



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



Consiglio Nazionale
delle Ricerche

Interrogazione e Utilizzo di Corpora Annotati

L'Esempio di ParlaMint



Formazione InfoText - Siena
Francesca Frontini, Giulia Pedonese
14 aprile 2025

Indice

- Principi FAIR e Open Science: **Quiz e ripasso**
- **Esercizio:** valuta il grado di aderenza ai principi FAIR di una risorsa
- Esempi di risorse FAIR: **lo standard di ParlaMint**
- **Ricerca** testuale su ParlaMint

Fonti

Questa presentazione è il risultato dell'adattamento delle seguenti fonti:

- van der Lek, Iulianna; Fišer, Darja. (2023). *Introduction to Language Data: Standards and Repositories*. In [UPSKILLS](https://upskillsproject.eu/project/standards_repositories/) Learning Content. https://upskillsproject.eu/project/standards_repositories/. [CC BY 4.0](#).
- van der Lek, I., Fišer, D., Samardzic, T., Simonovic, M., Assimakopoulos, S., Bernardini, S., Milicevic Petrovic, M., & Puskas, G. (2023). *Integrating research infrastructures into teaching: Recommendations and best practices* (Versione 2) <https://doi.org/10.5281/zenodo.8114407>. [CC BY 4.0](#).
- CLARIN ERIC Official Website: <https://www.clarin.eu/> . [CC BY 2.0](#)
- Frontini, Francesca; Pedonese, Giulia. *CLARIN per la Gestione FAIR dei Dati Linguistici*, Laboratorio Sperimentale, Università di Bologna, 20 giugno 2024.

FAIR corpora

- Chiare licenze di riuso
- Uso di formati di annotazione standard (TEI Parlamint)
- Uso di Tagset standard (UD)
- Disaccoppiamento tra dati e interfacce di interrogazione:
 - Deposito dei dati e possibilità di download dei sorgenti
 - Compatibilità con linguaggi di interrogazione standard – CQL
 - Interoperabilità con diverse interfacce:
 - SketchEngine & NoSketchEngine
 - Kontext
 - TEITok
 - ...
- Documentazione e materiali didattici



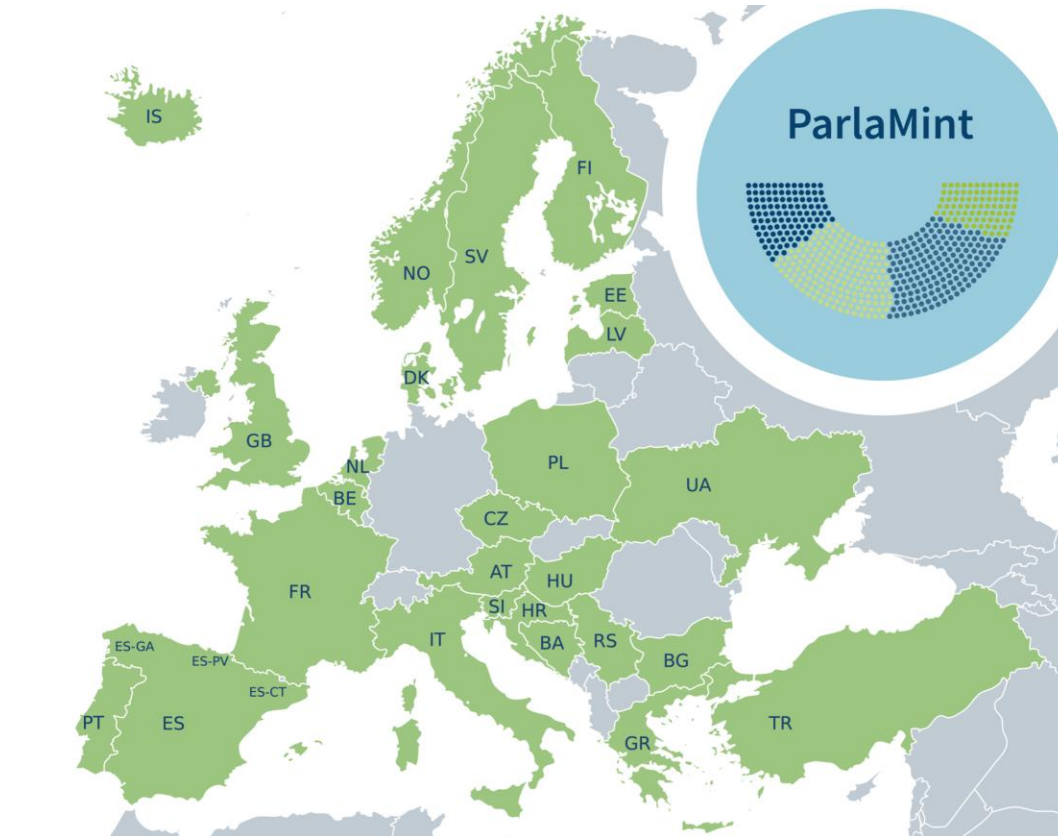
Un esempio di interoperabilità: ParlaMint

