

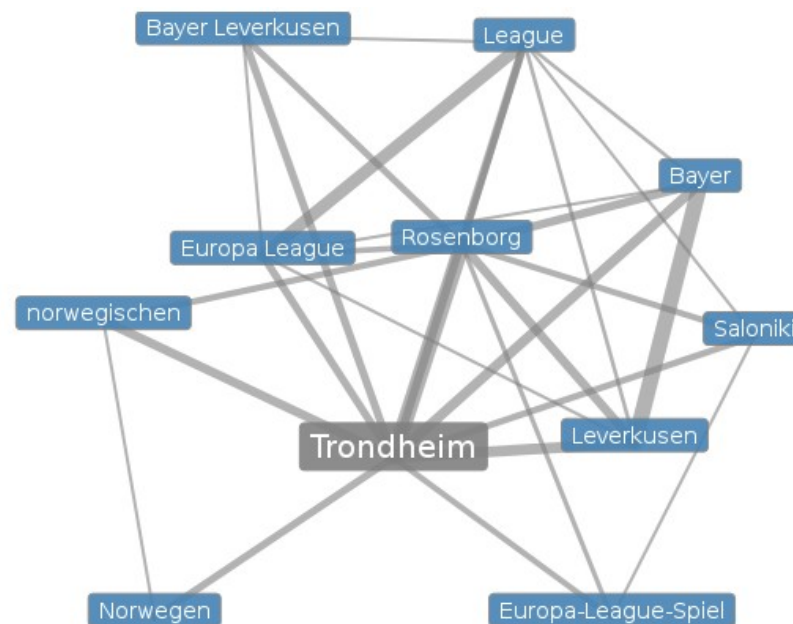
# Wortschatz / Leipzig Corpora Collection

## Resources & Access

- Project that offers
  - Corpus-based monolingual full form dictionaries of several languages
  - Web interface for general access
  - Web services
- Started as project „Deutscher Wortschatz“ more than 15 years ago by creating a corpus-based monolingual dictionary of the German language
- Since June 2006 Leipzig Corpora Collection (LCC) accessible at <http://corpora.uni-leipzig.de>



- Corpora in >300 languages
  - Text material
  - Word frequencies
  - Statistically significant word co-occurrences (based on left or right neighbours or whole sentences)
  - Semantic maps for visualizing strongest word co-occurrences
  - Subject areas for words (partially)
  - Morphological information (partially)
  - POS-tagged sentences (partially) etc.



- Generic web crawler
- RSS feeds
- (Distributed text crawling)
- Bootstrapping web corpora using search engines
- Text dumps

- Heritrix based web crawling
- Usage of URL lists provided by
  - AbyZNewslinks (directory of online news sources)
    - Around 100 languages
    - Ongoing process
  - List with generic URLs

- Download of text material from single domains:
  - Wikipedia dumps
    - 280 languages
  - Bibles (bible.is)
    - >800 languages
  - Watchtower (watchtower.org)
    - 438 languages
  - Project Gutenberg
    - 60 languages

- Periodical download of RSS feeds based on RSS feed lists
- Feed list:
  - Around 67,000 feed links for 137 languages
  - Steady extension
  - More?



- Example: 2016-09-19
  - 82 languages, 69,000 documents
  - eng: 26,000 docs (358K sentences)
  - deu: 6,000 docs (81K sentences)
  - nor (macro): 960 docs
- No feeds for aka, lug, etc.



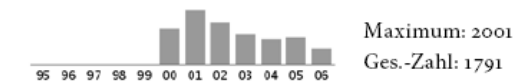
- Application 1: Identifying neologisms
  - Lemmas: Words and multiword units with zero or low frequency in the past and significantly higher frequency later.

## Babyklappe

GESUNDHEIT

(Auch: Baby-Klappe.)

Einrichtung, meist an Krankenhäusern, an der Neugeborene anonym abgegeben werden können.



