Data for AI development at CLARIN.SI

What the repository offers for pre-training, speech, post-training and evaluation

Nikola Ljubešić · Jožef Stefan Institute · CLARIN.SI
November 2026

What you will take away

Where the data is

1,038 resources in one certified repository, 790 of them openly licensed for training, including commercial use.

What it is good for

About 53B words of pre-training text, about 12,200 hours of speech, 2.2M Slovenian instruction pairs and 30 evaluation sets.

How to explore it

Browse the overview by lifecycle stage, then skim corpora in the concordancer before you download anything.

Plan

1. CLARIN.SI in five minutes
2. The repository through an AI lens
3. Pre-training data
4. Speech data
5. Post-training and evaluation
6. Exploring corpora in the concordancer
7. Gaps and opportunities

PART 1

CLARIN.SI in five minutes

The Slovenian node of the European research infrastructure CLARIN

CLARIN.SI: the Slovenian CLARIN consortium

- National consortium of **CLARIN ERIC**, the European research infrastructure for language resources and technologies
- Working since **2015**, with its seat at the **Jožef Stefan Institute**
- Focus on **Slovenian and the other South Slavic languages**, open to all languages
- Serves the humanities, the social sciences and everyone who works with language data, including AI developers
- Funded by the Slovenian Ministry of Education, Science and Sport and the Ministry of Culture

12 partners

4 universities · 4 research institutes · the National and University Library · the Slovenian Language Technologies Society · 2 companies

Sources: clarin.si/info/about, clarin.si/info/partners

What CLARIN.SI runs

Repository

Certified with the **CoreTrustSeal** and as a **CLARIN B-centre**. FAIR deposits with persistent handles, explicit licences and versioning. About 9 TB of data in more than 90 languages.

Concordancers

noSketch Engine and KonText over **100+ corpora in 30+ languages** (about 20B words). About **1 million queries** a year (Dec 2024 to Dec 2025, up 45%).

Tools and services

CLASSLA-Stanza pipeline, WebAnno, CQP/Tree syntactic search, SENTA, Korpusnik, an LLM benchmarks dashboard; code on GitHub.

Knowledge centres

CLASSLA (South Slavic languages), ELEXIS (lexicography), LLMs4SSH (large language models for the social sciences and humanities), CKCMC (computer-mediated communication).

Sources: clarin.si/info/about, clarin.si/info/concordances, CLARIN.SI presentation, January 2026

Why a research repository matters for AI

What you get

- **A licence on every record**, so you know whether you may train on it
- **Persistent handles** to cite and to pin in data cards
- **Versions** you can reproduce a training run against
- **Rich metadata**: size, languages, genre, annotation layers
- Data **curated by the people who built it**, often with a paper

What to watch for

- Several versions of the same corpus; take the latest
- Corpora that repackage other corpora; deduplicate
- Non-commercial and academic-only licences
- Records that only describe data hosted elsewhere

PART 2

The repository through an AI lens

1,293 records, classified by where they help in the model lifecycle

From 1,293 records to 1,038 resources

1,038
resources after collapsing versions and dropping hidden and superseded records

790
open licence that allows training and commercial use (CC0, CC BY, CC BY-SA, Apache, MIT, GPL)

102
non-commercial (79) or restricted to academic users and custom terms (23)

146
records without files: data hosted or queried elsewhere

- **754** resources include Slovenian, **124** cover only other South Slavic languages, **160** cover other languages
- **2026 is the busiest year so far**: 443 new records, including 402 e-book and audiobook samples and most of the new LLM benchmarks

Repository export of 14 October 2026. Full overview: clarinsi.github.io/clarinsi4ai

Six lifecycle stages

Stage	Resources	Open licence	Volume (estimate)
Pre-training data	384	329	≈ 53B words after removing overlaps (123B across all records)
Speech and multimodal data	200	186	≈ 12,200 hours of speech
Post-training data	80	63	≈ 2.2M instruction pairs (Slovenian) and 68 supervised task datasets
Evaluation and benchmarks	30	22	30 datasets covering 21 languages
Lexical and knowledge resources	262	110	≈ 50M lexicon entries; 134 ELEXIS records without files
Models and tools	82	80	68 models and 14 tools; no generative LLM weights

Every stage has its own page with all resources, links and licences. Volumes are rough: many corpora repackage others.

