=ADJA Schüler"]

Suche nach einer festen Phrase mit genau einem Wort Abstand, mit genauer Wortform „ein“: ["@ein #=1 Schüler"]

Beispiel-Spickzettel für DDC-Abfragen in D*

Diese Beispiele sind angelehnt an die Dokumentation der DWDS Korpussuche. Dort finden Sie weitere Beispiele: https://www.dwds.de/d/korpussuche#cheatsheet

Sucheingabe	Beschreibung des Gesuchten	Beispieltreffer
Schüler	lemmabasierte Suche	*Schüler, Schülern, Schülers …*
@Unterricht	exakte Wortform	*Unterricht*
Schul*	Präfixsuche	*Schule, Schulze, Schuljahr, …*
*kunst	Suffixsuche	*Dichtkunst, Baukunst, Rechenkunst,…*
schul	Infixsuche	*Mädchenschulen, Provinzialschulrat, verschuldet, …*
/ha[mu]s?t/	regulärer Ausdruck	*Schaute, Schaustellung, hauste, Hornhaut, …* [ha, gefolgt von m oder u, danach optional s, dann t]
/weg/gi	regulärer Ausdruck auf ganzem Tokenfeld mit Ignorieren der Groß-/Kleinschreibung	*weg, weG, wEg, wEG, Weg, WeG, WEg, WEG*
weg|case	Termexpansion (hier: alle Groß-/Klein-Schreibvarianten im Korpus)	*WEG, Weg, weg*
{Schule,Hof}	Tokens als Menge	*Schule, Hof, Hofe, Höfe, …*
Schule && Arbeit	Und-Suche	Sätze, in denen Formen von *Schule* **und** *Arbeit* vorkommen
Unterricht || lernen	Oder-Suche	Sätze, in denen Formen von *Unterricht* **oder** *lernen* vorkommen
Schule && !Schüler	Negation	Sätze, in denen eine Form von *Schule* **und keine** Form von Schüler vorkommt
"eine Schule"	Wortgruppe/Phrase	*eine Schule, einer Schule, …*
"eine #2 Schule"	Phrase mit Abstand (maximal 2)	*eine Schule, eine gute Schule, ein Bild der Schule, …*
"ein #>2 Schüler"	Phrase mit Abstand (mehr als 2)	Sätze, in denen zwischen einer Form von *ein* und einer Form von *Schüler* mehr als 2 Tokens stehen
"ein #=2 Schüler"	Phrase mit Abstand (genau 2)	Sätze, in denen zwischen einer Form von *ein* und einer Form von *Schüler* genau 2 Tokens stehen
NEAR(gut,Kind,3)	Abstandssuche ohne best. Reihenfolge	*guter Leute Kind, mein Kind mag besser sein, gefiel dem Kinde so gut, …*
NEAR(Hans,Hänschen,lernen,5)	Abstandssuche ohne best. Reihenfolge	*Was Hänschen nicht lernt, lernt Hans nimmermehr*
NEAR("wenn ich","werde ich",2)	Abstandssuche ohne best. Reihenfolge mit Wortgruppen	*Wenn ich wiederkomme, werde ich [...]*
$p=PPOSS	Abfrage nach Wortart	*seinen, seinigen, ihrigen, unsrige, meinen, …*
Schule WITH $.=0	Satzanfang	Sätze, die mit einer Form von *Schule* beginnen
Schule WITH $.=-2	Satzende (Hinweis: $.=-2 sucht nach dem vorletzten Token im Satz, meist ist ein Satzzeichen das letzte Token)	Sätze, die mit einer Form von *Schule* enden

Beispiele für DDC-Abfragen: Ergebnisse zählen mit COUNT()-Abfragen

COUNT(*#sep) #BY[date/1] — Tokenzahlen im Korpus, nach Publikationsjahr gruppiert

COUNT(*) #BY[date/1] — Satz-Anzahl im Korpus, nach Publikationsjahr gruppiert

Anzahl der Sätze im Korpus pro Publikationsjahr; hier angezeigt für die ersten 10 der insgesamt 377 Jahre, aus denen Publikationen im Korpus enthalten sind.

COUNT(* #in file) #BY[date/1] — Anzahl der Werke im Korpus, nach Jahr gruppiert

COUNT(* #in file) #BY[date/10] — Anzahl der Werke im Korpus, nach Dekade gruppiert

COUNT(* #in file) #by[schulform] — Metadatum "schulform" und die vergebenen Attribute

Achtung: Manche Felder für dieses Metadatum enthalten mehrere Attribute, z.B. Beispiel: Realschule:Gymnasium:Lehrerbildungseinrichtung:Berufliche Schule). In solchen Feldern sucht man mit DDC am besten mit Regulären Ausdrücken, z.B.:
count(* #in file #has[schulform,/\bRealschule\b/]) #by[schulform]

COUNT(Schule) #BY[date/10] — Treffer für Formen von *„Schule“*, nach Dekade gruppiert

COUNT(schule*) #BY[$p,$l] — Treffer für Wörter mit dem Präfix „schule“, nach Wortart und Lemma gruppiert

COUNT(Schule) #BY[$l-1] #DESC_COUNT — Treffer für Formen von „Schule“, nach Lemma des linken Nachbarn gruppiert, Gruppierungskriterium absteigend

Diese Beispiele sind angelehnt an die Dokumentation der DWDS Korpussuche. Dort finden Sie weitere Beispiele: https://www.dwds.de/d/korpussuche#count

Q&A: Knifflige Fragen und Antworten

z.B. betreffend…

Filtern mit Metadaten
Metadaten filtern mit Regulären Ausdrücken
Suchen in einzelnen Werken
Suchen in einem bestimmten Zeitraum
Unterschiede D* und DWDS
Frequenzabfragen mit verschiedenen Werkzeugen (und Indizes)

Wie kann ich im „GEI-Digital-2020“-Korpus nach Metadaten filtern?

Welche Arten von Metadaten (Metadaten-Attribute) kann ich für Suchabfragen nutzen?

Für Abfragen im „GEI-Digital-2020“-Korpus bieten sich die Folgenden an: *bildungslevel, dokumenttyp, editor, geiclass, land, place, ppn, publisher, schulform, unterrichtsfach*

Eine Liste aller Metadaten können Sie durch den Klick auf „Details“ in D* erreichen, und dort unter der Rubrik „Bibliographic Metadata Attributes“: http://diacollo.gei.de/gei-digital-2020/details.perl#bibl

Wie sehe ich die Metadaten eines bestimmten Buches, bzw. eines bestimmten Treffers meiner Suchabfrage?

Die Metadaten und ihre Attribute einzelner Werke werden bei D*Query-Ergebnissen im Format "HTML" angezeigt, wenn man die Ansicht der Einzelergebnisse durch Klick auf "more" erweitert (siehe Folie 33).

Wie kann ich Metadaten-Attribute für die Suche nutzen?

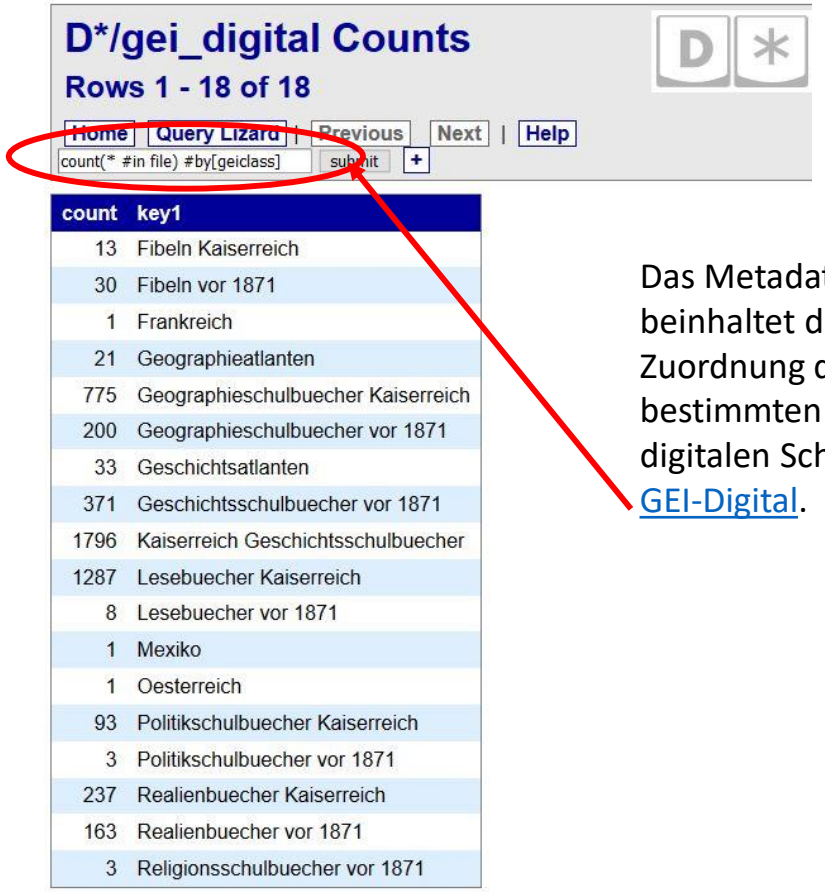
Zum Beispiel so:

Suchbegriff #HAS[geiclass,'Geschichtsschulbuecher vor 1871']

Welche Unterkategorien (Werte) finden sich in den verschiedenen Arten von Metadaten (Metadaten-Attributen) ?

Danach können Sie suchen mit: count(* #in file) #by[*metadatum*]

Hier als Beispiel: count(* #in file) #by[geiclass]



Das Metadatum "geiclass" beinhaltet die bibliothekarische Zuordnung der Werke zu bestimmten Sammlungen in der digitalen Schulbuchsammlung GEI-Digital.

NB: Beachten Sie die unterschiedliche Anzahl der verfügbaren Werke in GEI-Digital (http://gei-digital.gei.de/) im Gegensatz zum GEI-Digital-2020 Korpus. GEI-Digital wird laufend erweitert, während das Korpus statisch den Stand verfügbarer Volltexte von Ende 2020 abbildet. Zum anderen werden in GEI-Digital derzeit nicht für alle Werke automatisch generierte Volltexte erzeugt, weil z.B. bei Fibeln und Atlanten aufgrund vieler Abbildungen und uneinheitlichem Schriftbild die OCR-Ergebnisse zu hohe Fehlerraten aufweisen.