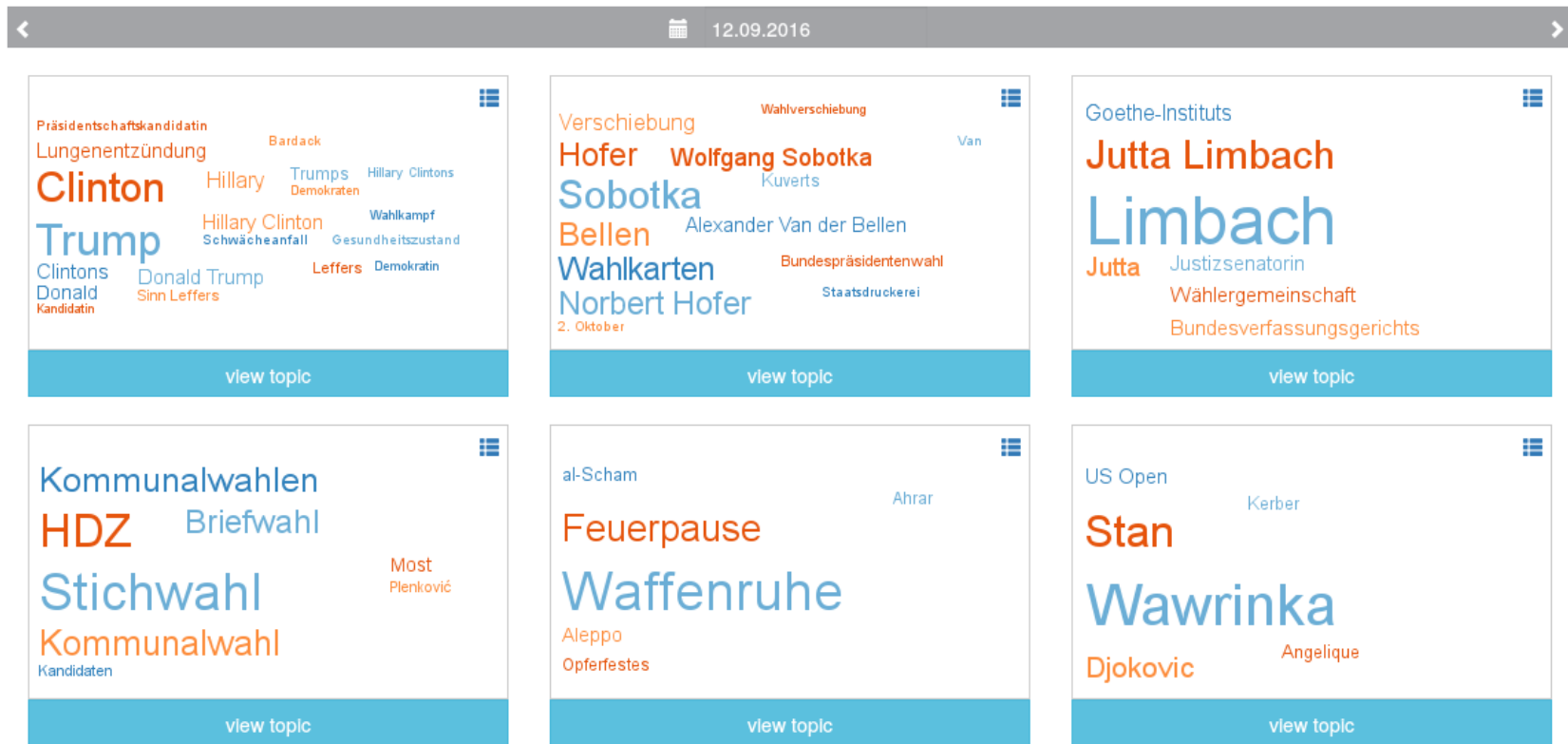
Babyklappen, Wärmebettchen, in die Säuglinge von außen gelegt werden können, gibt es in Deutschland seit 2000. Sie sollen das Aussetzen oder Töten von Säuglingen verhindern und der Mutter eine anonyme Abgabe des Kindes ermöglichen. Rechtlich ist die Situation jedoch umstritten, da sich bei Findelkindern aus der Babyklappe die Herkunft nicht feststellen lässt.

*Im Jahre 1709 hatte ein reicher holländischer Kaufmann eine »Babyklappe« am Waisenhaus eingerichtet. (Berliner Zeitung vom 07. 03. 2000) ■ »Für Frauen in extremen Notsituationen kann die Babyklappe eine große Entlastung sein.« (Berliner Zeitung vom 09. 03. 2000) ■ Dort wurde im März 2000 die bundesweit erste Babyklappe eröffnet, nachdem 1999 zwei Neugeborene in den Müll geworfen worden sind. (Berliner Zeitung vom 09. 02. 2001)*

- Application 2: Words of the day



The "words of the day" show which terms are particularly relevant today. These various newspapers and news services are evaluated daily. The "words of the day" are in the morning from about 7 o'clock available. The timeliness of a concept results from its frequency today, compared to its average frequency over a long period.