ParlaMint

ParlaMint, un progetto bandiera di CLARIN, ha portato alla creazione di corpora paralleli di dibattiti parlamentari di 29 paesi europei e regioni autonome, che coprono almeno il periodo dal 2015 al 2022 (alcuni corpora sono stati estesi fino al 2023) e contengono oltre 1 miliardo di parole.

- Corpora uniformemente codificati (TEI)
- Con ricchi metadati sui loro 24 mila parlanti
- Sono annotati linguisticamente fino al livello della sintassi UD e delle entità denominate
- Traduzione automatica in inglese per l'interrogazione parallela e l'annotazione semantica
- La versione più recente del corpus è la **4.1**, pubblicata nel 2024.
- Nel gennaio 2025 è stato avviato il progetto **ParlaCAP**, che mira ad arricchire oltre **8 milioni di discorsi** nei corpora ParlaMint con informazioni su **argomenti trattati** e **sentimenti espressi**, utilizzando modelli multilingue

ParlaMint



Annotazione linguistica

Attributes **ana** and **msd** both encode the morphosyntactic characteristics of the word in focus (in blue). The **ana** attribute contains Slovene-specific tags according to MULTEXT-East Specifications, while the **msd** attribute contains the Universal Dependency tagset which greatly simplifies cross-lingual comparison.

The attribute values Q and PART both stand for *particle*. The difference between the two tags is in the tagset used.

The **word form** also known as a token from the running text in the corpus.

The attribute **lemma** indicates the basic word form of the token, that is of the word in focus (in blue).

The corpus also encodes syntactic parses and named entities but since they are not used in this tutorial they were omitted from this illustrative example.

```
<s>
<w ana="mte:Q" msd="UposTag=PART" lemma="zlasti">Zlasti</w>
<w ana="mte:Sg" msd="UposTag=ADP | Case=Gen" lemma="glede">glede</w>
<w ana="mte:Ncmmsg" msd="UposTag=NOUN | Case=Gen | Gender=Masc | Number=Sing"
lemma="nadzor">nadzora</w>
<w ana="mte:Va-r3s-n"
msd="UposTag=AUX | Mood=Ind | Number=Sing | Person=3 | Polarity=Pos | Tense=Pres | VerbForm=Fin"
lemma="biti">je</w>
<w ana="mte:Pd-fsn" msd="UposTag=DET | Case=Nom | Gender=Fem | Number=Sing | PronType=Dem"
lemma="ta">ta</w>
<w ana="mte:Ncfnsn" msd="UposTag=NOUN | Case=Nom | Gender=Fem | Number=Sing"
lemma="stvar">stvar</w>
<w ana="mte:Rgp" msd="UposTag=ADV | Degree=Pos" lemma="zelo">zelo</w>
<w ana="mte:Agpfsn" msd="UposTag=ADJ | Case=Nom | Degree=Pos | Gender=Fem | Number=Sing"
lemma="kočljiv">kočljiva</w>
<pc ana="mte:Z" msd="UposTag=PUNCT">.</pc>
</s>
```

One pair of the opening and closing structural tags which, in our case, indicate punctuation (**pc**), word (**w**) or sentence (**s**).