Metadaten filtern mit Regulären Ausdrücken

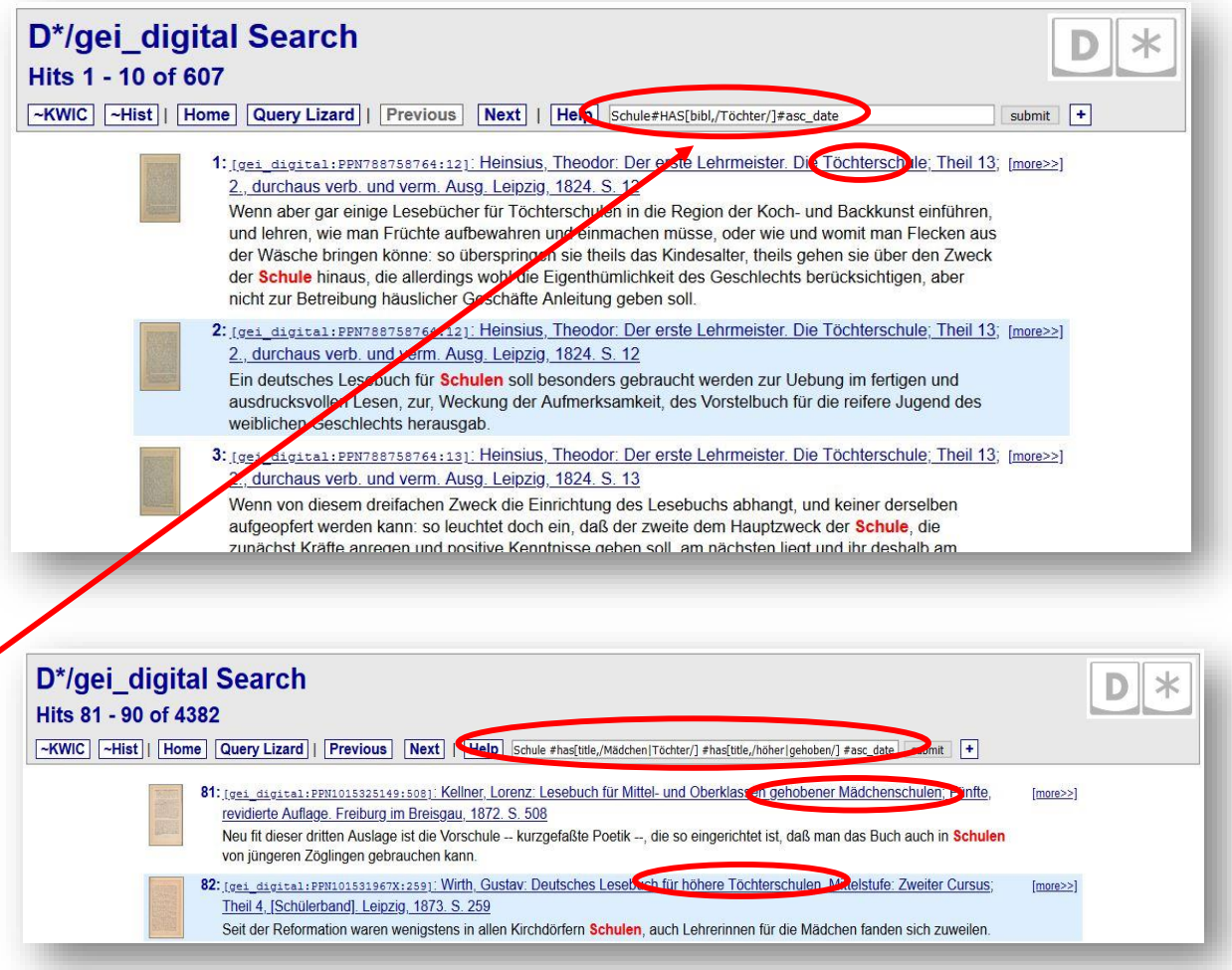
Kann ich einen Teil des Buchtitels als Filter für Abfragen benutzen?

Ja, aber hierfür kann man keine Elemente der DDC-Abfragesprache oder Operatoren nutzen, sondern muss Reguläre Ausdrücke formulieren.

Grund hierfür ist, dass die Werktitel nicht zu den Volltexten, sondern zu den Metadaten gehören (konkret zum Metadatum „bibl“). Metadaten wurden bei der Erstellung des Korpus nicht derselben computerlinguistischen Vorverarbeitung unterzogen wie die Volltexte. Sie wurden nicht tokenisiert, lemmatisiert etc. sondern sind für den Computer weiterhin einfache Zeichenfolgen (atomare strings). Deshalb kann man dort nicht mit DDC Abfrage-Suchoperatoren auf Tokenebene (wie NEAR()) suchen.
Merke: „Was nicht indiziert wurde, danach kann auch nicht gesucht werden.“

Im Beispiel rechts oben werden alle Vorkommen des Lemmas „Schule“ gesucht in Büchern, die folgendes Merkmal aufweisen (#HAS): Im Metadatum zu den bibliographischen Angaben (bibl) kommt die Zeichenfolge (/Töchter/) vor. Die Ergebnisse sollen aufsteigend nach dem Publikationsjahr sortiert werden (#asc_date).

Diese Anfrage verbindet also DDC-Suchoperatoren mit einem Regulären Ausdruck und bildet so einen Filter (#HAS) auf einen Teil des Buchtitels.



Unteres Beispiel:
Schule #has[title,/Mädchen|Töchter/] #has[title,/höher|gehoben/] #asc_date

NB: Unter diacollo.gei.de finden Sie eine Exceldatei mit den bibliographischen Angaben aller Werke im GEI-Digital-2020 Korpus. Die Titel und Untertitel der Werke geben z.T. Auskunft über die Adressaten, z.B. katholische oder evangelische Schulen, Lehrer- und Lehrerinnenbildungsanstalten, Stadt- oder Landschulen usw.

Kann ich auch Stichwörter oder Frequenzen in einem einzelnen Werk suchen?

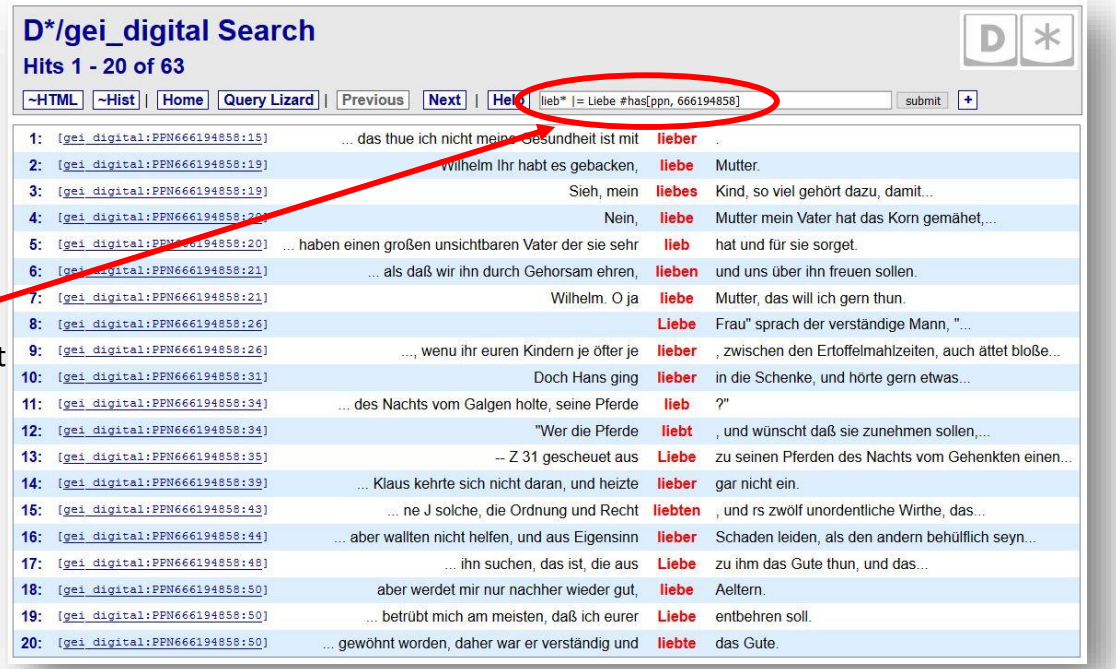
Ja, dafür nutzt man die PPN (Persistenter Identifier) zum filtern, z.B. so:
Suche nach Stichwort „Fleiß“ im „Kinderfreund“ von Rochow von 1798:
Fleiß #has[ppn,666194858]

Die PPN eines bestimmten Buches können Sie in GEI-Digital recherchieren oder in der auf diacollo.gei.de verlinkten Excelliste (dort als Bestandteil der URL der Digitalisate) nachschauen. In D* ist die PPN der Werke jeweils auch Bestandteil der Trefferanzeige.

Analog dazu:

Alle (indexierten=findbaren!) Eigennamen: $p=NE #has[ppn, 666194858]
Alle Nomen in diesem Buch: $p=NN #has[ppn, 666194858]

„lieb“ als Präfix in diesem Buch: lieb* #has[ppn, 666194858]
„lieb*“ oder „Liebe“ in diesem Buch: lieb* |= Liebe #has[ppn, 666194858] entspricht
lieb* WITHOR Liebe #has[ppn, 666194858]



Wie kann ich in einem bestimmten Zeitraum im „GEI-Digital-2020“-Korpus suchen?

count	key1
3	1600
68	1700
2365	1800
2600	1900

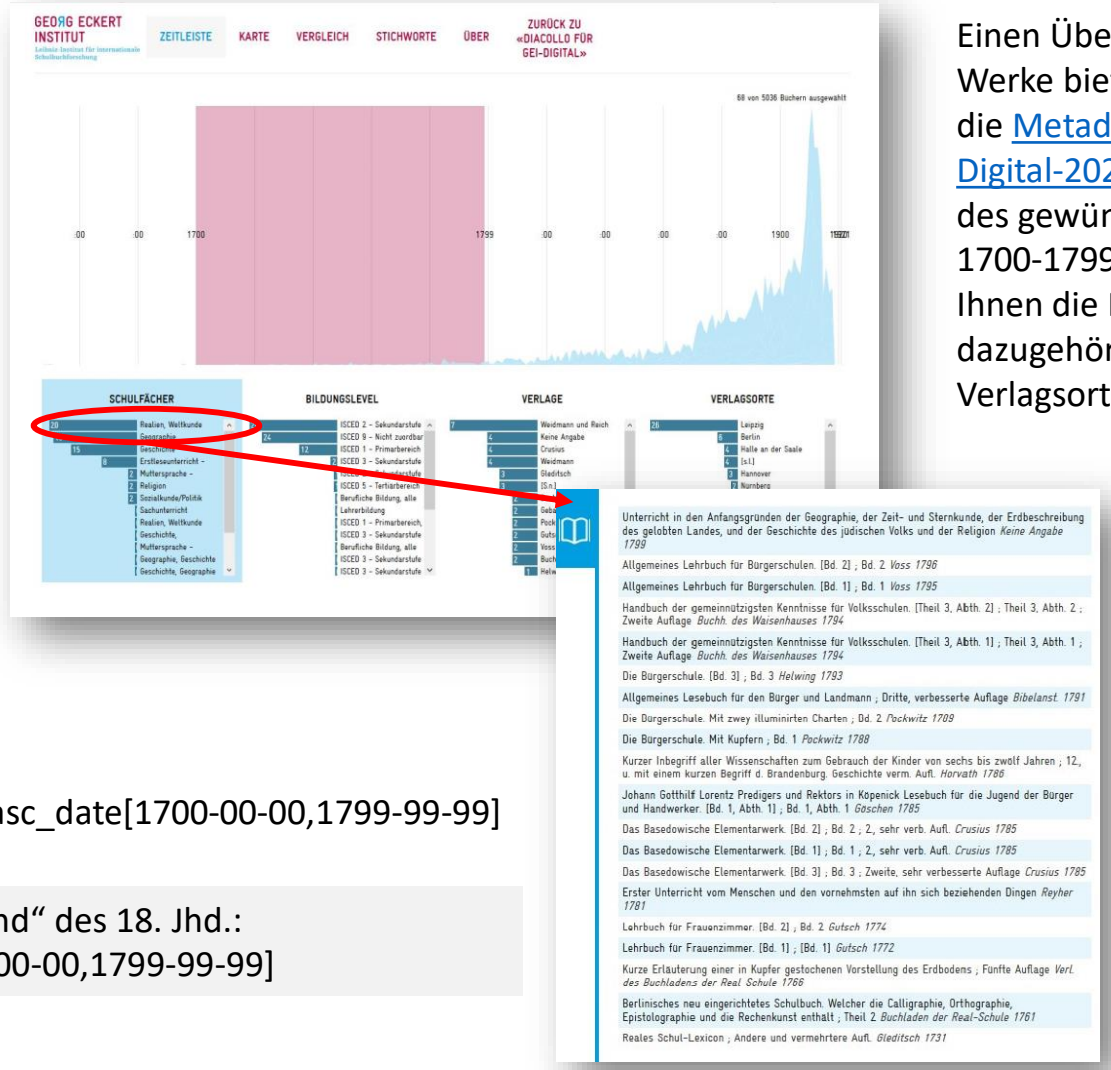
Eine COUNT-Abfrage nach Jahrhunderten ergibt für das 18. Jhd. (= 1700-1799) insgesamt 68 Werke im Korpus.