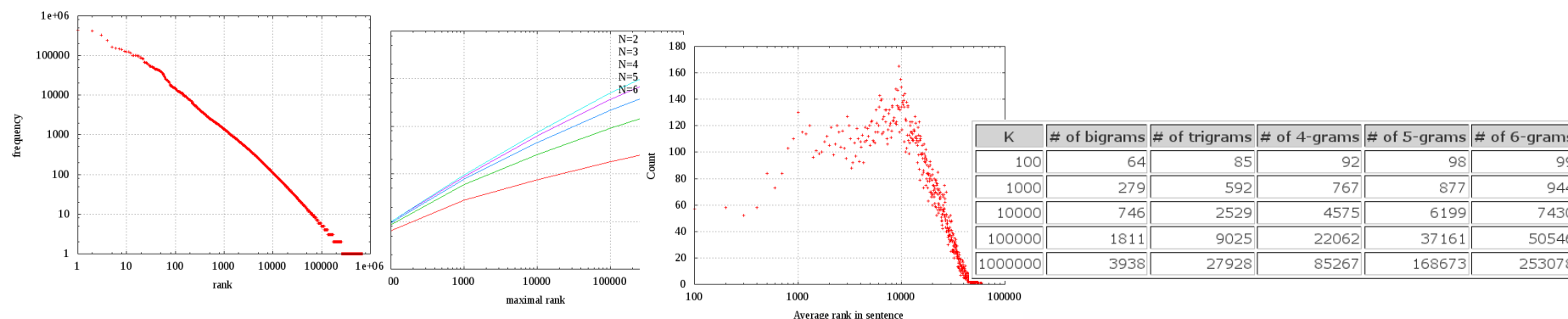


Card 1	Card 2	Card 3
<p>Präsidentenkandidatin Lungenentzündung <b>Clinton</b> <b>Trump</b> Hillary Trump Hillary Clintons Hillary Clinton Schwächeanfall Clinton Donald Donald Trump Sinn Leffers Bardack Hillary Trump Demokraten Wahlkampf Gesundheitszustand Leffers Demokratin</p> <p>view topic</p>	<p>Verschiebung Wahlverschiebung <b>Hofer</b> <b>Sobotka</b> <b>Bellen</b> <b>Wahlkarten</b> <b>Norbert Hofer</b> 2. Oktober Alexander Van der Bellen Bundespräsidentenwahl Staatsdruckerei Van Kuverts</p> <p>view topic</p>	<p>Goethe-Instituts <b>Jutta Limbach</b> <b>Limbach</b> Jutta Justizsenatorin Wahlgemeinschaft Bundesverfassungsgerichts</p> <p>view topic</p>
<p>Kommunalwahlen <b>HDZ</b> <b>Stichwahl</b> <b>Kommunalwahl</b> Kandidaten Briefwahl Most Plenković</p> <p>view topic</p>	<p>al-Scham <b>Feuerpause</b> <b>Waffenruhe</b> Aleppo Opferfestes Ahrar</p> <p>view topic</p>	<p>US Open <b>Stan</b> <b>Wawrinka</b> Djokovic Angelique Kerber</p> <p>view topic</p>

<http://wod.corpora.uni-leipzig.de>

- Processing steps:
  - Stripping
  - Language Separation (about 500 languages)
  - Sentence Segmentation
  - Cleaning
  - **Sentence Scrambling** (copyright reasons)
  - Tokenization
  - Database Creation
  - Statistical Evaluation

- Enrichment of corpora with statistical annotations
  - Word frequencies
  - Co-occurrence frequencies and significance (based on left or right neighbours or whole sentences)
- Creation of further corpora statistics
  - 230 statistical properties based on letter, word, sentence, sources level etc.



- Still open work in fields like automatic and manual quality assurance
  - Removal of near-duplicate sentences
- More languages
- More linguistic annotations
- Adjustment of methods to focus on rare languages
- Exploitation of new resources

<i>Language Code</i>	<i>Number of sentences</i>	<i>Rank</i>
eng	1100M	1
deu	1000M	2
rus	456M	3
nob	27M	31
nno	2M	54
<i>(nor)</i>	<i>1M</i>	<i>67</i>



- Data available via:
  - Web portal  
(<http://corpora.uni-leipzig.de>)
  - CQL-Interface  
(<http://cql.corpora.uni-leipzig.de>)
  - Download  
(<http://corpora.uni-leipzig.de/download.html>)
  - Webservices  
(<http://api.corpora.uni-leipzig.de>)

- Corpus query engine NSE

NoSketchEngine

**Concordance**

**Word list**

**Corpus info**

**My jobs**

?