Coverage by language and stage

Slovenian leads at every stage; Croatian, Serbian and English are the only others at all six

Downloadable resources per language and lifecycle stage. South Slavic languages first.

	Pre-training	Speech	Post-training	Evaluation	Lexical	Models & tools
Slovenian	297	192	60	23	119	43
Croatian	26	6	14	6	10	14
Serbian	22	8	11	3	6	11
Bosnian	13	2	7	1	3	–
Montenegrin	11	–	1	–	–	–
Macedonian	9	–	3	1	2	5
Bulgarian	11	–	3	–	3	5
English	39	1	14	11	19	7
German	7	–	4	2	6	–
Russian	8	–	3	3	3	1
Estonian	7	–	5	1	2	1
Spanish	6	–	4	2	2	–
Hungarian	5	–	3	–	5	–
Czech	4	2	3	1	2	–
French	4	–	3	2	3	–
Polish	4	2	3	–	3	–
Greek	7	–	2	–	1	1
Latvian	5	–	3	1	1	1
Ukrainian	7	–	1	–	2	1
Albanian	3	1	2	–	3	1

Reading the licences

Licence on the record	Training a model	Commercial use	Resources
CC0, CC BY, CC BY-SA, Apache, MIT, GPL	yes	yes	790
CC BY-NC, CC BY-NC-SA, CC BY-NC-ND, MULTEXT-East	research only	no	79
ACA (academic users), CLARIN.SI web-corpus licence	after sign-in, academic use	no	23
No licence, no files	access elsewhere	check provider	146

> A rough guide only. Check the licence text, and how attribution and share-alike apply to model weights and outputs, before you train.

