

Gebrauchsorientierte Sprachbeschreibung: Korpuslinguistische Analysen von (Fast-)Synonymen

Heike Zinsmeister
Seminar „Kreatives Schreiben“
Universität Vilnius
25.10.2017

Online: <https://www.korpuslab.uni-hamburg.de/lehrmaterialien/diverses.html>



Foto: UHH/Scharf-Bluhm

- Prof. Dr. Heike Zinsmeister
 - Linguistik des Deutschen mit dem Schwerpunkt Korpuslinguistik
 - Universität Hamburg



Hamburg



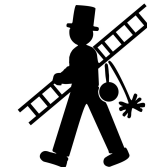
(Fast-)Synonyme

- Definition

Als **Synonyme** (von altgriechisch συνώνυμος *synōnymos*, bestehend aus σύν *syn* „zusammen“ und ὄνομα *ónoma* „Name“) bezeichnet man verschiedene sprachliche oder lexikalische Ausdrücke oder Zeichen, die den gleichen oder einen sehr ähnlichen Bedeutungsumfang haben. Die Beziehung zwischen Synonymen wird **Synonymie** genannt. (Wikipedia)



- Totale Synonymie: in allen syntagmatischen Kontexten austauschbar.
 - Gehweg, Bürgersteig, (Trottoir?)
 - Schornsteinfeger, Kaminkehrer
- Partielle Synonyme: nicht in allen Kontexten äquivalent
 - Polizist, Bulle



Synonym?