Eine DDC-Query für diesen Zeitraum sieht dann so aus:
Stichwort #asc_date[1700-00-00,1799-99-99]

Suche in Texten der 1870er Jahre:
Stichwort #asc_date[1870-00-00,1879-99-99]

Suche in Realienbüchern des 18. Jhd.:
Stichwort #HAS[geiclass,'Realienbuecher vor 1871'] #asc_date[1700-00-00,1799-99-99]

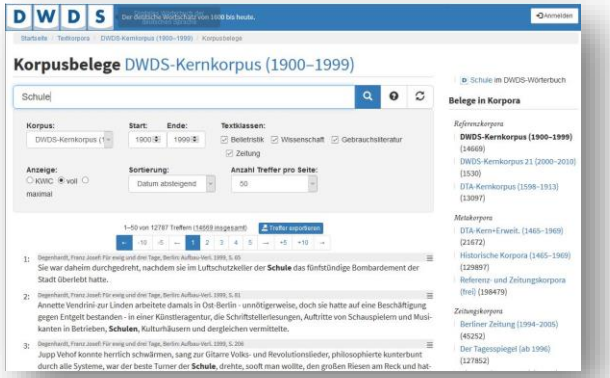
Suche in Bücher mit dem Titelbestandteil „Kinderfreund“ des 18. Jhd.:
Stichwort #HAS[bibl,/Kinderfreund/] #asc_date[1700-00-00,1799-99-99]



Einen Überblick über die vorhandenen Werke bietet auch ein externes Tool, die Metadaten-Visualisierung des GEI-Digital-2020 Korpus. Durch die Auswahl des gewünschten Zeitraumes (hier 1700-1799) auf der Zeitleiste werden Ihnen die Metadaten der dazugehörigen Werke, Schulfach, Verlagsort etc. angezeigt.

Durch Klick auf eine Facette wie z.B. „Realien“ lässt sich die ausgewählte Menge weiter filtern. In der rechten Seitenleiste können die Titel der jeweils gewählten Menge eingeblendet werden. Sie sind mit den Digitalisaten verlinkt.

Entsprechen Suchen in D* – bzw. der D*-Instanz des Georg-Eckert-Instituts – exakt den Suchen in Korpora des Digitalen Wörterbuchs der deutschen Sprache (DWDS)?



Nein, es gibt gewisse Unterschiede, vgl. https://www.dwds.de/d/korpussuche#do

dwds-Korpussuchen nutzen eine zusätzliche Komponente zum Auffinden, Reparieren oder Ablehnen von fehlerhaften Anfragen.
Die D*-Instanzen verwenden diese Abfragebereinigungskomponente nicht. Einige "benutzerfreundliche" dwds-Abfrage-Sanierungs-Transformationen werden von D* und DWDS unterschiedlich behandelt (insbesondere implizite Phrasen-Abfragen)

dwds: das Haus --> "das Haus" (Phrasenabfrage)
D*: das Haus --> das && Haus (logische Konjunktion)

dwds.de fügt außerdem implizit das Schlüsselwort #sep zu Benutzerabfragen hinzu, es sei denn, der Benutzer gibt #join an; bei dstar ist #join die Vorgabe und #sep muss explizit hinzugefügt werden, falls gewünscht (dies kann sich ändern). Ein Beispiel hierfür sind die beiden ersten COUNT-Abfragen auf Folie 38.

Ich kann Frequenzen mit verschiedenen Werkzeugen abfragen, richtig?

Ja, dafür gibt es mehrere Möglichkeiten:

1. Mittels COUNT() Abfragen mit DDC: count(Schule) submit

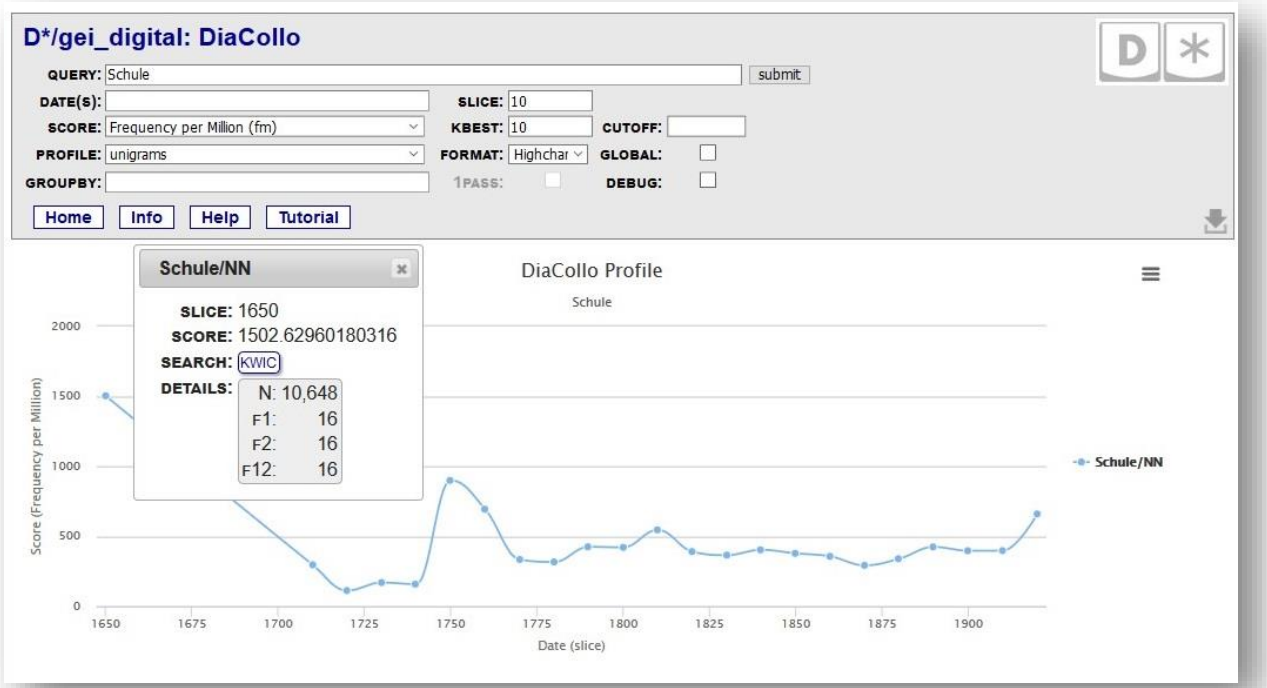
count	key1
74244	*

(vgl. Folie 38)

2. Mittels LexDB (vgl. Folie 23-26).

3. Per Time Series. Dabei ist zu beachten, dass die Ergebnisse in der Standard („default“) Einstellung in 5-Jahreschritten („SLICE“ 5) und geglättet („WINDOW“ 1) dargestellt werden. Für eine jahresgenaue Anzeige setzen Sie die Werte Slice=1 und Window=0. Per Klick auf TSV („raw tab-seperated data“) kann man die Zahlen der einzelnen Jahre genauer sehen (vgl. Folie 19-21).

4. In DiaCollo: Score: Frequency (=absolute Häufigkeit) oder Frequency per million (relative Häufigkeit). Profile: Unigramms (für Frequenz des Suchbegriffs) oder ddc (für Frequenzen der Kollokate zum Suchbegriff) und Format: Highchart. Slice=1 für ungeglättete, jahresgenaue Anzeige (vgl. Folie 58).
Beispiel rechts:

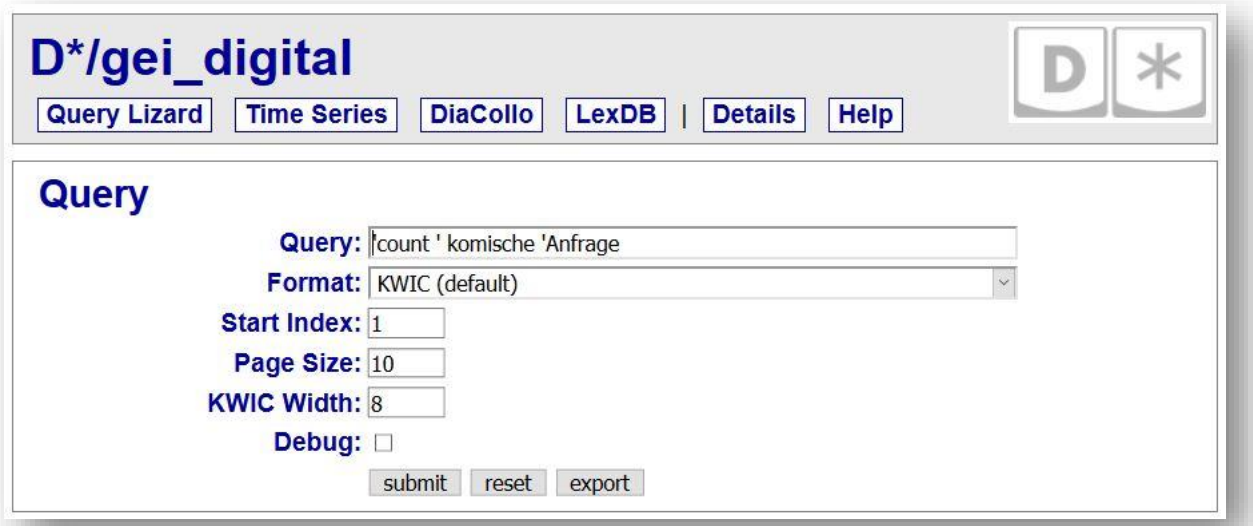


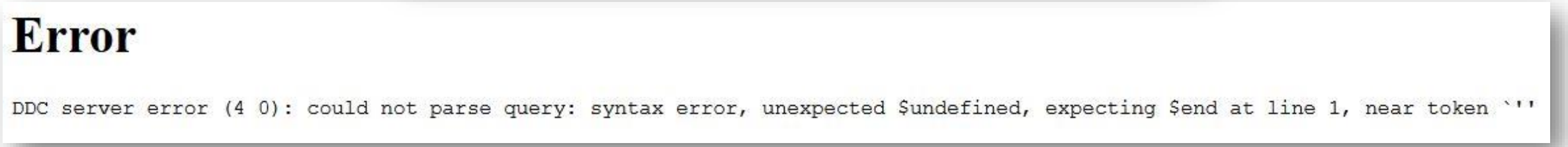
ACHTUNG 1: Beachten Sie bitte, das von den Werkzeugen ggf. verschiedene Indizes genutzt werden (vgl. Folie 9 und Folie 12). Der Index von DDC (der auch für die Time Series genutzt wird) enthält *alle* Tokens aller Texte; der von DiaCollo genutzte Index hingegen nur eine Auswahl: hier werden bestimmte Wortarten nicht mit einbezogen, da diese normalerweise als überflüssig und/oder hinderlich für die Berechnung von Kollokationen angesehen werden. Die Gesamtzahl N der Token in den Indizes ist deshalb unterschiedlich groß, und deshalb auch die berechnete Frequenz der gesuchten Begriffe pro Mio. Token.

ACHTUNG 2: Die Anzahl der untersuchten Tokens (N) ist im „GEI-Digital-2020“- Korpus nicht für jedes Jahr gleich. Vor 1871 und nach 1918 sind vergleichsweise weniger Daten vorhanden (vgl. http://diacollo.gei.de/gei-digital-2020/visualized/#/Stream), so dass einzelne nicht/vorhandene Worte aufgrund der kleineren Grundgesamtheit stärker ins Gewicht fallen (siehe hier oben die „hohe Frequenz“ durch 16 Treffer für „Schule“ in Texten von 1650).

ACHTUNG 3: Durch Fehler bei der automatischen Volltexterkennung der historischen Werke dieses Korpus erlauben die Frequenzberechnungen nur näherungsweise Aussagen über die tatsächliche Häufigkeit der gesuchten Begriffe im Quellenmaterial (vgl. Folie 10) .

Fehlermeldungen





Fehlermeldungen werden ausgegeben, wenn DDC die gestellte Anfrage nicht „versteht“, d.h. sie nicht verarbeiten kann, weil sie nicht den Konventionen der DDC-Abfragesprache entspricht. Die Fehlermeldung gibt Hinweise auf die Art des Fehlers.

Zu Fehlermeldungen in DiaCollo gibt es hier eine Reihe von Erklärungen und Hilfestellungen:
https://diacollo.gei.de/dstar/gei_digital/diacollo/help.perl#faq-errors

Teil 3:

DiaCollo – Parameter, Ergebnisansichten und Beispielanfragen

Kollokationen

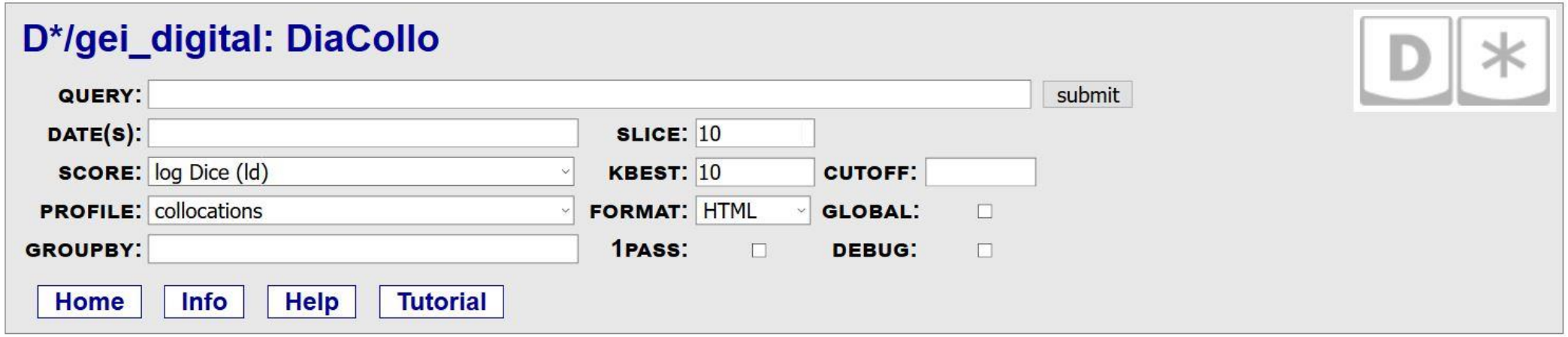
Kollokation ist ein Fachbegriff der Linguistik. In der computergestützten Korpusanalyse bezeichnet Kollokation ein statistisch auffälliges gemeinsames Vorkommen von Wörtern innerhalb eines vordefinierten Abstands. Über Kollokationsabfragen (wie *DiaCollo*) kann z. B. herausgefunden werden, dass ein Wort X häufig in einem Umkreis von 5 Wörtern vor oder nach einem Stichwort Y vorkommt.

Die Gründe für das gemeinsame Vorkommen können unterschiedlicher Natur sein. Für die historische Forschung ist ein Vergleich dieser häufigen Wortverbindungen über die Zeit interessant, um z.B. Sprach- bzw. Bedeutungswandel auf die Spur zu kommen.

”You shall know a word by the company it keeps.“

John Rupert Firth

Dies ist die Oberfläche zur Formulierung von Anfragen in DiaCollo:



DATE(S): Hier kann eingegeben werden, welcher (Publikations-) Zeitraum untersucht werden soll; wenn das Feld leer bleibt, wird der gesamte vorhandene Zeitraum genutzt

SCORE: Art der gewünschten Berechnung/Bewertung

PROFILE: Hier wird gewählt, auf welche „Darreichungsform“ der Rohdaten DiaCollo zugreift, um die Anfrage auszuführen, z.B. Profiltyp „collocations“ für den DiaCollo-Index zur Ermittlung und Bewertung von Kollokationen

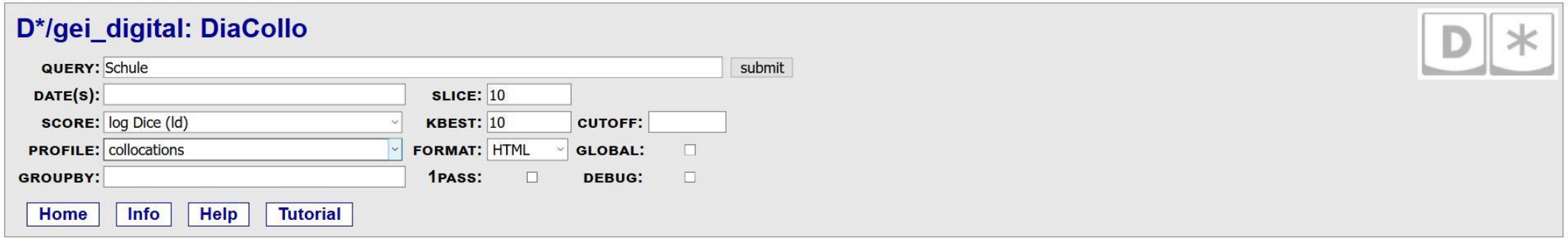
SLICE: Frei wählbares Zeitintervall (Zeitschnitt) für den Vergleich, standardmäßig in Dekaden

KBEST: Maximale Anzahl der zu ermittelnden „besten Kollokate“, standardmäßig bis zu 10

FORMAT: Gewünschtes Ausgabeformat der Ergebnisse, standardmäßig ist hier „HTML“ (eine Liste) voreingestellt

GROUPBY: Hier kann angegeben werden, welche indizierten Attribute der Kollokationskandidaten für die Berechnung genutzt und in der Ergebnismenge gezeigt werden sollen. Es kann auch gewählt werden, ob z.B. nur bestimmte Wortformen in Betracht gezogen werden sollen

Beispielanfrage und Standard-Ergebnisanzeige in DiaCollo



Raw URL: http://diacollo.gei.de:8082/dstar/gei_digital/diacollo/profile.perl?profile=2&slice=10&format=html&score=ld&query=Schule&kbest=10&eps=0&diff=adiff

N	f1	f2	f12	score	label	lemma	pos	
89674	147	19	2	8.6250	1650	kriegen	VVFIN	KWIC
89674	147	25	2	8.5737	1650	falsch	ADJA	KWIC
89674	147	312	4	8.1576	1650	kommen	VVFIN	KWIC
89674	147	88	2	8.1235	1650	Schlag	NN	KWIC
89674	147	203	2	7.5488	1650	lernen	VVINF	KWIC
89674	147	205	2	7.5406	1650	lernen	VVPP	KWIC
89674	147	262	2	7.3240	1650	Herr	NN	KWIC
89674	147	430	2	6.8276	1650	sagen	VVFIN	KWIC
89674	147	506	2	6.6491	1650	klein	ADJA	KWIC
89674	147	664	2	6.3364	1650	E	NE	KWIC
459360	152	45	2	8.3779	1710	Jus	NE	KWIC
459360	152	64	2	8.2451	1710	Jugend	NN	KWIC
459360	152	70	2	8.2056	1710	gebürtig	ADJD	KWIC
459360	152	115	2	7.9393	1710	Präge	NN	KWIC

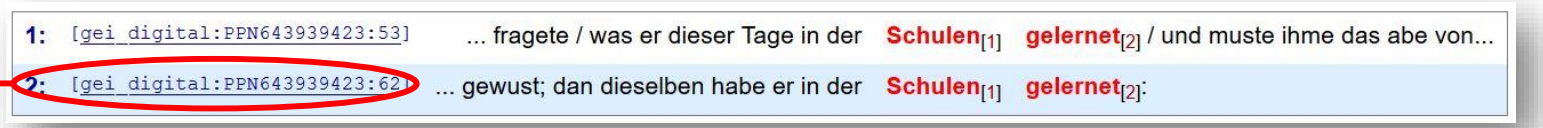
In diesem Beispiel wurden die Standardeinstellungen der Parameter genutzt, um im „GEI-Digital-2020“-Korpus Kollokationen zum Stichwort „Schule“ zu ermitteln.