PART 3

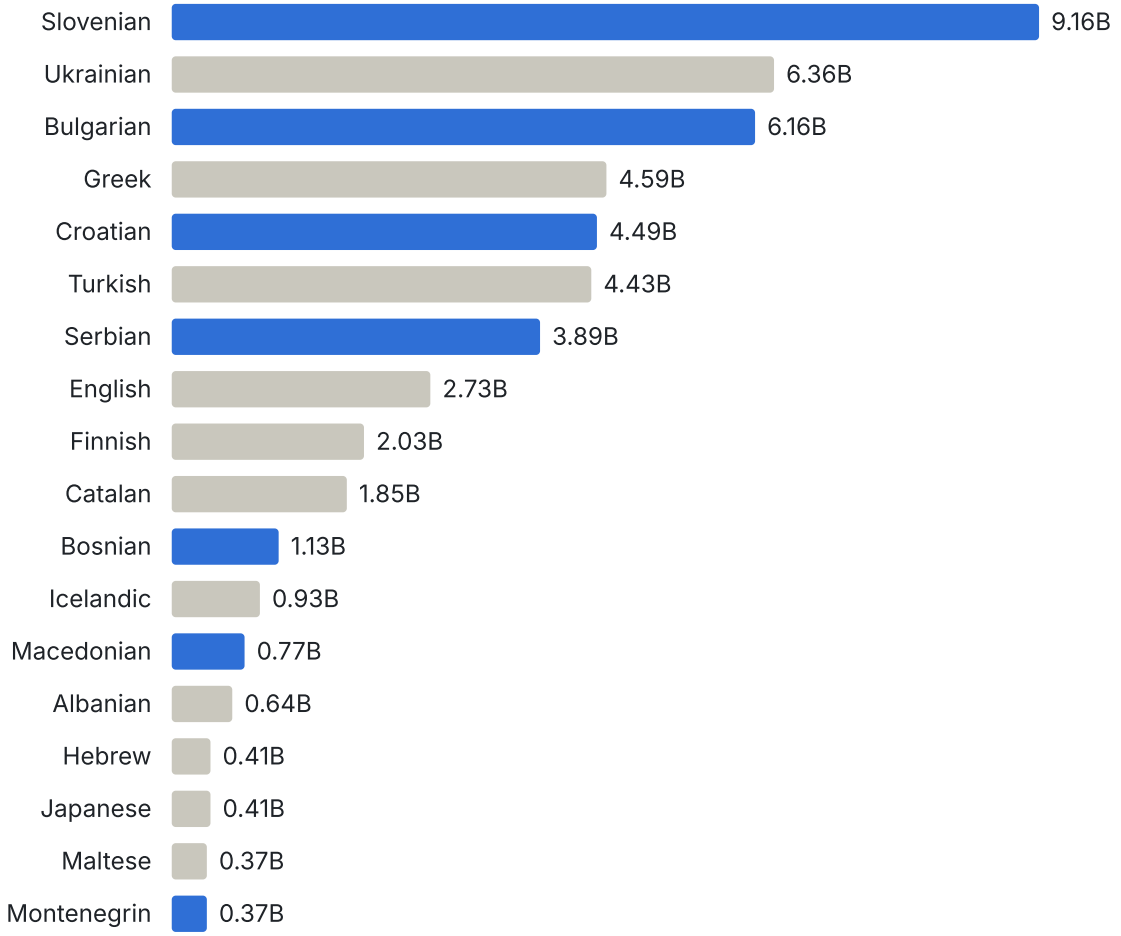
Pre-training data

About 53B words after removing overlaps, 50B of them openly licensed

Pre-training text per language, roughly

Slovenian has about 9B words of pre-training text; five languages have 4–6B

Rough estimate of words per language after removing known overlaps, billion words. South Slavic languages in blue.



CLASSLA-web 2.0: the current South Slavic reference

Language	Words
Bulgarian	5.99B
Serbian	3.71B
Croatian	3.01B
Slovenian	2.31B
Bosnian	1.01B
Macedonian	691M
Montenegrin	294M
Total	**17.0B**

- Crawled in **2024** from national top-level domains
- CC0 released in 2026
- **Genre and topic labels** on every text: filter your pre-training mix
- Linguistically annotated with CLASSLA-Stanza
- JSONL and vertical formats

hdl.handle.net/11356/2079

MaCoCu and older crawls

MaCoCu (CC0, 2023)

- Ukrainian 6.2B · Greek 4.4B · Turkish 4.3B · Catalan 1.7B · Icelandic 887M · Albanian 626M · Maltese 348M words
- Plus South Slavic crawls (2.5B Serbian, 2.4B Croatian); MaCoCu-sl is near-identical to CLASSLA-web.sl 1.0
- **MaCoCu-Genre**: 28.5B words in 13 languages with genre labels
- English-paired parallel versions for 14 languages

Older and overlapping collections

- **BERTić-data**: 8.4B words of Bosnian, Croatian, Montenegrin and Serbian, deduplicated, used to train BERTić
- **WaC corpora** (hrWaC, srWaC, bsWaC, fiWaC): crawls from before 2016
- **CLASSLA-web 1.0**: superseded by 2.0

> Combining crawls adds diversity but also near-duplicates. Deduplicate across collections, not only within one.

Slovenian beyond the web

Resource	Content	Words	Licence
metaFida 1.0	Combined, deduplicated Slovenian corpora	4.7B	concordancer only
OSS 1.0	Scientific texts from the Open Science Slovenia portal	2.6B	CC BY-SA
KAS 2.0	Academic theses (BA, MA, PhD)	1.5B	ACA
Gigafida 2.2	Reference corpus of written standard Slovenian	1.1B	concordancer only
Trendi 2026-09	Monitor corpus of news, updated monthly	1.1B	concordancer only
COLESLAW 1.0	Legislative, judicial and governmental legal texts	772M	CC BY
sPeriodika 1.0	Periodicals 1771–1914 (OCR)	708M	CC BY-SA

The three largest curated collections can be searched online but not downloaded. OSS and COLESLAW are the largest open domain corpora.

Parliamentary debates: ParlaMint and friends

- **ParlaMint 5.0**: 29 parliaments, **1.2B words**, speaker-attributed, rich metadata, CC BY
- **ParlaMint.ana**: the same, linguistically annotated
- **ParlaMint-en.ana**: machine translation of all 29 into English, **1.4B words**
- **siParl 4.0**: Slovenian parliament 1990–2022, 239M words
- Historical: yu1Parl (1919–1939), Kranjska provincial assembly, SlovParl
- New ParlaMint corpora: Israel (409M), Ukraine, Canary Islands

> Formal spoken-written register, balanced topics, clean transcripts and speaker metadata: useful for pre-training, for dialogue-style data and for building evaluation sets.

Books, user-generated and historical text

Books (2026)

231 e-book samples from 6 Slovenian publishers, the opening part of each book, EPUB, CC BY-SA .

The only sizeable block of contemporary edited long-form prose and poetry.

User-generated

Janes (2017): Slovenian tweets, forums, blogs, news comments, Wikipedia talk, 240M words.