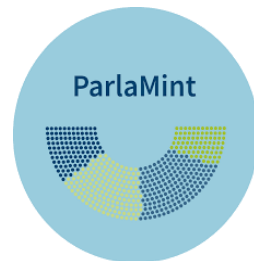


Metadati specifici

Metadati: informazioni descrittive sulle fonti dei dati e sul contenuto (manuale) provenienti da fonti parlamentari ufficiali e da fonti esterne. Facilitano ad es. analisi contrastive, diacroniche (creazione di sottocorpora) ecc. Sono relativi ad aspetti come:

- **Aspetti del parlamento:** unicamerale o bicamerale; termini, sessioni e riunioni con indicazione temporale
- **Relatori:** nomi, genere, status di deputato, ministri, affiliazione al partito
- **Partiti:** coalizione/opposizione, orientamento politico destra-sinistra
- **Discorsi:** contrassegnati dall'oratore e dal suo ruolo: ad esempio, presidente, oratore regolare, ospite o contenenti commenti marcati del trascrittore, come lacune nella trascrizione, interruzioni, applausi, ecc.

<https://github.com/clarin-eric/ParlaMint/blob/main/TEI/ParlaMint.odd.xml>





DASHBOARD

ParlaMint-AT 4.0 (Austrian parliament)



PARLAMINT-AT 4.0 (AUSTRIAN PARLIAMENT)

CORPUS INFO

MANAGE CORPUS

Concordance

Examples of use in context

Parallel Concordance

Translation search

Wordlist

Frequency list

Keywords

Terminology extraction

Trends

Diachronic analysis, neologisms

Text type analysis

Statistics of the whole corpus



You are using NoSketch Engine. These tools are only available in Sketch Engine

TEXT TYPE ANALYSIS

ParlaMint-AT 4.0 (Austrian parliament)



Structures and text types

speech - speaker_party

speech - speaker_party_name

speech - speaker_role

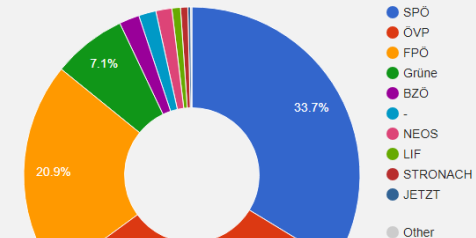
speech - subcorpus

speech - term

name - type

Items: 11, Total frequency: 231,758

speech - speaker_party

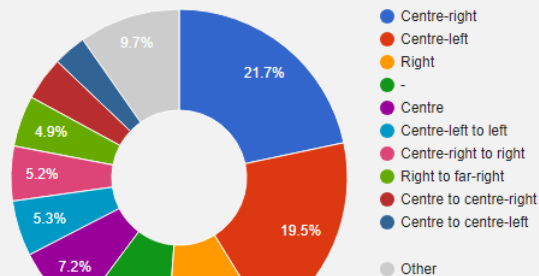


ParlaMint-XX 4.0 (European parliaments)