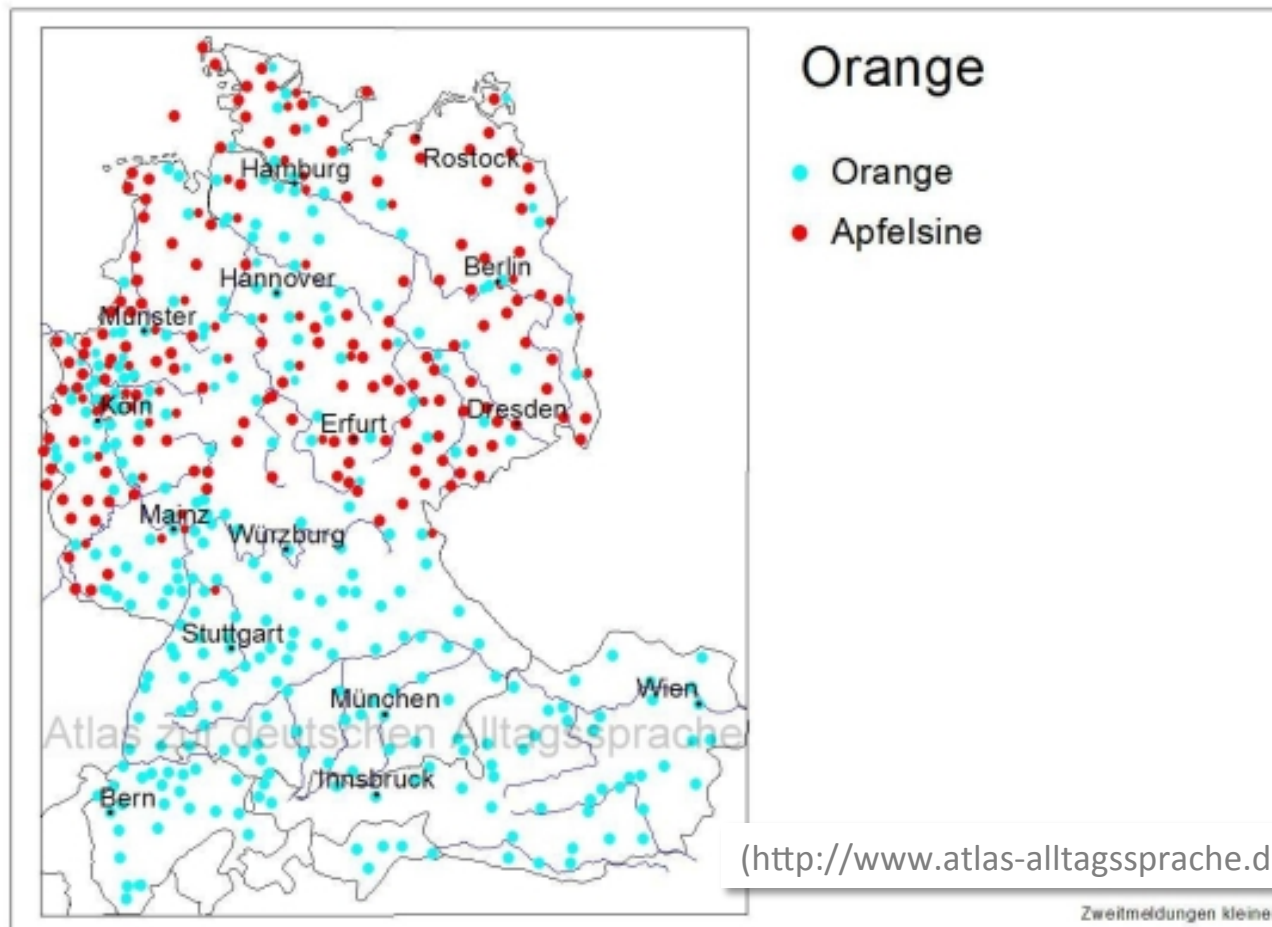
- Apfelsine
- Orange





Synonym?

- Im nieder- und mitteldeutschen Sprachgebiet isst man *Apfelsinen* und im oberdeutschen und österreichischen Sprachraum *Orangen*.





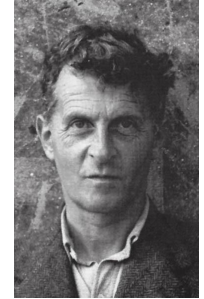
Synchrone Fast-Synonymie diachron erklärt

- Mittel- und Norddeutschland
 - Einfuhr durch den Seehandel der Holländer; Früchte stammen ursprünglich aus China
 - (1) niederländisch *sinas appel* ‚China Apfel‘
 - (2) niederdeutsch *Appelsina*
- Oberdeutscher und österreichischer Sprachraum
 - Einfuhr durch den Fernhandel der Mittelmeerländer; Übernahme der romanischen Form, die auf Arabisch/Persisch basierte
 - (1) arabisch/persisch *nāranġ(a)*
 - (2) italienisch *arancia*; französisch *orange*

(Pfeifer 2005: 50 u. 953, nach Ganslmayer 2014: 16)

Bedeutung

- Ludwig Wittgenstein (1953):
„Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache“
- John Rupert Firth (1957):
„You shall know a word by the company it keeps.“
- Korpuslinguistische Methode:
 - Ermittlung der Verteilung (Distribution) i Korpora
 - Kollokationen
 - Das überproportionale Miteinandervorkommen zweier oder mehrerer beliebiger Wörter oder lexikaler Einheiten



Korpuslinguistische Ressourcen

- DWDS: Kernkorpus des 20. Jahrhunderts
 - www.dwds.de
- Deutsches Referenzkorpus DeReKo des Instituts für deutsche Sprache in Mannheim (IDS)
 - Onlinezugriff:
<https://cosmas2.ids-mannheim.de/cosmas2-web/>

Definition: Korpus

Ein Korpus ist eine **Sammlung schriftlicher oder gesprochener Äußerungen**. Die Daten des Korpus sind typischerweise **digitalisiert**, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den **Daten** selber sowie möglicherweise aus **Metadaten**, die diese Daten beschreiben, und aus linguistischen **Annotationen**, die diesen Daten zugeordnet sind.