FRENK comment corpora: 325M (sl) and 408M (hr) words.

Large Croatian, Estonian and Latvian comment sets are ND .

Historical

sPeriodika 708M words, IMP 18M, KDSP 12M, plus manuscripts and legal texts.

A post-OCR correction set exists for sPeriodika.

Parallel data

- **MaCoCu English pairs** for 14 languages, CC0 : Ukrainian 239M, Greek 151M, Catalan 143M, Croatian 100M, Serbian 96M, Slovenian 85M words ...
- **DGT-UD**: EU translation memories, 23 languages, 2.1B words, parsed
- **RSDO4 2.0**: 3.1M Slovenian–English sentence pairs from public institutions and donors, for MT, CC BY-SA
- **KAS-MT**: Slovenian–English abstracts from theses, 22M words

> Use these for translation, cross-lingual alignment and for adding English context during continued pre-training. ParlaMint-en offers a 29-language set of machine-translated parallel data.

Pre-training checklist

Choosing

1. Prefer the **latest version** of each resource
2. Check the **licence tier** against your use
3. Pick the **newest crawl** per language, add older ones for diversity
4. Use **genre and topic labels** to balance the mix

Processing

5. **Deduplicate** across collections
6. Expect formats: vertical, CoNLL-U, TEI, JSONL, plain text
7. Strip or keep annotation layers (lemmas, UD, NER) as needed
8. Cite the **handle** in your data card

All pre-training resources: clarinsi.github.io/clarinsi4ai/01-pretraining.html

PART 4

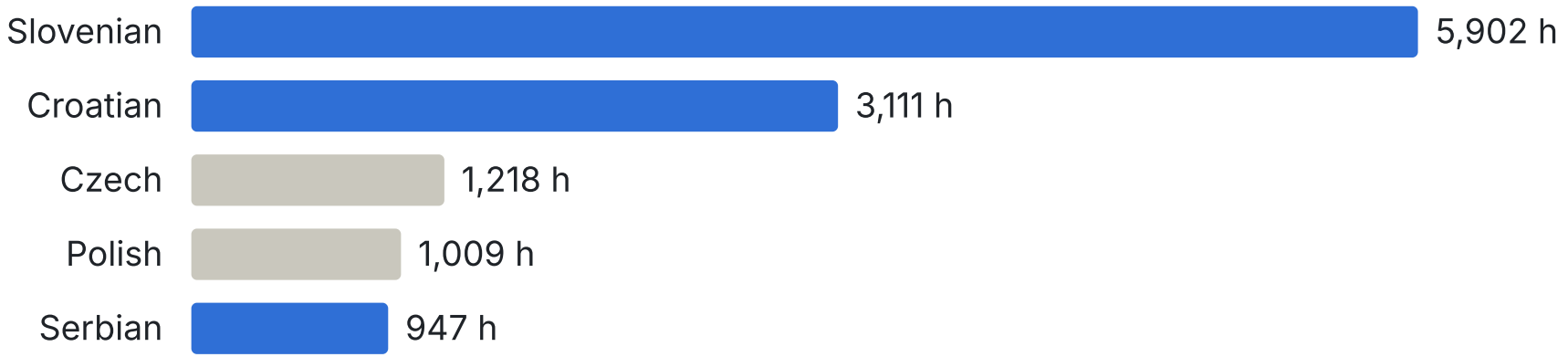
Speech data

About 12,200 hours of recorded speech, most of it aligned with transcripts

Hours of speech per language

Slovenian has about 5,900 hours of speech and Croatian 3,100; the others about 1,000 each

Hours of recorded speech per language. South Slavic languages in blue.



Each record's duration counted once; transcript-only and derived releases left out.

ParlaSpeech 4.0: 10,313 hours in five languages

Language	Hours	Sentences
Slovenian	4,080	1,435k
Croatian	3,110	923k
Czech	1,218	718k
Polish	1,009	535k
Serbian	896	291k

Transcriptions · FLAC audio · CC BY-SA

- Parliamentary transcripts aligned with the parliaments' public recordings, one sentence per instance with speaker metadata
- Filled and silent pauses for all languages, word stress for Croatian and Serbian

Slovenian speech: about 5,900 hours

Resource	What it is	Hours	Licence
ParlaSpeech 4.0	Slovenian parliament, aligned transcripts and audio	4,080	CC BY-SA
ARTUR 1.0	Speech database built for ASR; 884 of the 1,067 hours transcribed	1,067	CC BY-SA
Gos 2.1	Reference spoken corpus with video	319	ACA
Audiobook samples	171 samples from 11 publishers, MP3, no transcripts	304	CC BY-SA
Zila 1.0	Gail Valley dialect, transcribed	100	CC BY-NC-SA

Smaller sets: SOFES, Berta, MoJezik, Krohot, GORDAN, EVA, ROG-Dialog (dialogue acts, humour, multimodal conversation).