Items: 41, Total frequency: 7,851,329

speech - party_orientation



speech - speaker_party

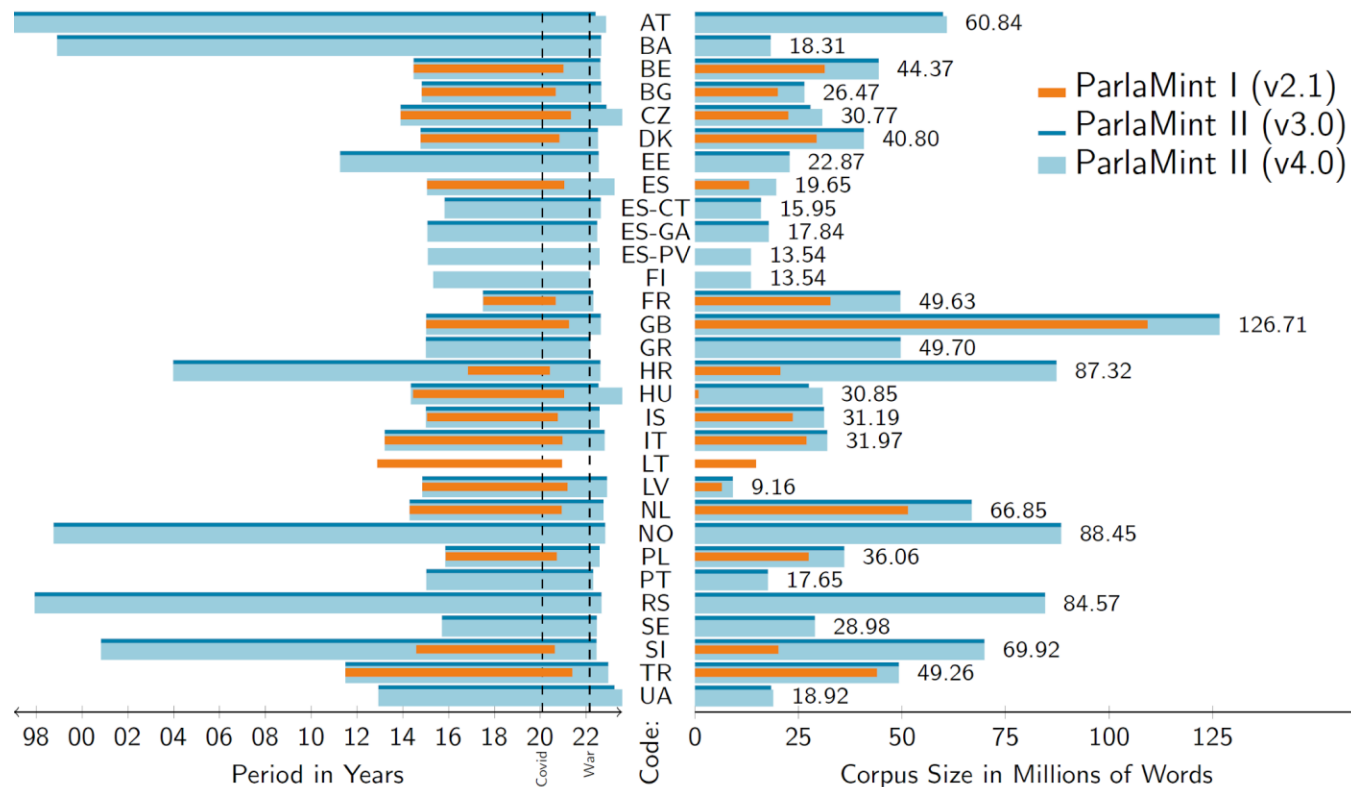
speech - speaker_party

speech - speaker_party

speech - speaker_party

Attribute value		Structure frequency ?		Attribute value		Structure frequency ?	
1	SPÖ	78,091	...	11	GRÜNE	279	...

COPERTURA: periodo e volume



Come usare ParlaMint

Vari tipi di concordancers:

1. **Sketchengine**
2. NoSketchEngine (CLARIN.si)
3. TeiTok
4. Kontext
5. ...
6. **Scaricando i dati e caricandoli su altri strumenti, o scrivendo il proprio codice (modulo di domani!)**

Dai repositories:

1. Tomaž Erjavec et al. (2023) Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. <http://hdl.handle.net/11356/1859>
2. Tomaž Erjavec et al. (2023) Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. <http://hdl.handle.net/11356/1860>
3. Taja Kuzman et al. (2023) Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 4.0. <http://hdl.handle.net/11356/1864>

Materiali didattici aggiuntivi

Attraverso il **Learning Hub** sono disponibili due tutorial su ParlaMint, di cui uno su topic modeling, che i corpus query systems tradizionali non permettono di fare, attraverso l'installazione di un CQS desktop <https://www.clarin.eu/content/learning-hub>

Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research

Authors: Darja Fiser and Kristina Pahor de Maiti

Faculty of Arts, University of Ljubljana, Slovenia

Keywords: *parliamentary proceedings, parliamentary corpora, language and gender, digital humanities*

What's on the agenda? Topic modelling parliamentary debates before and during the COVID-19 pandemic

Authors: Ajda Pretnar Žagar, Kristina Pahor de Maiti, Darja Fišer

Institute of Contemporary History, Ljubljana, Slovenia

Keywords: *topic modelling, LDA, parliamentary debates, text mining*

Parte pratica

Sketch Engine

Software per la gestione dei corpora pensato per linguisti

- **Word Sketch** per fare ricerche sulle collocazioni e presentarle graficamente
- **Word Sketch Difference** per confrontare due parole e il loro utilizzo
- **Thesaurus** per trovare sinonimi del lemma ricercato
- **Concordance search** per cercare il contesto di una parola o una frase in una sola lingua. Qui è disponibile la ricerca tramite **Corpus Query Language**
- **Parallel Concordance** per fare ricerche su corpora di 2 o più lingue diverse e confrontare i risultati
- **Keyword (Term Extraction)** per estrarre terminologia e unità di una o più parole che sono tipiche di un corpus/documento/testo