DWDS: Kernkorpus des 20. Jahrhunderts

- Zeitlich und nach Textsorten ausgewogenes Korpus des gesamten 20. Jahrhunderts
- Textsorten: Belletristik, Gebrauchsliteratur, Wissenschaft, Journalistische Prosa
- Umfang: ca. 100 Millionen Textwörter in 79.830 Dokumenten.
- Die Annotation der Daten und Metadaten erfolgt in XML nach den TEI-Richtlinien

Beispiel: DWDS-Kernkorpus 20

← → ↻ www.dwds.de/?qu=Gutmensch ☆ ⬇ ☰

Anmelden | Registrierung beim DWDS Mobil | Hilfe zur Suche | Kontakt | Impressum

D W D S Ressourcen ▼ Erschließung ▼ Projekt ▼ Aktuelles ▼

Gutmensch 🔍 DWDS Standardsicht +Ressourcen ↻

Kernkorpus 20 🗑️ 🖨️ ? ✕

Treffer: 14, davon anzeigbar: 8 Filter

KWiC	Datum ↓	Datum ↑	Zufällig	Links	Rechts
1 1999 G kurz					
2 1999 G kurz					
3 1999 G schwani					
4 1999 G schwani					
5 1999 Z zeit199					
6 1998 G kellner					
7 1998 G kellner					
8 1997 Z zeit199					

Version: 1.1 Optionen

KWIC-Darstellung = Keyword in Context / Konkordanz

DWDS: Mustersuche zu Kollokationen

\$p=ADJA Europa 🔍 ?

Korpus: DWDS-Kernkorpus
Start: 1900 **Ende:** 1999
Textklassen: Belletristik Wissenschaft Gebrauchsliteratur Zeitung
Anzeige: KWIC voll maximal
Sortierung: Datum aufsteigend
Anzahl Treffer pro Seite: 10

1–10 von 553 Treffern [Export als: [TSV](#)]

⏪ ⏩ 1 2 3 4 5 → →

1:	w	1901	altertum_me	...piel diplomatischer Intrigen, derselben Art, wie es das moderne Europa im 17. und 18. Jahrhundert und auch in unsere...	☰
2:	z	1902	bt_m190202	Man hat es heute mit einem neuen Europa und einem alten Amerika zu thun . Das neue E...	☰
3:	z	1902	bt_m190202	...neuen Europa und einem alten Amerika zu thun . Das neue Europa ist ein Land von noch nicht entwickelten Möglic...	☰
4:	B	1902	suttner_roma	... Szene der fernen Zukunft, da man in dem vergletschert gewesenen Europa alte Funde ausgräbt und darüber Forschungen ...	☰
5:	w	1902	altertum_me	...er Usurpatoren und Monarchen, welche die Staaten des modernen Europa geschaffen haben.	☰
6:	G	1902	peters_goldla	...olchen Ländern ist die Gastfreundschaft grösser als im harten Europa , und im Macequege-Distrikt herrscht ein starke...	☰
7:	B	1903	heyking_roma	... persönliches Verantwortungsgefühl ausgebildet, als im alten Europa .	☰
8:	G	1903	autobiograph	In diesem zur Zeit nationalistisch zerrissenen Europa käme ich mir selbst wie ein Judas an meinem F...	☰

Suche im DWDS-Wortprofil

Lemma:

Anzahl: **min. logDice:**

Überblick

Erziehung Fächern **Kopftuch Kultus** Qualität
Schulen Schüler ausfällt bilingualen erteilt
fächerübergreifenden
muttersprachlichen
naturwissenschaftlichen normalen
polytechnischen regulären schwänzen
theoretischen zweisprachigen

Suchwort: „Kralle“ –Typische Verbindungen

Überblick

Pfoten Raubvögel Reißzähne
Samtpfoten **Schnabel** Tatzen
Zehen Zähne **ausfährt**
ausgefahrenen
eingezogenen gebogene gekrümmten
lackierten **messerscharfen**
schärfen sichelförmigen spitzen
wetzen zurückziehbaren

ist in Präpositionalgruppe

Bären mit Füße mit Katze mit
befindet in befreien aus festhalten mit greift mit
hält in packt mit trägt in verteidigt mit

ist Subjekt von

hat Adjektivattribut

ausgefahrenen
ausgestreckten blutigen
eingezogenen eisernen
gebogene **gekrümmten**
gelben gewaltigen kräftige **lackierten**
langen **messerscharfen**
scharfen **sichelförmigen**
spitzen stählerne tödliche
zarten **zurückziehbaren**

hat Genitivattribut

Adlers Katze Tieres Tigers

hat Präpositionalgruppe

ist Akk./Dativ-Objekt von

ausfährt **ausstreckte** **eingezogen**
schlägt schneiden schärfen spüren
stutzt trägt **wetzen** zeigen