Dialects, multimodal data and speech models

Dialect and heritage speech

Gail Valley (Zila), Torlak (Serbian), Chakavian "Mići Princ", heritage BCMS in Germany, Albanian in Kosovo, primary-stress datasets.

Multimodal (2025–2026)

SLO-VLM-IT: 1.1M items for vision-language instruction tuning NC .

Ilustrirani Slovenec: historical newspaper pages for document understanding NC .

Speech models

RSDO and PROTOVERB Conformer ASR for Slovenian, a Gail Valley dialect ASR model, punctuation and text (de)normalisation, and an **ASR model evaluator**.

All speech resources: clarinsi.github.io/clarinsi4ai/02-speech-multimodal.html

PART 5

Post-training and evaluation

Instruction data, supervised task data and benchmarks

Instruction data: Slovenian only, mostly medical

Dataset	Pairs	How it was made
MED-Termset	975k	Wikidata terms, LLM-enriched
MED-Anatomy	712k	Expert terminology, LLM-enriched
PHARMA	482k	Medicinal product databases
MED 2.0	25k	Generated and translated, validated
SloLexQA	16.5k	Lexicographic QA
DH	10.2k	Written by hand, digital humanities
GEN	6.8k	GPT-4, manually validated
SAFE 0.5	501	Human prompts, refusals

≈ 2.2M

prompt–response pairs, 98% medical or pharmaceutical

- No preference data; text instructions are single-turn (only the vision-language set SLO-VLM-IT has conversations)
- No instruction data for other South Slavic languages

All rows except SloLexQA are GaMS-Instruct datasets.

Hidden supervised fine-tuning material

Expert Q&A archives

- **ŠUSS** (1998–2010): questions about Slovenian answered by linguists, 365k words
- **TermSvet** (2026): terminology consulting service, 282k words
- **Terminologišče**: terminological counselling, 509 answers

Real user questions with expert answers, all CC BY .
Not yet in instruction format.

Dictionaries as seed data

- Definitions, examples, collocations and synonyms turn into QA pairs; SloLexQA shows how
- Sloleks 3.1 (372k lemmas), Thesaurus of Modern Slovene (102k), KSSS collocations (78k), two wordnets

68 task-specific supervised datasets

10

Text-to-text: grammatical error correction (Šolar 3.0), normalisation (Janes-Norm), simplification, summarisation (KAS-Sum), post-OCR.

21

Linguistic annotation: SUK, hr500k, SETimes.SR, ReLDI (tagging, parsing, NER), coreference, multiword units.

10

Semantics: word-sense disambiguation, relation extraction, definitions, commonsense (SloATOMIC).