---

**Home**

**User guide**

---

**Save**

as subcorpus

**View options**

KWIC

Sentence

**Sort**

Left

Right

Node

References

Shuffle

**Sample**

Filter

Overlaps

Frequency

Query **basinga** 23 (124.86 per million)

Page  of 2  [Next](#) | [Last](#)

621	mu ntubwe olwo nga baseka ! Abambutu be <b>basinga</b> okubeera abampi . Abawala abawanvu akutunuulira
1891	nga mu bonna nze mumpi nzekka alimu . Bano <b>basinga</b> kubaako ffuuti nnya zokka ! Bazadde bange
838	muganda . Abatambala n'abo abasibuka e Gomba , <b>basinga</b> kumanyibwa mu kumeggana era ne kyampiyoni
961	Buganda 18 n'emizannyo abantu baayo gye <b>basinga</b> okwettanira . Bano ku mirembe egyayita
4690	ekibagatta . Kino ky'ekifo abaagalana kye <b>basinga</b> okukozesa mu myaka egisooka mu mukwano
3310	ebinyirira nga bino mubyerinde . Emirimu gye <b>basinga</b> okugaba gya kuvuga mmotoka olw'okutya enguudo
9708	Olwokuna ekiro ne Mmande ze zimu ku nnaku ze <b>basinga</b> okukola olw'abawala abangi ababa bajja
247	tebayamba nnyo kubanga abamenyi b'amateeka be <b>basinga</b> poliisi obungi . Abasula mu Nakawa mulimu
1844	okuyimirira e Kyabakuza . Bannakyabakuza <b>basinga</b> kwettanira ntambula ya bodaboda era emmotoka
3568	aba bodaboda . Eno y' emu ku nsonga lwaki <b>basinga</b> kubba pikipiki mpya eziba n' ebyuma ebiramu
7981	mabanja na nsonga za maka ate bamanelenda <b>basinga</b> kufera bakyala . "oLUUSI nze mbatiza abasibe
415	ofuuse mulwadde . Abaana b'amasomero be <b>basinga</b> okukozesa ebiragalalagala . Abudaabuda
429	wonna w'oyagala , ojja kufunamu . Abaayo <b>basinga</b> kulima nva ndiirwa omuli nnakati , obutungulu
8144	ebintu bingi , omuli okumanya ebintu bye <b>basinga</b> okwettanira , enneeeyisa yaabwe , essaawa
8144	okwettanira , enneeeyisa yaabwe , essaawa ze <b>basinga</b> okujjirako ewuwo n'ensonga , bizibu ki
926	Abayimbi ne bannakatamba ebuwulu byabwe <b>basinga</b> kubissaayo ku nkomerero ya mwezi kuba olwo
781	gwe batayinza kumala gasuula . Abasajja be <b>basinga</b> okutya n'okwewala okugenda mu bifo gye
3341	bumulumulu kubanga abantu bangi . Emmere <b>basinga</b> kugiggya Budondo , Budiope , Namayingo
9492	Robin van Persie ne Wayne Rooney aba ManU <b>basinga</b> Sir Alex ferguson omusaala omusava . Singa
973	b'emidudu n'abasazi b'ensawo be babadde <b>basinga</b> okubabba . Bbulooka wa takisi ng'akozesa

Page  of 2  [Next](#) | [Last](#)

<http://cql.corpora.uni-leipzig.de>

lug\_newscrawl\_2013

16



Corpus: German (deu\_newscrawl\_2011)

Simple query:

[Query types](#) [Context](#) [Text types](#) [?](#)

Query type  simple  lemma  phrase  word  character  CQL

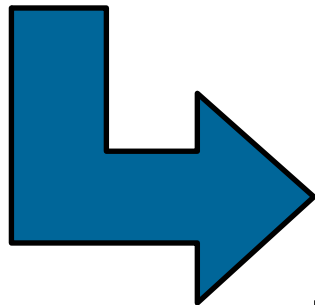
Lemma:

Phrase:

Word form:   match case

Character:

CQL:  Default attribute:




Query **d.\*, Hund, VERB** 1,778 (3.80 per million)