Die Ergebnisse werden im Format „**HTML**“ in Listenform dargestellt. Untersucht wurde das gesamte Korpus in 10-Jahres-Intervallen.

Link zum „keyword-in-context“ im D*-Index

NB: Die Ermittlung von f12 erfolgt im DiaCollo-Index. Die hier verlinkte Darstellung der Stichworte im Kontext wird nicht von DiaCollo selber ausgeführt, sondern jeweils als Anfrage an die DDC-Korpussuche weitergereicht. Diese bezieht sich auf den DDC-Index und kommt deshalb ggf. zu leicht abweichenden Ergebnissen, vgl: http://diacollo.gei.de/dstar/gei_digital/diacollo/help.perl#faq-runtime, "Why don't the corpus KWIC links always return exactly f12 hits?“ Sowohl die DiaCollo-Häufigkeiten als auch die DDC-Häufigkeiten sind also exakt und korrekt – sie zählen nur leicht unterschiedliche Dinge.

Link zum Digitalisat



Die DiaCollo-HTML-Ergebnisansicht im Detail:

N: Gesamtanzahl der untersuchbaren Tokens im Index im gewählten Zeitintervall (hier: Dekaden)

label: Name des Zeitabschnitts (hier „1560“ für die Jahre 1560-1569, und „1710“ für die Jahre 1710-1719). Für die Jahre 1570-1709 standen in diesem Beispiel nicht genug Daten (=zu wenige Texte) zur Verfügung, um statistisch signifikante Kollokationen zu berechnen.

f1: Häufigkeit des gewählten Stichwortes (hier: „Schule“) im gewählten Zeitintervall

f2: Häufigkeit des Kollokates (hier: **lemma, pos**) im gewählten Zeitintervall

f12: Häufigkeit des gemeinsamen Vorkommens von Stichwort und Kollokat im jeweiligen Zeitabschnitt im DiaCollo-Index

score: berechnete Assoziationsstärke von Stichwort und Kollokat; die Größe ist Kriterium für die Reihenfolge in der Listenansicht, für die Farbkodierung, und in anderen Ansichten auch der Größe der Darstellung.

N	f1	f2	f12	score	label	lemma	pos	
89674	147	19	2	8.6250	1650	kriegen	VVFIN	KWIC
89674	147	25	2	8.5737	1650	falsch	ADJA	KWIC
89674	147	312	4	8.1576	1650	kommen	VVFIN	KWIC
89674	147	88	2	8.1235	1650	Schlag	NN	KWIC
89674	147	203	2	7.5488	1650	lernen	VVINF	KWIC
89674	147	205	2	7.5406	1650	lernen	VVPP	KWIC
89674	147	262	2	7.3240	1650	Herr	NN	KWIC
89674	147	430	2	6.8276	1650	sagen	VVFIN	KWIC
89674	147	506	2	6.6491	1650	klein	ADJA	KWIC
89674	147	664	2	6.3364	1650	E	NE	KWIC
459360	152	45	2	8.3779	1710	Jus	NE	KWIC
459360	152	64	2	8.2451	1710	Jugend	NN	KWIC
459360	152	70	2	8.2056	1710	gebürtig	ADJD	KWIC
459360	152	115	2	7.9393	1710	Präge	NN	KWIC

lemma: die jeweils stärksten ermittelten Kollokate zum Suchbegriff pro Zeitabschnitt

pos: Part-of-Speech = Wortart des Kollokates; z.B. NN=Nomen, NE= Eigenname, ADJA= Adjektiv usw.

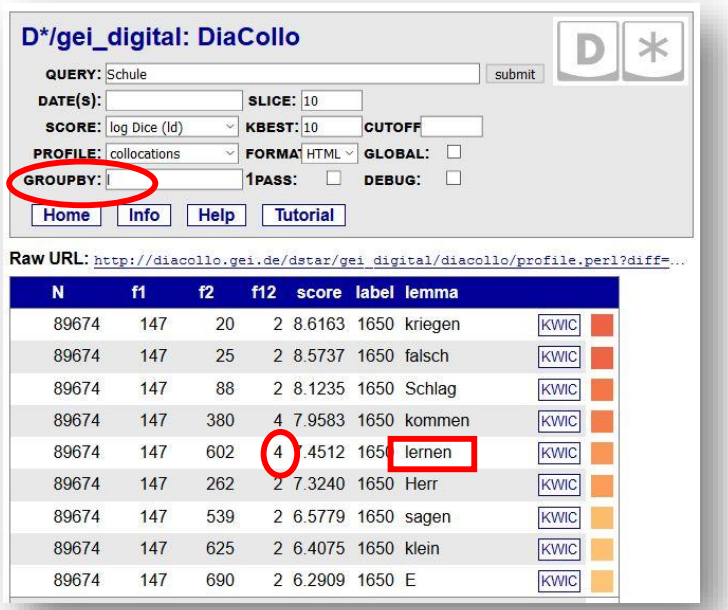
Der GROUPBY-Parameter

GROUPBY: [ ] = **GROUPBY:** l,p

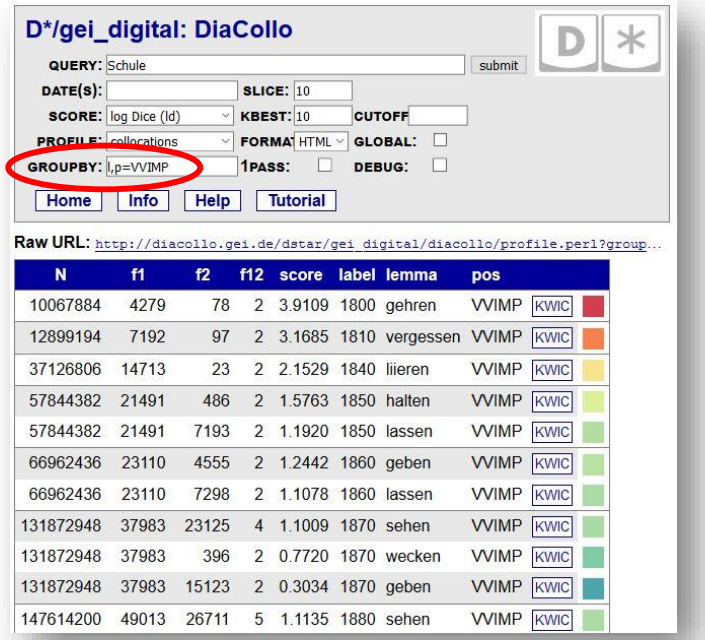
Standardmäßig werden die indexierten Attribute *Lemma* und *Wortart* (Part-of-Speech, PoS) für die Berechnung der Kollokationskandidaten genutzt und in der Ergebnismenge gezeigt (**GROUPBY**: l,p).
Im Beispiel unten ist das Lemma „lernen“ zweimal unter den *k* besten Kollokationen: einmal als Infinitiv (VVINF) und einmal als Partizip Perfekt (VVPP):

N	f1	f2	f12	score	label	lemma	pos	
89674	147	19	2	8.6250	1650	kriegen	VVFIN	KWIC
89674	147	25	2	8.5737	1650	falsch	ADJA	KWIC
89674	147	312	4	8.1576	1650	kommen	VVFIN	KWIC
89674	147	88	2	8.1235	1650	Schlag	NN	KWIC
89674	147	203	2	7.5488	1650	lernen	VVINF	KWIC
89674	147	205	2	7.5406	1650	lernen	VVPP	KWIC
89674	147	262	2	7.3240	1650	Herr	NN	KWIC
89674	147	430	2	6.8276	1650	sagen	VVFIN	KWIC
89674	147	506	2	6.6491	1650	klein	ADJA	KWIC
89674	1							
459360	1							
459360	152	64	2	8.2451	1710	Jugend	NN	KWIC

... fragete / was er dieser Tage in der **Schulen**[1] **gelernet**[2] / und muste ihme das a
... gewust; dan dieselben habe er in der **Schulen**[1] **gelernet**[2].
... etwas tüchti- Zes **lernen**[2] wolte / in die **Schule**[1] gehen / und sich unn lassen.
... agten er brächte nur die Zeit mie dem **Schule**[1] ngehen in/ und würde doch uichts **lernen**[2]...



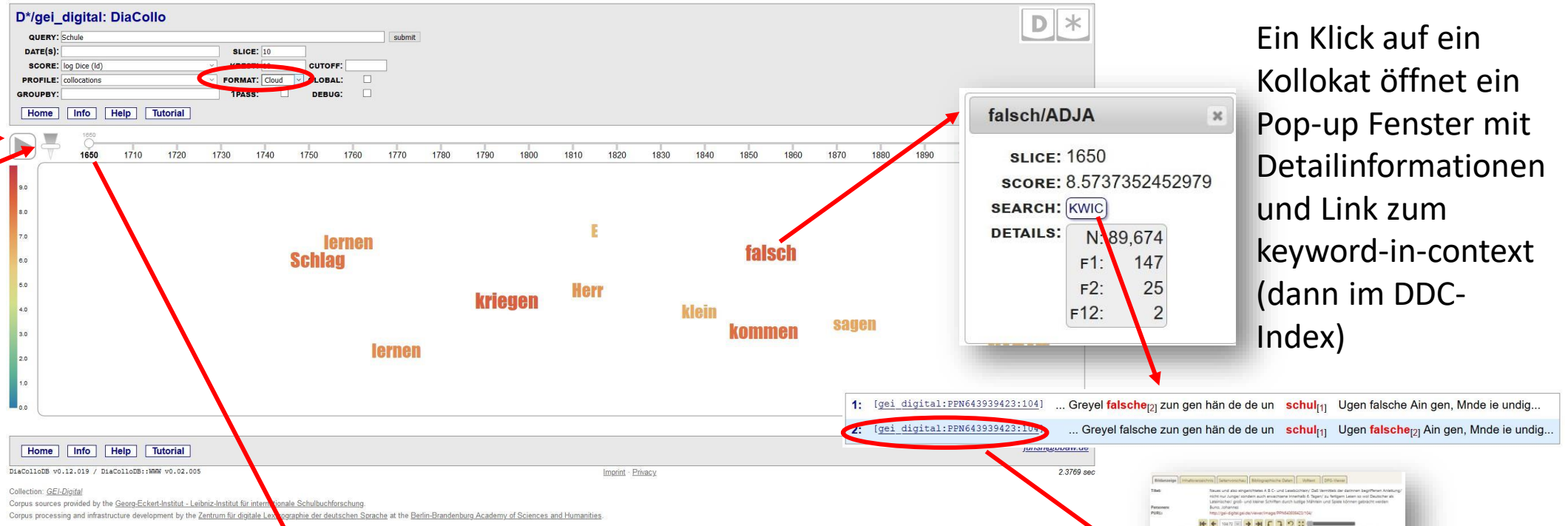
Alternativ können Lemmata zusammengefasst werden, auch wenn sie mit mehreren PoS-Tags vorkommen (**GROUPBY**: l).