in Koordination mit

Augen Beine **Federn** **Fell** **Flügel** **Kopf**
Pfoten **Schnabel** **Schwanz**
Zehen **Zunge** **Zähne**

ist Genitivattribut von

Syntaktische Dependenzrelationen

- Relationstypen innerhalb von Phrasen
 - Adjektiv-Nomen (Etikett: ATTR): *klares Feindbild, schöne Bescherung*
 - Nomen-Koordination-Nomen (CJ): *grün und blau, Kopf oder Zahl*
 - Nomen-Nomen (im Genitiv) (GMOD): *Abbau des Feindbildes, Abbau von Vorurteilen*
 - Adverb-Adjektiv (ADVA): *sehr intelligent, hoch erfreut*
 - Phrasenübergreifende Relationen
 - Nomen -Verb (SUBJ): *das Feindbild verblasst, das Badewasser läuft aus*
 - Nomen-Verb (PSUBJ): *die Flasche wurde entkorkt*
 - Nomen -Verb (OBJA): *ein Feindbild abbauen, eine Rede halten*
 - Nomen -Verb (OBJD): *dem Publikum verkünden*
 - Verb-Präposition-Nomen (V PP): *zur Verfügung stehen, auskommen ohne*
- Feindbild
- Verb-Verb (INFOBJ): *aufgehen sehen, (auf etw.) zu sprechen kommen*
 - Adverb Verb (VADV): *schallend lachen, freimütig zugeben*

Hintergründe des Wortprofils

- Automatische Ermittlung von syntakt. Relationen
 1. Satzsegmentierung und morphologische Analyse mit der TAGH-Morphologie (Geyken & Hanneforth 2005)
 2. Ermittlung der Phrasen und Annotation der darin enthaltenen Modifizierer/Koordinationsfunktionen
 3. Ermittlung von dependenten bzw. koordinierenden Phrasen
 4. Ermittlung von Nebensätzen und Annotation der darin enthaltenen syntaktischen Köpfe
 5. Ermittlung von Hauptsätzen und Annotation der darin enthaltenen syntaktischen Köpfe

Dependenzbeispiel



COSMAS II



Online-Korpusschnittstelle der geschriebenen
Korpora des Instituts für Deutsche Sprache
(IDS) in Mannheim

<http://www.ids-mannheim.de/cosmas2/>

COSMAS II - Übersicht

COSMAS II_{web}

Online-Hilfe

Entwicklung

COSMAS II_{web} starten

Browserunterstützung


Die Applikation COSMAS II_{web}

Übersicht

COSMAS II_{web} ist eine betriebssystemunabhängige WWW-Applikation, die die Ko-
WINDOWS-Betriebssystem gebunden zu sein oder COSMAS II_{win} installieren zu

COSMAS II_{web} verfügt - mit Ausnahme der grafischen Suchanfragekomponente -

Aktuelle Version: 1.7

 Applikation **starten**.

Detaillierte Informationen

- Freigabeinformationen: **aktuelle Version**, ältere Versionen
- Liste der **unterstützten Browser**



		Einstieg Anmeldung Recherche Optionen			
Portal COSMAS II		Liebe Benutzer des COSMAS II-Webdienstes,			
		Einstieg Anmeldung Recherche Optionen			
Online-Hilfe		<p>Benutzungshinweise</p> <p>Die Benutzung von COSMAS II ist kostenlos und nur für nicht-kommerzielle Zwecke erlaubt. Beacht "Registrieren" die Nutzungsbeschränkungen.</p>			
Forum	Login				
Kontakt	Logout				
Browserunter	Registrieren				
	Passwort vergessen ?				
		Einstieg Abmeldung Recherche Optionen Hilfe			
Archiv		<p>Recherche</p> <p>Das normale Vorgehensmodell bei einer Recherche ähnelt einem Wasserfallmodell. Die einzelnen Schritte sind unten dargestellt. Selbstverständlich können sie im Modell zurückspringen. Änderungen in voran gestellten Aktionen wirken sich aber auf nachstehende aus. So löscht die Auswahl eines neuen Archivs beispielsweise die frühere Auswahl eines Korpus, sodass sie ein neues Korpus wählen müssen.</p> <ol style="list-style-type: none"> 1. Archiv wählen ↳ 2. Korpus wählen ↳ 3. Anfrage formulieren ↳ 4. Wortformliste bearbeiten ↳ 5. Ergebnisanzeige ↳ 6. KWIC ↳ 7. Export von Ergebnissen 			
Korpus					
Suchanfrage					
Wortformliste					
Ergebnisse					
Kookkurrenzanalyse					
Gesamt-KWIC					
Gesamt-Volltext					
Export					