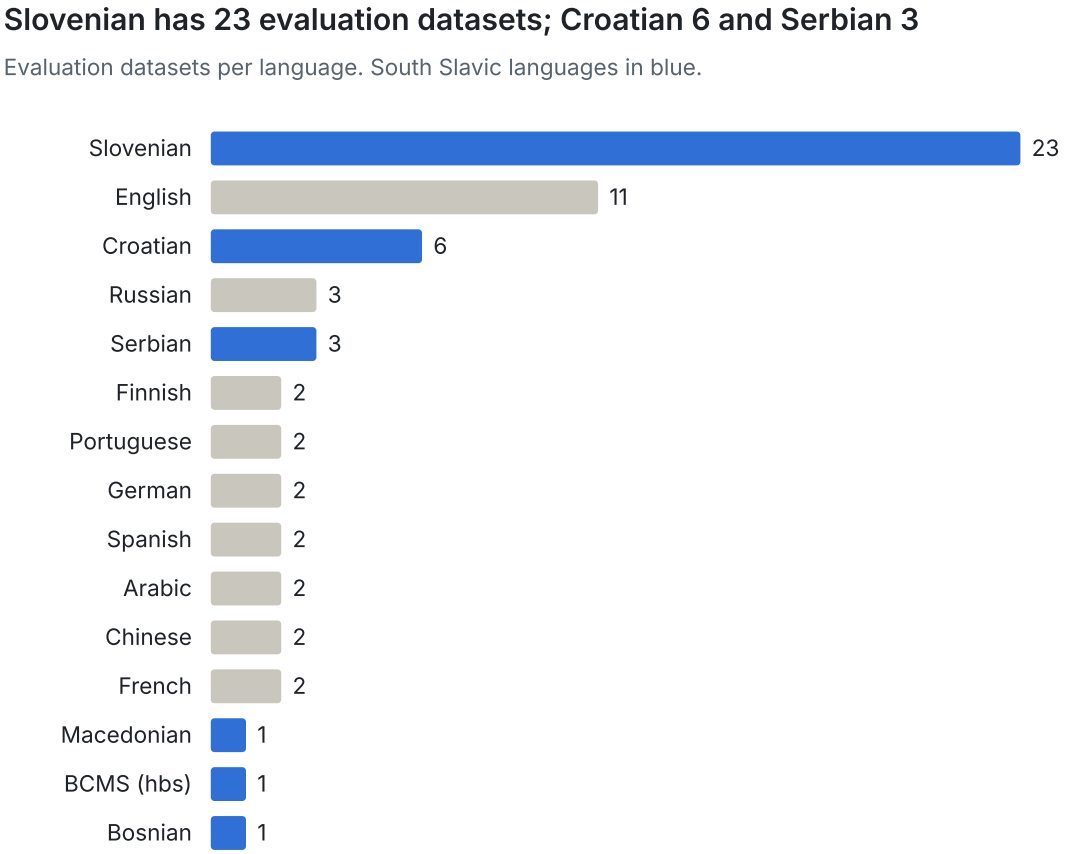
27

Classification: sentiment (ParlaSent, SentiNews), hate speech (FRENK, IMSyPP), genre (X-GENRE), topic, language identification.

Convert them into instructions for task-aware fine-tuning, or use their test splits for evaluation. Croatian and Serbian are well covered here.

clarinsi.github.io/clarinsi4ai/03-post-training.html

Evaluation: a Slovenian suite, thinner for related languages



Slovenian

- SuperGLUE and SQuAD 2.0 translations, SI-NLI, SloWiC, KE-WSC
- Bias: EEC-SL · grammar checking: Šolar-Eval, DASSLE
- Lexical knowledge with distractors: SYNDIST, COLLDIST

Other South Slavic

- COPA for Croatian, Serbian, Macedonian and dialects
- PragMega, CoSimLex (Croatian) and AspectBench (bs, hr, sr)

The **SloBENCH** framework runs the evaluation for a public leaderboard.

New in 2026: LLM benchmarks

WALS-Bench

Metalinguistic knowledge across many languages, 77k items, CC BY .

GramQA

Corpus-grounded grammatical question answering, multilingual, 1.9k items.

PragMega

Pragmatics (metaphor, irony, humour) for Czech, Slovenian and Croatian.

SLawNLI

Natural language inference over Slovenian law, 2.2k pairs.

ETHICS-EN-SL

Moral judgement sample in English and Slovenian, 500 items.

AI-GenT

Human vs machine-generated text in English and Slovenian, 3.3k texts.

clarinsi.github.io/clarinsi4ai/04-evaluation.html

Lexical resources and models, briefly

Lexical and knowledge resources (262)

- Morphology: Sloleks 3.1, hrLex, srLex
- Semantics: sloWNet, OSWN, thesaurus, collocations, sentiment lexicons
- Terminology, frequency lists, n-grams
- For AI: tokeniser checks, synthetic data, retrieval grounding, lexical probes

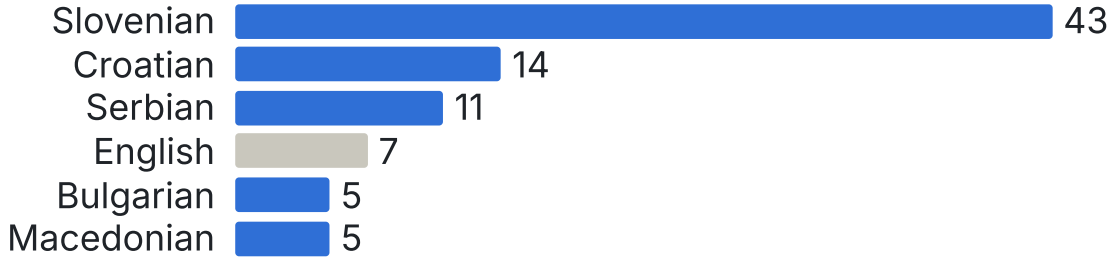
Downloadable resources per language (128 with files)