Es kann auch ausgewählt werden, welche Wortarten überhaupt in die Berechnungen einbezogen werden.
In diesem Beispiel z.B. nur Imperative
(**GROUPBY**: l,p=VVIMP)

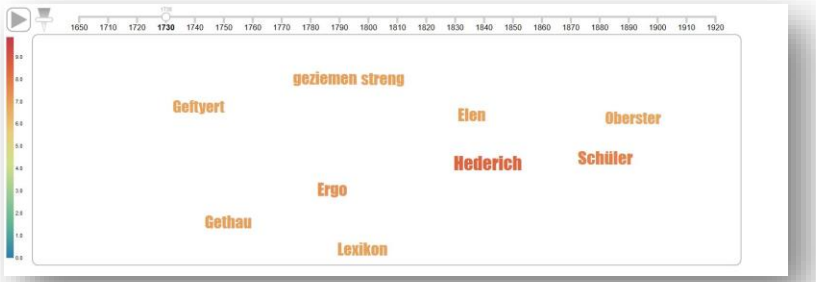
Beispielanfrage und Ergebnisanzeige im Cloud-Format in DiaCollo

Start/Stopp und Geschwindigkeit der Animation



Ein Klick auf ein Kollokat öffnet ein Pop-up Fenster mit Detailinformationen und Link zum keyword-in-context (dann im DDC-Index)

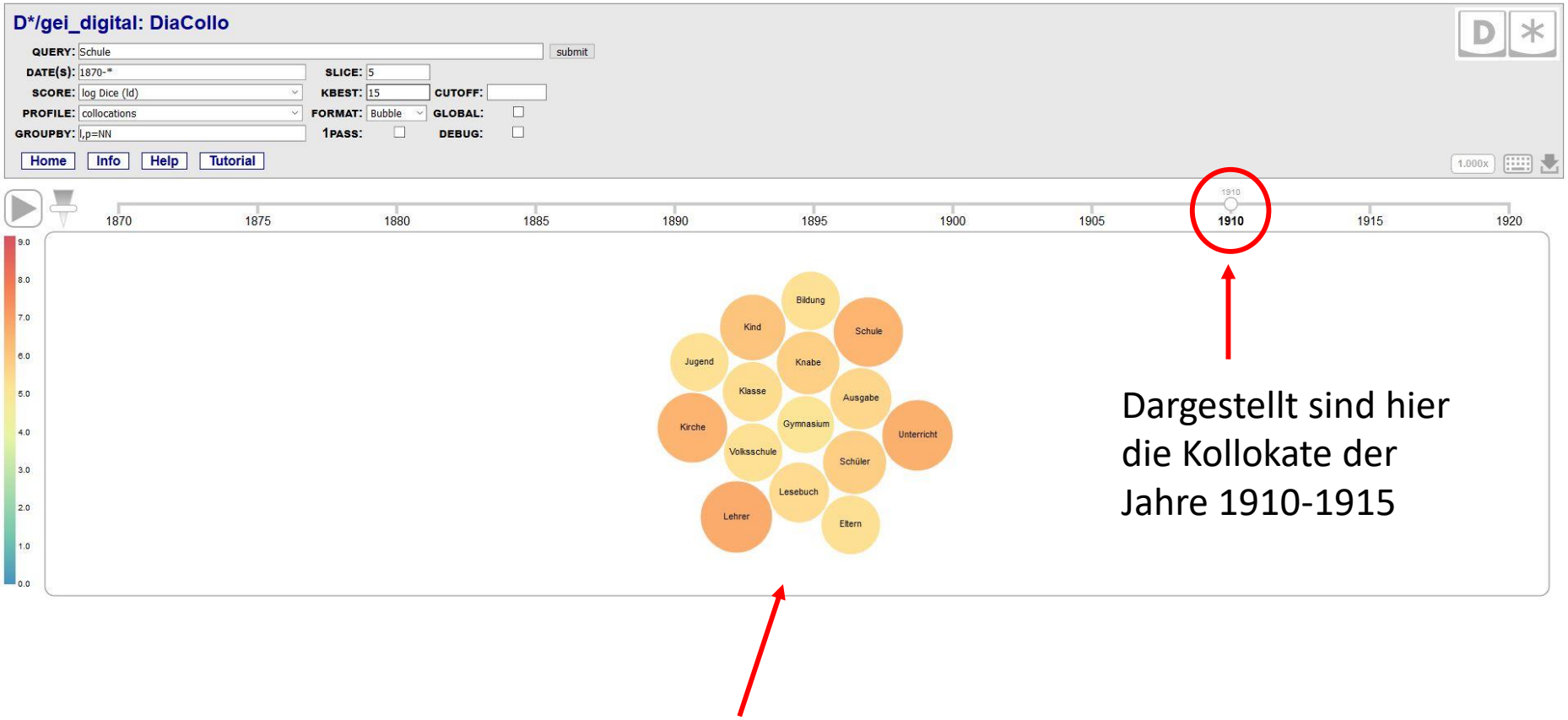
Per Schieberegler auf der Zeitleiste kann die Visualisierung auch interaktiv gesteuert werden



Beispielanfrage und Ergebnisanzeige im Bubble-Format in DiaCollo

Hier formulierte Beispiel-Anfrage:

Suche nach Kollokaten (**SCORE**=log Dice, **PROFILE**= collocations) von „Schule“ (**QUERY**= Schule) in Werken mit Publikationsdatum von 1870 bis zum spätesten Publikationsdatum im Korpus (**DATE(S)**= 1870-*) in 5-Jahres-Abschnitten (**SLICE**=5). Analysiere nur Nomen (**GROUPBY**= l,p=NN). Stelle die maximal 15 engsten Kollokate (**KBEST**=15) als Kreise (**FORMAT**=bubbles) auf der interaktiven Zeitleiste dar.

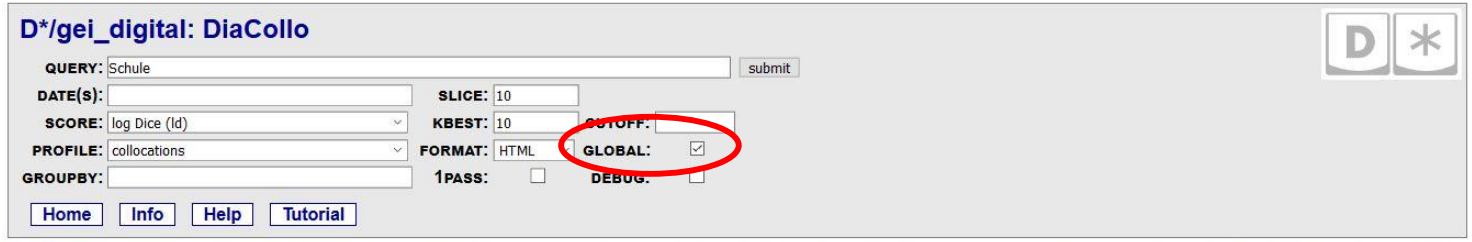


Dargestellt sind hier die Kollokate der Jahre 1910-1915

Doppelklick auf ein Kollokat führt wiederum zur Detailansicht

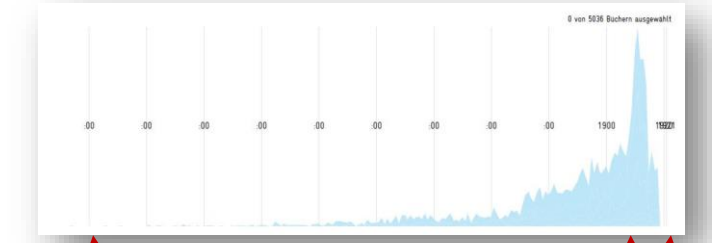
Die GLOBAL-Option

GLOBAL ermittelt die *k* besten Kollokationen zum gewählten Stichwort *im gesamten Korpus* statt in den unter **SLICE** gewählten Zeitabschnitten. Angezeigt wird dann die jeweilige Stärke dieser *k* Kollokationen *innerhalb* der im Parameter **SLICE** gewählten Zeitabschnitte.



Raw URL: http://diacollo.gei.de:8082/dstar/gei_digital/diacollo/profile.perl?query=Schule&format=html&score=ld&eps=0&slice=10&diff=adiff&global=1&kbest=10&profile=2

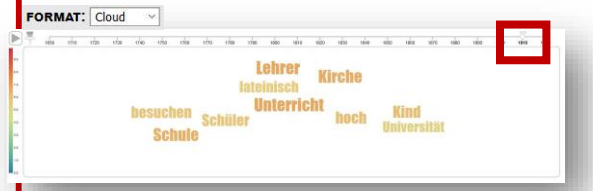
N	f1	f2	f12	score	label	lemma	pos	
89674	147	0	0	0.0000	1650	Kirche	NN	KWIC
89674	147	0	0	0.0000	1650	Kind	NN	KWIC
89674	147	0	0	0.0000	1650	Lehrer	NN	KWIC
89674	147	0	0	0.0000	1650	hoch	ADJA	KWIC
89674	147	0	0	0.0000	1650	Schule	NN	KWIC
89674	147	0	0	0.0000	1650	besuchen	VVFIN	KWIC
89674	147	0	0	0.0000	1650	Schüler	NN	KWIC
89674	147	0	0	0.0000	1650	lateinisch	ADJA	KWIC
89674	147	0	0	0.0000	1650	Unterricht	NN	KWIC
89674	147	0	0	0.0000	1650	Universität	NN	KWIC
348554948	30769	86306	732	6.7871	1910	Lehrer	NN	KWIC
348554948	30769	56303	621	6.7652	1910	Unterricht	NN	KWIC
348554948	30769	263665	1184	6.6200	1910	Kirche	NN	KWIC
348554948	30769	130769	728	6.5111	1910	Schule	NN	KWIC
348554948	30769	681932	1756	6.1457	1910	hoch	ADJA	KWIC
348554948	30769	621940	1585	6.1085	1910	Kind	NN	KWIC
348554948	30769	65030	354	5.8886	1910	Schüler	NN	KWIC
348554948	30769	30225	291	5.8882	1910	besuchen	VVFIN	KWIC
348554948	30769	26838	241	5.6469	1910	lateinisch	ADJA	KWIC
348554948	30769	40717	202	5.2705	1910	Universität	NN	KWIC
244620	163	163	4	8.6513	1920	Schule	NN	KWIC
244620	163	1013	7	7.6077	1920	Kind	NN	KWIC
244620	163	0	0	0.0000	1920	lateinisch	ADJA	KWIC
244620	163	0	0	0.0000	1920	Schüler	NN	KWIC
244620	163	0	0	0.0000	1920	besuchen	VVFIN	KWIC



1650er

1910er

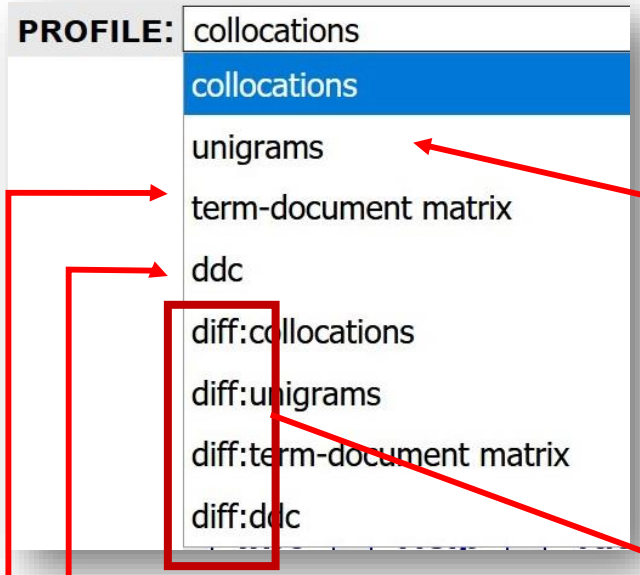
1920er



Bitte beachten Sie: Die Anzahl der untersuchten Tokens (*N*) ist im „GEI-Digital-2020“- Korpus nicht für jedes Jahr gleich. Vor 1871 und nach 1918 sind vergleichsweise weniger Daten vorhanden (vgl. die hier dargestellte Visualisierung der Metadaten http://diacollo.gei.de/gei-digital-2020/visualized/#/Stream). Deshalb sind die Kollokationen aus datenreichen Jahren bei der Berechnung im GLOBAL-Modus immer „im Vorteil“ und erscheinen aus statistischen Gründen stärker, bzw. „globaler“.