Suche

Lemmatisierung

Korpuspräsentation

Ergebnispräsentation

KWIC/Volltext

Newsletter

Die Einstellungen, die Sie in COSMAS II vornehmen können, sind in unterschiedliche Kategorien gegliedert:

- **Suche**
Damit können Sie darauf Einfluss nehmen, ob die Groß-/Kleinschreibung berücksichtigt werden soll, ob Expansionslisten (Wortformlisten) beim Einsatz von Platzhalter- und Lemmatisierungsoperatoren erzeugt und wie sie sortiert werden sollen; ob die Treffer mit einem Zufallsgenerator reduziert werden sollen.
- **Lemmatisierung**
Damit können Sie Einfluss nehmen auf die Arbeitsweise des Lemmatisierungsoperators &. Die Flexionsformen bleiben immer eingeschaltet.
- **Korpuspräsentation**
Mit diesen Einstellungen können Sie Einfluss auf die Sortierung der Korpuspräsentation nehmen.
- **Ergebnispräsentation**
Mit diesen Einstellungen können Sie Einfluss auf die Sortierung und die Darstellung der Ergebnispräsentation nehmen.
- **KWIC/Volltext**
Mit den KWIC- und Volltextoptionen können Sie Einfluss auf die Anzeige und die alphabetische Sortierung des KWICs und Volltexts nehmen.
- **Newsletter**
Wenn Sie den COSMAS II-Newsletter abonnieren, erhalten Sie eine Benachrichtigung, wenn eine neue Version von COSMAS II erschienen ist, die IDS-Korpora ergänzt wurden, etc.

Ihre Einstellungen werden beim Abmelden vom COSMAS II-Service **sitzungsübergreifend gespeichert.**

	Einstieg	Abmeldung	Recherche	Optionen
Archiv	Aktuelles Archiv	<input type="text" value="-"/>	Aktuelles Korpus	<input type="text" value="-"/>
Korpus	Suchanfrage	<input type="text" value="-"/>		
Suchanfrage	Archivauswahl			
Wortformliste	Öffnen Sie bitte eines der folgenden Archive:			
Ergebnisse	W - Archiv der geschriebenen Sprache			
Kookkurrenzanalyse	Alte Version von "W - Archiv der geschriebenen Sprache" (verfügbar bis 13.05.2012)			
Gesamt-KWIC	W-ÜBRIG - Archiv der aussortierten geschriebenen Korpora			
Gesamt-Volltext	TAGGED-C - Archiv morphosyntakt. annotierter Korpora (CONNEXOR)			
Export	TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)			
	TAGGED-M - Archiv morphosyntakt. annotierter Korpora (MECOLB; ehem. TAGGED)			
	HIST - Archiv der historischen Korpora			
	GFDS - Kartei der Gesellschaft für deutsche Sprache			
	WK-PH - Archiv der phasengegliederten Wendekorpora			



Informationen zu den Archiven finden Sie unter [Organisation des Textmaterials.](#)



Einstieg

Abmeldung

Recherche

Optionen

Archiv

Korpus

→ vordefiniert

benutzerdefiniert

Korpusverwaltung

Suchanfrage

Wortformliste

Ergebnisse

Kookkurrenzanalyse

Gesamt-KWIC

Gesamt-Volltext

Export

Aktuelles
Archiv

TAGGED-T - Archiv morphosyntakt. annotierter Korpora (T

Aktuelles Korpus

-

Suchanfrage

-

Korpusauswahl

Laden Sie bitte eines der folgenden virtuellen Korpora:

Vordefinierte Korpora

TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T

brz - Braunschweiger Zeitung, September 2005 - Dezember 2009

bvz - Burgenländische Volkszeitung, Januar 2007 - Dezember 2009

haz - Hannoversche Allgemeine, August 2007 - Dezember 2009

hmp - Hamburger Morgenpost, Januar 2006 - Dezember 2009

mm - Mannheimer Morgen, Januar 2006 - Dezember 2009

non - Niederösterreichische Nachrichten, Januar 2007 - Dezember 2009





Einstieg

Abmeldung

Recherche

Optionen

Hilfe

Archiv

Korpus

Suchanfrage

Wortformliste

Ergebnisse

Kookkurrenzanalyse

Gesamt-KWIC

Gesamt-Volltext

Export

Aktuelles Archiv TAGGED-T - Archiv morphosyntakt. annotiert **Aktuelles Korpus** hmp - Hamburger Morgenpost, Januar

Suchanfrage &licht

Suche formulieren **Ergebnisse:**

Eingabe:
&licht

Weggelassener Verknüpfungsoperator bedeutet:

Wortabstand /+w1 logisches "ODER"

MORPH-Assistent **Suchen** **Suchanfragesyntax**

Beispiele:
direkt nacheinander [Hilfe zu den Suchanfragebeispielen](#)



Alle aktivieren

Alle deaktivieren

Liste zu: &licht, 12 Wortformen, alfab. aufsteigend

Öffnen

Ergebnisse

Wortformenliste: Deaktivierung von irrelevanten Formen

Alle aktivieren

Alle deaktivieren

Liste zu: &licht, 12 Wortformen, alfab. aufsteigend

Schließen

<input type="checkbox"/>	LICHT	16
<input type="checkbox"/>	Licht	2215
<input checked="" type="checkbox"/>	licht	7
<input type="checkbox"/>	Lichte	28
<input checked="" type="checkbox"/>	lichte	13
<input type="checkbox"/>	Lichten	1
<input checked="" type="checkbox"/>	lichten	27
<input type="checkbox"/>	LICHTER	7
<input type="checkbox"/>	Lichter	268
<input checked="" type="checkbox"/>	lichter	5
<input type="checkbox"/>	Lichtes	6
<input checked="" type="checkbox"/>	lichtes	3

Alle aktivieren

Alle deaktivieren

Schließen

Ergebnisse



- Archiv
- Korpus
- Suchanfrage
- Wortformliste

Aktuelles Archiv: TAGGED-T - Archiv morphosyntakt. annotiert
 Aktuelles Korpus: hmp - Hamburger Morgenpost, Januar
 Suchanfrage: &licht
 Korpusansicht Ergebnisse: 55

- Gesamt-KWIC**
- Gesamt-Volltext
- Export
- Kookkurrenzanalyse

- Ergebnisse**
- Quellenansicht
 - **Korpusansicht**
 - Dokumentansicht
 - Ansicht vor/seit
 - Zeitpunkt
 - Jahrzehntansicht
 - Jahresansicht
 - Monatsansicht
 - Tagesansicht
 - Länderansicht
 - Stat. KWIC-Ausw.


Treffer	Texte	von	bis	Korpus
<input type="checkbox"/>	17	17	2006	2006 HMP06 Hamburger Morgenpost 2006
<input type="checkbox"/>	19	19	2007	2007 HMP07 Hamburger Morgenpost 2007
<input type="checkbox"/>	7	7	2008	2008 HMP08 Hamburger Morgenpost 2008
<input type="checkbox"/>	12	11	2009	2009 HMP09 Hamburger Morgenpost 2009
	55	54	2006 2009	4 Korpora

- Gesamt-KWIC
- Gesamt-Volltext
- Export
- Kookkurrenzanalyse

- Ergebnisse
- Kookkurrenzanalyse
- Gesamt-KWIC**
- **unsortiert**
- alphabetisch
- chronologisch
- zufällige Sort.
- nach Kook.
- Gesamt-Volltext
- Export