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fske_parlamint21_it

Esempio 1: Farsi gli affari propri

The screenshot displays the SketchEngine Concordance interface. At the top, the title "CONCORDANCE" is shown next to a search bar containing "ParlaMint-IT 4.1 (Italian parliament)". Below this, a status bar indicates the query "CQL 'fare'[{0,3} 'affare' 'proprio'] • 6" with a frequency of "0.16 per million tokens • 0.000016%". The main section is titled "CHANGE CRITERIA" and has three tabs: "BASIC", "ADVANCED", and "ABOUT". The "BASIC" tab is active. On the left, a "Query type" dropdown menu is open, showing options: "simple", "lemma", "phrase", "word", "character", and "CQL" (which is highlighted). The main query input field contains the CQL query: "fare"[] {0,3} "affare" "proprio". Below the input field is an "Insert" toolbar with buttons for brackets, curly braces, less-than/greater-than, quotes, ampersand, backslash, pipe, tilde, and hash. To the right of the toolbar is a "CQL BUILDER" button. Below the toolbar, a "Default attribute?" dropdown menu is set to "lemma".

SketchEngine <https://www.clarin.si/ske/#corpus?tab=basic&cat=all&sketches=0&lang=&lang2=&query=&showOld=0>

Esempio 2: Farsi gli affari propri

CONCORDANCE Italian Web 2020 (itTenTen20)

CQL [word="si"] "fare" [][0,3] "affare" "proprio" • 231
0.02 per million tokens • 0.0000016%

CHANGE CRITERIA

BASIC ADVANCED ABOUT

Query type ②

- simple
- lemma
- phrase
- word
- character
- CQL

CQL

[word="si"] "fare" [][0,3] "affare" "proprio"

Insert [] { } < > " " & \ | ~ # TAGS CQL BUILDER

Default attribute ?

lemma

Subcorpus ②

none (the whole corpus) +

Macro ?

none

Filter context ② ▾

Text types ? ▾

GO

Get more space +

KWIC +

Q1: Complex cor...
an introduction to
corpus language
SKETCH ENGINE
www.sketchengine.eu

CQL manual

SketchEngine

<https://www.clarin.si/ske/#corpus?tab=basic&cat=all&sketches=0&lang=&lang2=&query=&showOld=0>

Esempio 3: piuttosto che

- *Piuttosto che: dalla preferenza all'esemplificazione di alternative / Caterina Mauri, Anna Giacalone Ramat. - In: CUADERNOS DE FILOLOGÍA ITALIANA. - ISSN 1133-9527. - STAMPA. - 20:(2015), pp. 49-72.*
- https://dx.doi.org/10.5209/rev_CFIT.2015.v22.50951
- *Uso di Piuttosto che con valore disgiuntivo, Ornella Castellani Pollidori, Accademia della Crusca* <https://accademiadellacrusca.it/it/consulenza/uso-di-piuttosto-che-con-valore-disgiuntivo/11>

SketchEngine

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fske_parlamint21_it

Piuttosto che...

🔍 CHANGE CRITERIA

BASIC

ADVANCED

ABOUT

Query type ?

simple

lemma

phrase

word

character

CQL

CQL

[tag="V.*"][]{0,3} "piuttosto" "che" [tag="V.*"][]{0,3}|

Insert

[]

{ }

< >

" "

&

\

|

~

#

TAGS

CQL BUILDER []

Default attribute ?

lemma ▼

Subcorpus ?

none (the whole corp... ▼



Macro ?

none



Piuttosto che...

🔍 CHANGE CRITERIA

BASIC

ADVANCED

ABOUT

Query type ?

simple

lemma

phrase

word

character

CQL

CQL

[tag="V.*"] [tag="D.*"] [tag="N.*"] "piuttosto" "che" [tag="D.*"] [tag="N.*"]

Insert

[]

{ }

< >

" "

&

\

|

~

#

TAGS

CQL BUILDER []

Default attribute ?

lemma ▼

Subcorpus ?

none (the whole corp... ▼



Macro ?

none



Piuttosto che...