Models and tools (82)

- Encoders: SloBERTa, CroSloEngual BERT
- **CLASSLA-Stanza**: lemmas, tags, parses, NER for 5 languages; annotate your own data
- ASR, MT, punctuation, summarisation, QA models
- No generative LLM weights here; the GaMS models are on Hugging Face

Downloadable resources per language (82 with files)



PART 6

Look before you download

The CLARIN.SI concordancers

The CLARIN.SI concordancers

- **100+ corpora, 30+ languages**, about 20B words, one back-end
- **noSketch Engine**: open, no login; a login version adds subcorpora
- **KonText**: plays audio for spoken corpora
- **CQP/Tree**: syntactic queries over parsed corpora
- Concordances, frequency lists, collocations, keywords, text-type filters

Source: clarin.si/info/concordances

Why AI developers should care

- Skim a corpus before downloading gigabytes
- Check domains, genres and noise in a web crawl
- Find examples for prompts, tests and error analysis
- Query corpora that cannot be downloaded: **metaFida, Gigafida, Trendi**

Corpora from this overview you can search online

Stage	Corpora that mention concordancer access in their repository record
Web	CLASSLA-web 1.0 for all seven South Slavic languages (sl, hr, sr, bs, cnr, mk, bg), meWaC
Reference	metaFida (4.7B), Gigafida 2.2, ccGigafida
Domains	KAS and its BA, MA and PhD sub-corpora, KZB, ccMAKS, SBSJ
Historical	sPeriodika, IMP, goo300k, CVET, Trubar
Parliament and literature	SlovParl, Maj68
Speech	Gos, ROG, ROG-Dialog

36 records in the overview mention the concordancers. Many more corpora are online: open the corpus list to see them all.

Example: is this web corpus right for my domain?

1. Open noSketch Engine and select a CLASSLA-web corpus
2. Search a domain term, e.g. `[lemma="zdravilo"]` (medicine)
3. Open **text types**: which genres and domains does it come from?
4. Read 50 concordance lines: boilerplate, machine translation, spam?
5. Compute **collocations** and a **frequency list** by genre
6. Download the corpus from its repository handle, filtered by the genres that worked

> **Live demo**
> Ten minutes in noSketch Engine are often enough to decide which corpus and which genres to include in a training mix.

PART 7

Gaps and opportunities

What is missing for modern AI development

1. **No preference data and little dialogue data.** Text instructions are single-turn and 98% medical or pharmaceutical; only SLO-VLM-IT has (machine-generated) conversations.
2. **Instruction data exists only for Slovenian.** Croatian, Serbian, Bosnian, Montenegrin, Macedonian and Bulgarian have large web corpora but no instruction or safety data.
3. **Evaluation for other South Slavic languages is thin.** Beyond COPA, only PragMega, CoSimLex, AspectBench and word analogies cover them.
4. **Almost no speech data for Bosnian, Montenegrin, Macedonian or Bulgarian.** ParlaSpeech 4.0 covers Croatian, Czech, Polish, Serbian and Slovenian; Bosnian has only small heritage-speaker sets.
5. **The largest curated Slovenian corpora cannot be downloaded.** metaFida, Gigafida and Trendi are concordancer-only.
6. **Metadata does not describe AI readiness.** 725 free keywords, and formats often hidden in zip files.

How you can help, and how to start

Contribute

- Deposit your training and evaluation data, with a clear licence
- Build instruction and preference data for South Slavic languages
- Port existing Slovenian benchmarks to related languages
- Report what made a resource hard to use for training

Start here

- Overview by lifecycle stage: clarinsi.github.io/clarinsi4ai
- Repository: clarin.si/repository
- Concordancer: clarin.si/ske
- About CLARIN.SI: clarin.si/info/about



Thank you. Questions?

clarinsi.github.io/clarinsi4ai · www.clarin.si

Nikola Ljubešić · Jožef Stefan Institute · CLARIN.SI