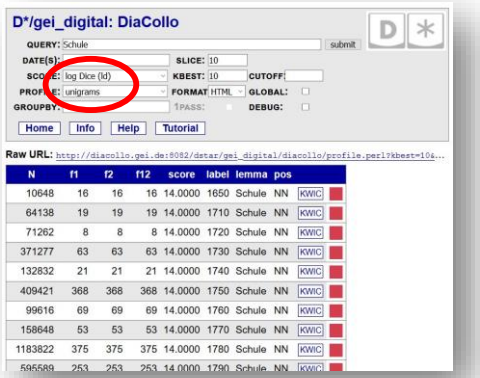
PROFILE-Optionen in DiaCollo

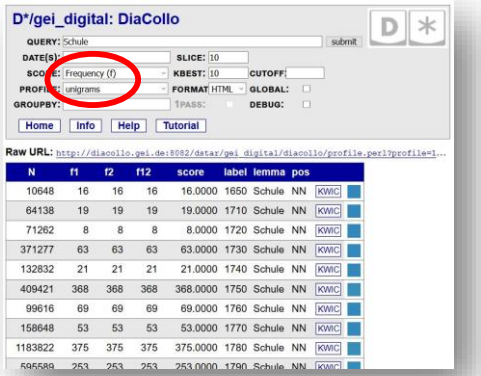
DiaCollo bietet verschiedene sog. Profile, d.h. Methoden zur Erfassung der Rohdaten (Volltexte), auf deren Basis dann (mit den gewählten SCORE-Funktionen) statistisch signifikante Kollokationen bewertet, eingestuft und ausgewählt werden können (vgl. http://diacollo.gei.de/gei-digital-2020/diacollo/help.perl#profiles). Derzeit unterstützt DiaCollo die folgenden Profiltypen:



collocations: Dies ist das Basisprofil für die Kollokationsanalyse: Ermittelt werden die Wörter, die am häufigsten und (vor allem) auch häufig innerhalb eines bestimmten Abstandes vom Suchbegriff vorkommen.

unigrams: Ermittelt werden alle Vorkommen des Suchbegriffs. Beispiel rechts: Im *GEI-Digital-2020* Korpus kommt „Schule“ in den 10.648 Token des Publikationszeitraums 1650-1659 insgesamt 16 mal vor, in den Jahren 1710-1719 in 64.138 Token 19 mal usw. Entspricht der absoluten Frequenz, die auch mit **SCORE: Frequency (f)** ermittelt wird.



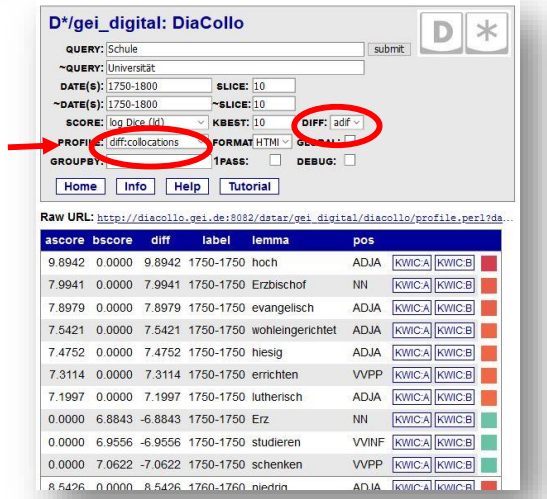


ddc: die Suchanfragen werden an eine DDC Suchmaschine geschickt, die den DDC-Index nutzt, siehe Beispiel auf der nächsten Folie.

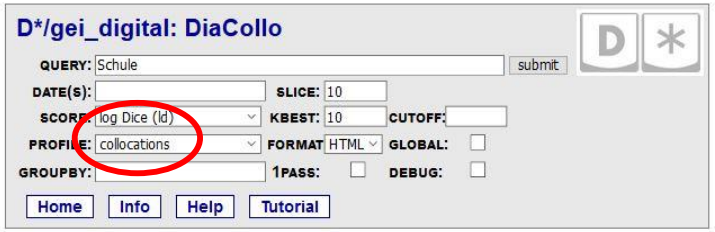
Term-document matrix (tdf): Ermittelt Kollokationen unter Nutzung einer Term-Dokument-Matrix. Ermöglicht flexiblere Abfragen und Ergebnismengenaggregation als die einfachen Kollokationsprofile, ist aber im Allgemeinen langsamer in der Auswertung und weniger empfindlich gegenüber Proximity-Effekten.

diff:... Für Vergleiche zweier Stichwörter mit den jeweils gewählten Profilen.

Rechts ein Beispiel für **DIFF:collocations** mit der Standardeinstellung **DIFF: adiff.** Dabei vergleicht **DIFF:collocations** die Kollokate zweier Stichworte, hier „Schule“ (Stichwort A) und „Universität“ (Stichwort B). Die Auswahl von **adiff** hebt die deutlichsten Unterschiede der Kollokationsstärke hervor (bei der Berechnung wird kein Pruning, d.h. keine Anwendung von „kbest“ genutzt, bevor die Unterschiede berechnet werden). Wenn Sie Kollokationen berechnen wollen, die beiden Stichworten gleichzeitig besonders nah stehen, sollten Sie eine **diff**-Operation wie **min** (wenn Sie sicher sind, dass Sie genug Daten haben) oder **havg** (wenn Ihr Korpus spärlich ist oder Ihre Suchbegriffe und/oder deren Kollokate keine hochfrequenten Elemente sind) wählen, siehe: http://diacollo.gei.de/dstar/gei_digital/diacollo/help.perl#diffs



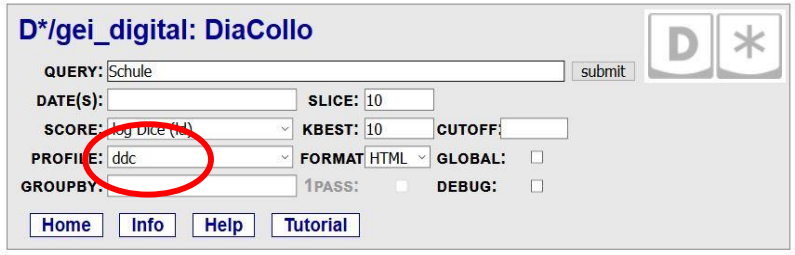
Unterschiedliche Ergebnisse bei der Nutzung von PROFILE: collocations und PROFILE: ddc



Raw URL: http://diacollo.gei.de:8082/dstar/gei_digital/diacollo/profile.perl?eps=0&que...

N	f1	f2	f12	score	label	lemma	pos	
89674	147	19	2	8.6250	1650	kriegen	VVFIN	KWIC
89674	147	25	2	8.5737	1650	falsch	ADJA	KWIC
89674	147	312	4	8.1576	1650	kommen	VVFIN	KWIC
89674	147	88	2	8.1235	1650	Schlag	NN	KWIC
89674	147	203	2	7.5488	1650	lernen	VVINF	KWIC
89674	147	205	2	7.5406	1650	lernen	VVPP	KWIC
89674	147	262	2	7.3240	1650	Herr	NN	KWIC
89674	147	430	2	6.8276	1650	sagen	VVFIN	KWIC
89674	147	506	2	6.6491	1650	klein	ADJA	KWIC
89674	147	664	2	6.3364	1650	E	NE	KWIC
459360	152	45	2	8.3779	1710	Jus	NE	KWIC
459360	152	64	2	8.2451	1710	Jugend	NN	KWIC
459360	152	70	2	8.2056	1710	gebürtig	ADJD	KWIC
459360	152	115	2	7.9393	1710	Präge	NN	KWIC
459360	152	126	2	7.8811	171	öffentlich	ADJA	KWIC
459360	152	683	2	6.2944	1710	R	NE	KWIC
459360	152	1593	3	5.8160	1710	berühmt	ADJA	KWIC
459360	152	1073	2	5.7414	1710	Kirche	NN	KWIC
459360	152	1840	2	5.0400	171	T	NE	KWIC
459360	152	3542	2	4.1490	1710	Jahr	NN	KWIC
597446	78	434	3	7.5850	1720	hoch	ADJA	KWIC
597446	78	283	2	7.5041	1720	Buch	NN	KWIC
597446	78	336	2	7.3065	1720	Gesetz	NN	KWIC

In diesem Beispiel zeigt die Spalte **N**: Für die 1710er Jahre enthält der DiaCollo Index (links) 459.360 Token, der DDC Index (rechts) hingegen 2.408.950 Token.



Raw URL: http://diacollo.gei.de:8082/dstar/gei_digital/diacollo/profile.perl?diff=adif...