CHANGE CRITERIA

BASIC

ADVANCED

ABOUT

Query type ?

simple

lemma

phrase

word

character

CQL

CQL

`[]{0,3} "piuttosto" "che" []{0,3} "piuttosto" "che" []{0,3}`

Insert

[]

{ }

< >

" "

&

\

|

~

#

TAGS

CQL BUILDER []

Default attribute ?

word

Subcorpus ?

none (the whole corp... ▾



Macro ?

none



Creazione di sottocorpora

Create subcorpus

Subcorpus name *

- ☒ Subcorpus from text types
☐ Subcorpus from concordance

expand all collapse all

speech.subcorpus ▾

speech.from ▾

speech.to ▾

speech.sitting ▾

speech.speaker_type ▾

speech.speaker_role ▾

speech.speaker_party_name ▾

speech.party_status ▾

speech.speaker_name ▾

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fske_parlamint21_it

Keyword extraction

Term		
1	torno a ripetere	...
2	centro antiviolenza	...
3	violenza sulle donne	...
4	violenza contro le donne	...
5	casa rifugio	...
6	facoltà di illustrare	...
7	donna uccisa	...
8	senatore nannicini	...
9	senatore patriarca	...
10	visita al senato	...
11	benvenuto al senato	...
12	benvenuto agli allievi	...

Term		
18	inchiesta sul femminicidio	...
19	forma di violenza	...
20	contrasto alla violenza	...
21	donna vittime	...
22	votazione con procedimento elettronico	...
23	votazione con procedimento	...
24	torno a ripeterlo	...
25	benvenuto in senato	...
26	senatore laforgia	...
27	scuola statale	...
28	lavoro di cura	...
29	vittima di violenza	...

Altri portali per la ricerca sui corpora in CLARIN

kon|text

Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: online2_now | Query: piuttosto, che (3 hits) ~ Details

Hits: 3 | i.p.m.: 0.01 (related to the whole corpus) | ARF: 1.54 | Result is sorted 1 / 1

Line selection: simple ▾

<input type="checkbox"/> voxeurop.eu	avere questo tipo di pressione da parte dei nostri lettori	piuttosto che	dipendere da una dozzina di aziende pubblicitarie ". Per
<input type="checkbox"/> voxeurop.eu	burnout . I terapeuti consigliano di controllare periodicamente le notizie	piuttosto che	gli aggiornamenti costanti , mentre alcune Ong stanno riconoscendo il
<input type="checkbox"/> voxeurop.eu	è) , anche se è stata fatta per paura	piuttosto che	per sincero rammarico (non lo sappiamo) - il

1 / 1

<https://contentsearch.clarin.eu/>

Altri portali per la ricerca sui corpora in CLARIN

Update existing operation "query"

Execution options:

- ☒ Perform automatically also subsequent operations
- ☐ Skip subsequent operations

Advanced query ☒ | Insert within | Keyboard

....[word="piuttosto" | lemma="piuttosto"] [word="che" | lemma="che"]

Syntax error - unexpected character " " at position 1.

TIP In case of compatibility problems with advanced CQL editor and your browser, you can switch back to basic CQL editor in "View" → "General view options" (next tip)