Page 1 of 89  [Next](#) | [Last](#)

13006090	Ihrem Bedarf liegen . Die Spaziergänge mit	dem /DET Hund /NOUN können /VERB	Sie dann als Goodie obendrauf betrachten
24392201	nun mal auch ihre Vorgaben und müssen mit	den /DET Hunden /NOUN heulen /VERB	, eben wie überall . Sachlich bleiben kann
24862536	Kinder . Mir tut das so leid , was du mit	dem /DET Hund /NOUN erlebt /VERB	hast . Mir war es wieder zu warm gewesen
20777499	Nicky , dass ist sehr schlimm , was dir	der /DET Hund /NOUN angetan /VERB	hatte und deine Mutter dabei zusah . Nicky
19920375	die feine/prominente Gesellschaft . Nur	dumme /ADJ Hunde /NOUN beissen /VERB	die Hand , die sie füttert . Nur ein Nein
13149039	sind immer Mittel vorhanden . Dort liegt	der /DET Hund /NOUN begraben /VERB	und denn kann der BR auch nicht mit mehr
25954215	aber bitte sagen Sie mir nicht , wie ich	den /DET Hund /NOUN erziehen /VERB	soll . Danke fürs Erzählen , ich sehe Dich
14886679	heraus gesund wird und im w.S.d.W . nicht vor	die /DET Hunde /NOUN geht /VERB	. Glaube dem Leben , es lehrt mehr als
18618812	würde versuchen , es ihr als das Beste für	den /DET Hund /NOUN beizubringen /VERB	und es 1 Tick langsamer angehen , sonst
25456660	ja das Prbl . Ja eben , gerade deshalb ,	der /DET Hund /NOUN stammt /VERB	vom Wolf ab und man hat ihm seine Freiheit
1019904	Wo , na ja , ein bisschen ausschalten ,	der /DET Hund /NOUN kommt /VERB	schon klar , baba . So manche Frau würde
13041491	Krank . Warum auch das bißchen Freizeit	dem /DET Hund /NOUN widmen /VERB	! Warum laufen so viele Menschen in irgendwelche
21883369	die sosnt keinen Lebensinhalt haben . Wenn	der /DET Hund /NOUN wählen /VERB	muss , kommt er zu mir . Wenn die Anschauung
15628783	an Sie zurück : - ) Gute Nacht . Gehe mit	dem /DET Hund /NOUN spazieren /VERB	, der sagt nix dazu . Geht ein Kind zur
17360726	Lustig im Hundekorb , sonst ist es umgekehrt	der /DET Hund /NOUN schleicht /VERB	sich in's Bett . : - ) ) ) Mache ich eh
1856152	Welt 10 mal so viele an echtem Mangel vor	die /DET Hunde /NOUN gehen /VERB	? Und wozu brauchen wir Goldstone , wenn
4965931	nicht dazu hinreißen lassen . Muss mich um	den /DET Hund /NOUN kümmern /VERB	. Nach dem . was du schreibst . und wie

- Several services that are steadily extended
- OpenAPI documentation (allows testing services directly in browser):

 Leipzig Corpora Collection
default (/v2/api-docs) ▾

**REST API of the Leipzig Corpora Collection / Projekt Deutscher Wortschatz**

This is the REST API of the [Leipzig Corpora Collection \(LCC\)](#) at the [Natural Language Processing Group, University of Leipzig](#).

With these Web services you have direct access to the data of the Leipzig Corpora Collection (LCC) by using a software of your choice. This Web page also allows to test the provided services directly.

Please be aware that these services have still **BETA** status. There may be downtimes and changes to the interfaces for the time being.

Created by the Natural Language Processing Group, University of Leipzig  
 See more at <http://wortschatz.uni-leipzig.de>  
[Contact the developer](#)  
[Terms of Usage](#)