N	f1	f2	f12	score	label	lemma	pos	
326080	70	400	2	7.4285	1650	kommen	VVFIN	KWIC
2408950	110	160	2	7.9232	1710	öffentlich	ADJA	KWIC
2408950	110	2510	2	4.6446	1710	T	NE	KWIC
1955520	30	480	2	7.0656	1720	hoch	ADJA	KWIC
14533590	400	90	2	7.0634	1730	Oberster	NN	KWIC
14533590	400	440	2	6.2858	1730	Ergo	NE	KWIC
14533590	400	650	2	5.9638	1730	ober	ADJA	KWIC
14533590	400	870	2	5.6894	1730	richten	VVFIN	KWIC
14533590	400	870	2	5.6894	1730	treiben	VVPP	KWIC
14533590	400	1130	2	5.4207	1730	recht	ADJD	KWIC
14533590	400	1170	2	5.3835	1730	fangen	VVFIN	KWIC
14533590	400	1530	2	5.0856	1730	Sprache	NN	KWIC
14533590	400	1660	2	4.9916	1730	Lateinische	NN	KWIC
14533590	400	3170	2	4.1983	1730	Athen	NE	KWIC
4330010	100	350	2	7.1862	1740	Kloster	NN	KWIC

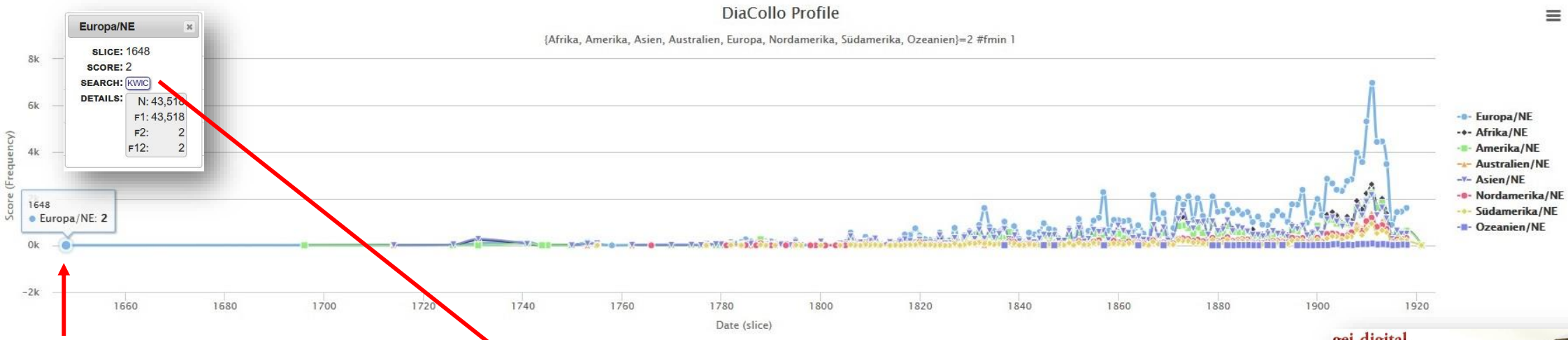
PROFILE: ddc

Die Suchanfragen werden an eine DDC Suchmaschine geschickt, dürfen also Elemente der DDC-Abfragesprache beinhalten; gesucht wird zudem im vom DDC genutzten Index, der im Gegensatz zum DiaCollo-Index *alle* Token – also auch für DiaCollo meist wenig relevante aber sehr häufige (Stop-)Wörter – umfasst. Dies verringert die relative Häufigkeit der untersuchten Begriffe. Der DDC- Profiltyp ist erheblich langsamer und aufwändiger in der Berechnung.

Vgl. hierzu auch die Informationen zum DDC-Profiltyp unter „Help“

Frequenzvergleich in DiaCollo

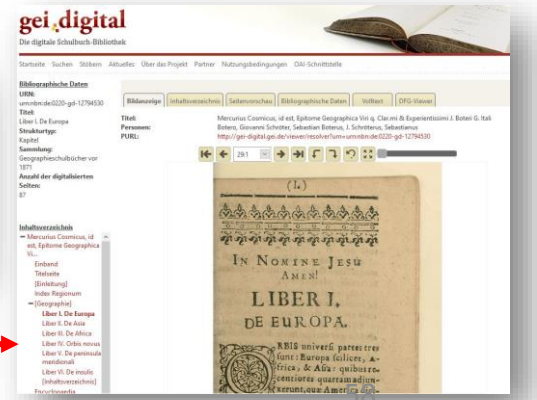
FORMAT: Highchart, **SCORE**: Frequency, **PROFILE**: ddc, **GROUPBY**: l,p=NE [Link zu diesem Beispiel]



Beim **Darüber-fahren** mit der Maus werden zu den einzelnen Punkten weitere Informationen eingeblendet.

Ein **Klick** auf die Datenpunkte öffnet Detailinformationen und Link zum KWIC (über eine DDC-Query)

KWIC-Ansicht mit Link zur Quelle

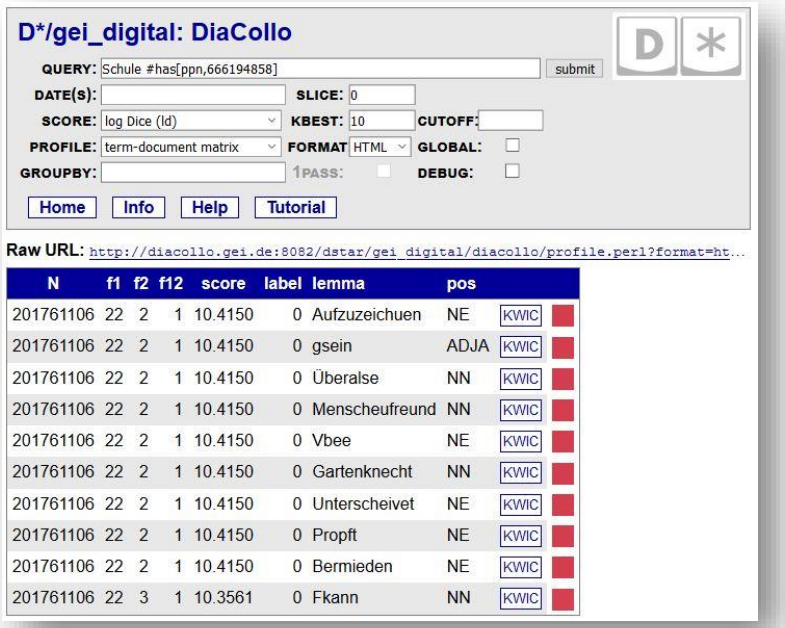


Kollokationen innerhalb eines Werkes / synchrone Kollokationen

Kollokationen in Werken eines bestimmten Jahres können durch entsprechende Auswahl des Jahres im DATE(S) Parameter gesucht werden. Wenn nur innerhalb *eines* bestimmten Werkes gesucht werden soll, ist dies möglich mit dem **TDF** oder **DDC**-Profiltyp und entsprechender #HAS Klauseln. Hier z. B. die Suche nach Kollokaten zu „Schule“ in Rochows Kinderfreund von 1798 mit dem Persitensten Identifier ppn 666194858:

TDF:
- QUERY: Schule #has[ppn,666194858]
- SLICE: 0
- PROFILE: term-document matrix (tdf)

... auf Paragraphen-Ebene. N ist immer noch die Größe des Gesamtkorpus; f1, f2, f12 sind auf das Buch (hier per PPN) eingeschränkt.

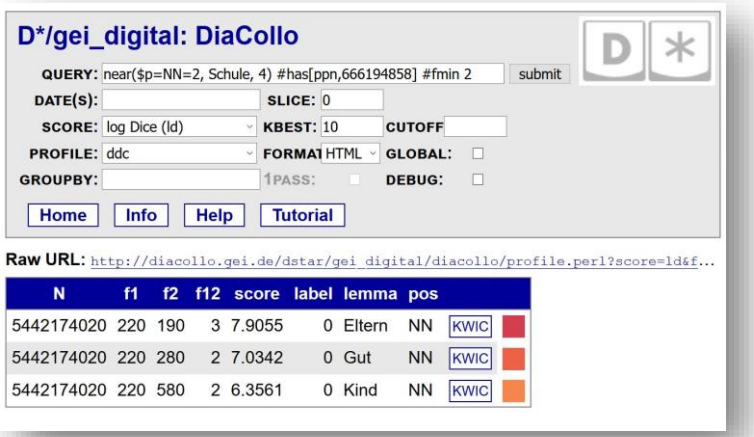


N	f1	f2	f12	score	label	lemma	pos	
201761106	22	2	1	10.4150	0	Aufzuzeichuen	NE	KWIC
201761106	22	2	1	10.4150	0	gsein	ADJA	KWIC
201761106	22	2	1	10.4150	0	Überalse	NN	KWIC
201761106	22	2	1	10.4150	0	Menscheufreund	NN	KWIC
201761106	22	2	1	10.4150	0	Vbee	NE	KWIC
201761106	22	2	1	10.4150	0	Gartenknecht	NN	KWIC
201761106	22	2	1	10.4150	0	Unterscheivet	NE	KWIC
201761106	22	2	1	10.4150	0	Propft	NE	KWIC
201761106	22	2	1	10.4150	0	Bermieden	NE	KWIC
201761106	22	3	1	10.3561	0	Fkann	NN	KWIC

Link zum TDF-Beispiel

DDC:
- QUERY: near($p=NN=2, Schule, 4) #has[ppn,666194858] #fmin 2
- SLICE: 0
- PROFILE: ddc

Gesucht werden hier nur Nomenkollokate ($p=NN=2), höchstens 4 Tokens zwischen Kollokat & Kollokant (NEAR(...,4)), mit minimaler Kookkurrenzfreqenz 2 (#fmin 2 --> f12 >=)



N	f1	f2	f12	score	label	lemma	pos	
5442174020	220	190	3	7.9055	0	Eltern	NN	KWIC
5442174020	220	280	2	7.0342	0	Gut	NN	KWIC
5442174020	220	580	2	6.3561	0	Kind	NN	KWIC

Hier ist wieder N die Tokenanzahl des Gesamtkorpus; f1, f2, und f12 sind auf das Buch eingeschränkt (x10, weil es 10 Ko-okkurrenz Paare pro Token gibt: 5 links + 5 rechts von „Schule").

Weil "**N**" bei diesen Beispielen auf das Gesamtkorpus bezogen wird, sind die "**score**" Werte u.U. nicht direkt vergleichbar mit denjenigen, die bzgl. anderer Korpus-Teilmengen (z.B. andere Einzelbücher, oder das Gesamtkorpus selber) berechnet wurden; sie sollten aber untereinander vergleichbar sein.
Wann und warum dies jeweils (nicht) der Fall ist, hängt von der verwendeten Score-Funktion ab.

Wir hoffen, Ihnen mit diesen Folien Lust auf die Erkundung von GEI-Digital-2020 und anderen Korpora mit der D*-Umgebung gemacht zu haben. Vielen Dank für Ihr Interesse! Das „DiaCollo für GEI-Digital“-Team wünscht viel Spaß und interessante Befunde!

Weiterführende Links:

- DiaCollo Tutorial von Lothar Lemnitzer, Bryan Jurish und Daniel Burkhardt
- „Korpussuche – Suchmaschine und Suchabfragesprache“ des Digitalen Wörterbuchs der Deutschen Sprache (DWDS)
- Dokumentation zur DDC Abfragesprache von Bryan Jurish
- Antworten auf knifflige Fragen zu DiaCollo von Bryan Jurish
- Projekt „DiaCollo für GEI-Digital“ am Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
- Verschiedene (Referenz-) Korpora des Zentrums Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften

Feedback und Fragen gerne an: nielaender@leibniz-gei.de

gei.Digital

GEORG ECKERT INSTITUT
Leibniz-Institut für internationale Schulbuchforschung