☐ Specify parameters

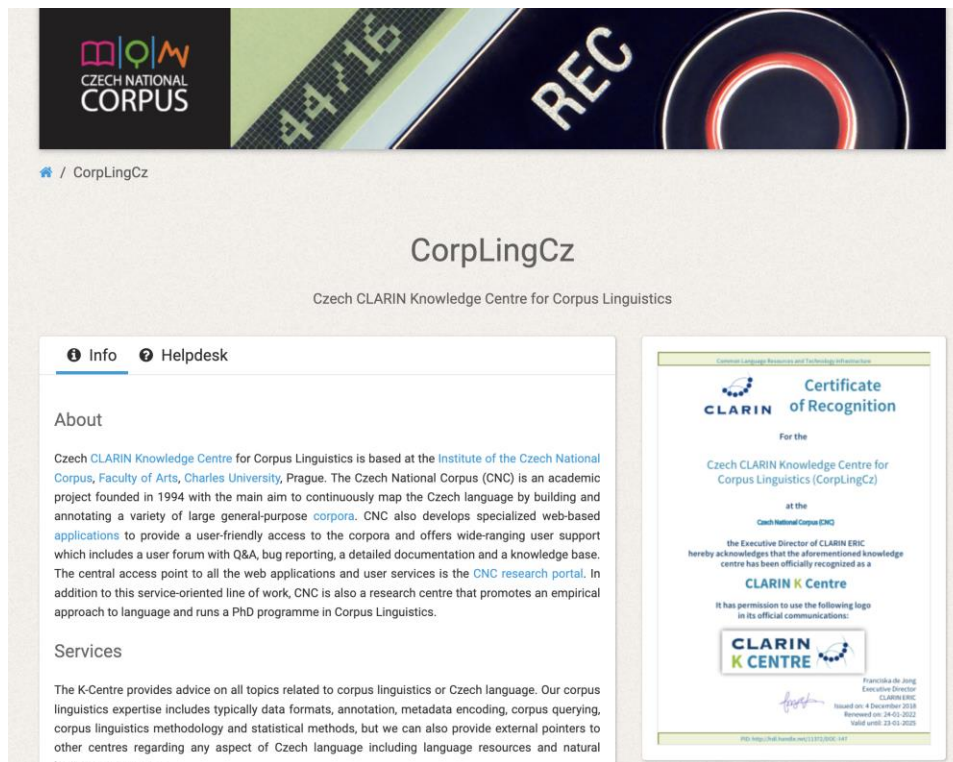
Default attribute: word

☐ Specify context

Proceed

<https://www.korpus.cz/kontext/query?corpname=syn2020>

Altri portali per la ricerca sui corpora in CLARIN



The image shows a screenshot of the CorpLingCz website and a CLARIN Certificate of Recognition. The website header features the 'CZECH NATIONAL CORPUS' logo and a banner with 'REC' and a circular graphic. Below the header, the site is identified as 'CorpLingCz' and 'Czech CLARIN Knowledge Centre for Corpus Linguistics'. The main content area has two tabs: 'Info' and 'Helpdesk'. The 'Info' tab is active, showing an 'About' section with text about the center's foundation in 1994 at Charles University, its mission to build and annotate the Czech National Corpus (CNC), and its role as a research center. A 'Services' section is also visible. To the right of the website content is a 'Certificate of Recognition' from CLARIN. The certificate is issued to the 'Czech CLARIN Knowledge Centre for Corpus Linguistics (CorpLingCz)' at the 'Czech National Corpus CNC'. It acknowledges the center's official recognition as a 'CLARIN K Centre' and grants permission to use the CLARIN K Centre logo in official communications. The certificate is signed by Francisco de Jong, Executive Director of CLARIN ERIC, and dated 4 December 2018.

CZECH NATIONAL CORPUS

CorpLingCz

Czech CLARIN Knowledge Centre for Corpus Linguistics

Info Helpdesk

About

Czech CLARIN Knowledge Centre for Corpus Linguistics is based at the [Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague](#). The Czech National Corpus (CNC) is an academic project founded in 1994 with the main aim to continuously map the Czech language by building and annotating a variety of large general-purpose [corpora](#). CNC also develops specialized web-based [applications](#) to provide a user-friendly access to the corpora and offers wide-ranging user support which includes a user forum with Q&A, bug reporting, a detailed documentation and a knowledge base. The central access point to all the web applications and user services is the [CNC research portal](#). In addition to this service-oriented line of work, CNC is also a research centre that promotes an empirical approach to language and runs a PhD programme in Corpus Linguistics.

Services

The K-Centre provides advice on all topics related to corpus linguistics or Czech language. Our corpus linguistics expertise includes typically data formats, annotation, metadata encoding, corpus querying, corpus linguistics methodology and statistical methods, but we can also provide external pointers to other centres regarding any aspect of Czech language including language resources and natural language processing.

Certificate of Recognition

For the

Czech CLARIN Knowledge Centre for Corpus Linguistics (CorpLingCz)

at the

Czech National Corpus CNC

the Executive Director of CLARIN ERIC hereby acknowledges that the aforementioned knowledge centre has been officially recognised as a

CLARIN K Centre

It has permission to use the following logo in its official communications:

CLARIN K CENTRE

Francisco de Jong
Executive Director
CLARIN ERIC
Issued on: 4 December 2018
Renewed on: 24-01-2022
Valid until: 23-01-2025

PDF: http://clarin.eu/clarin-eric/11712/ERIC_147

<https://korpus.cz/clarin>

Eventi di formazione H2IOSC

<https://www.h2iosc.cnr.it/news/>

Aggiornamento su iniziative di training di tutte le Infrastrutture coinvolte nel progetto



Iniziative della comunità CLARIN

CLARIN Newsflash

aggiornamento mensile sulle attività dei consorzi nazionali

CLARIN Café

webinar informali e interattivi per discutere su temi di interesse



Grazie per l'attenzione!

Contatti:

[https://forum.clarin.eu/
formazione@clarin-it.it](https://forum.clarin.eu/formazione@clarin-it.it)