**available-corpora-service : Available Corpora Service** Show/Hide | List Operations | Expand Operations

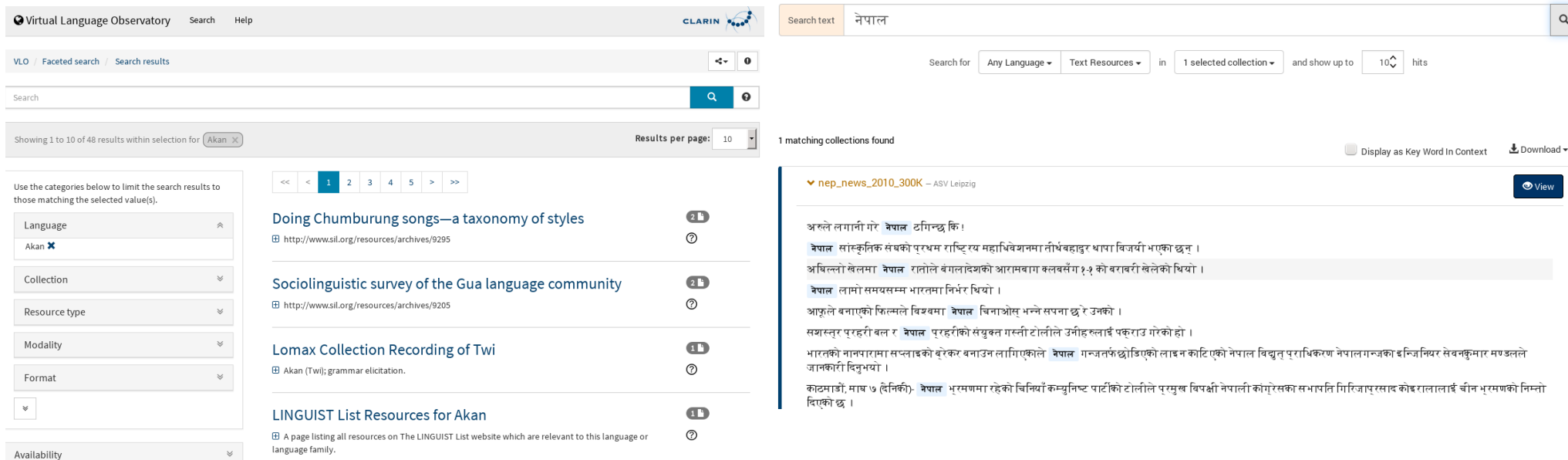
GET	/corpora/availableCorpora	Get available corpora
-----	---------------------------	-----------------------

**cooccurrences-service : Cooccurrences Service** Show/Hide | List Operations | Expand Operations

GET	/cooccurrences/{corpusName}/cooccurrences/{word}	Get sentence cooccurrences
GET	/cooccurrences/{corpusName}/cooccurrencesgraph/{word}	Get cooccurrences graph
GET	/cooccurrences/{corpusName}/leftcooccurrences/{word}	Get left neighbor cooccurrences
GET	/cooccurrences/{corpusName}/rightcooccurrences/{word}	Get right neighbor cooccurrences

<http://api.corpora.uni-leipzig.de>

- Participation in CLARIN-D
- Research infrastructure for language resources as Service-oriented architecture (SOA)
- European language bias



Virtual Language Observatory Search Help

CLARIN

Search text: नेपाल

Search for: Any Language Text Resources in 1 selected collection and show up to 10 hits

Showing 1 to 10 of 48 results within selection for Akan

Results per page: 10

1 matching collections found

Display as Key Word In Context Download

▼ nep\_news\_2010\_300K – ASV Leipzig View

अरुले लगानी गरे **नेपाल** टगिन्द्र कि !  
**नेपाल** सांस्कृतिक संघको प्रथम राष्ट्रिय महाभियेनमा तीर्थबहादुर थापा विजयी भएका छन् ।  
 अधिल्लो खेलमा **नेपाल** रातोले बंगलादेशको आरामबाग क्लबसँग १-१ को बराबरी खेलेको थियो ।  
**नेपाल** लामो समयसम्म भारतमा निर्भर थियो ।  
 आफूले बनाएको फिल्मले विश्वमा **नेपाल** चिनाओस् भन्ने सपना छरे उनको ।  
 सशस्त्र प्रहरी बल र **नेपाल** प्रहरीको संयुक्त गस्ती टोलीले उनीहरूलाई पक्राउ गरेको हो ।  
 भारतको नानपारामा सप्लाईको ब्रेकर बनाउन लागिआएको **नेपाल** गन्तव्य छ।  
 कोटमाडौं, माघ ७ (दिनिकी) - **नेपाल** भ्रमणमा रहेको चिनियाँ कम्पनिष्ट पार्टीको टोलीले प्रमुख विपक्षी नेपाली कांग्रेसका सभापति गिरिजाप्रसाद कोइरालालाई चीन भ्रमणको निम्तो दिएको छ ।

<https://vlo.clarin.eu/?fq=collection:Leipzig+Corpora+Collection>