Auswahl aktiv. Alle aktiv. Alle deaktiv. |< << Seite 1 von 1 >> >| Volltext

1	<input type="checkbox"/>	HMP09/FEB.01071	"Die Augenpartie wird mit einem lichten Zitruston in Kombination mit K
2	<input type="checkbox"/>	HMP09/FEB.01071	Sie einen Lidschatten in einem lichten Sommerton (Nr. 222).
3	<input type="checkbox"/>	HMP09/FEB.01325	bei Numerus-Clausus Fächern zu lichten .
4	<input type="checkbox"/>	HMP09/MAR.03275	Das Angebot ist groß: Unter dem lichten Glasdach der ehemaligen Maschi
5	<input type="checkbox"/>	HMP09/APR.00258	uchtungskonzepte kann unter www. licht . de
6	<input type="checkbox"/>	HMP09/JUN.01711	ist dank üppiger Verglasung so licht und luftig wie in kau...
7	<input type="checkbox"/>	HMP09/JUL.02943	als Rezitatoren, möchte er ein lichtes , poetisches Gesam...



WARNUNG:

Die statistische Kookkurrenzanalyse einer geringen Anzahl von Belegen (hier: 55) führt nur selten zu wirklich aussagekräftigen Ergebnissen.

Einstellungen

Kontext: Wörter links höchstens 1 Satz
 Wörter rechts Autofokus

Granularität:

Zuverlässigkeit:

Clusterzuordnung:

Lemmatisierung verwenden:

Funktionswörter ignorieren:

LLR-Wert anzeigen:

Nummerierung des Hauptkookkurrenzators:

Kookkurrenzanalyse

#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
+	1	46	3	Schilderwald	100% den Schilderwald [...] lichten
+	2	34	6	Anker	100% Anker lichten
+	3	23	9	Szenen	100% lichte [...] Szenen
+		0	55	46	statistisch unspezifisch



Archiv

Korpus

Suchanfrage

Wortformliste

Ergebnisse

Kookkurrenzanalyse

Gesamt-KWIC

Gesamt-Volltext

Export

Aktuelles Archiv TAGGED-T – Archiv morphosyntakt. annotie

Aktuelles Korpus hmp – Hamburger Morgenpost, Januar

Suchanfrage &licht

Exportkonfiguration

Ergebnisse: 55



Gesamt-KWIC



Gesamt-Volltext



Export



Kookkurrenzanalyse

Allgemeines

Dateiname: Ergebnisse_licht

Schriftgröße: 8

Bitte keine Umlaute, Sonderzeichen oder Leerzeichen im Dateinamen verwenden.

Dateiformat: ASCII RTFOrientierung: Querformat Hochformat Ergebnisübersicht Korpusansicht KWIC

vorher 0

nachher 0

Sätze

KWIC / Volltext

Sortierung: unsortiert

 Nur aktive Treffer

Korpusgrundlagen

- Öffentlich recherchierbar ist das „Wortprofil 2012“: dieses basiert auf Korpora im Umfang von etwa 1,8 Milliarden Textwörtern und enthält 11,9 Millionen verschiedene Kookkurrenzen für insgesamt 105 000 Lemmaformen.
- Intern für die lexikografische Arbeit verfügbar ist das „Wortprofil 2016“: dieses basiert auf Korpora im Umfang von etwa 2,7 Milliarden Textwörtern und enthält 21,2 Millionen verschiedene Kookkurrenzen für insgesamt 140 000 Lemmaformen.

Literaturangaben und URL

- Didakowski, Jörg und Alexander Geyken (2014): From DWDS corpora to a German word profile—methodological problems and solutions. In: OPAL – Online publizierte Arbeiten zur Linguistik 2/2014, S. 39–47.
- Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In: A. Abel & R. Zanin (Eds.), Korpora in Lehre und Forschung (pp. 115–137). Bozen, Italien: Bozen University Press.
- Meißner, Cordula (2008): Eine gebrauchorientierte Beschreibung des Sprachsystems mit Hilfe der Korpuslinguistik - das Beispiel der Synonyme ewig und unendlich. Deutsch als Fremdsprache 45: 8-13.
- Zu Wittgenstein (1953): Schulte, Joachim. Hrsg. (2001): Ludwig Wittgenstein: Philosophische Untersuchungen. Kritisch-genetische Edition. Frankfurt: Wissenschaft-liche Buchgesellschaft.