Slovenian Lexicographic Datasets

dict-conversions

Two LM-ready datasets built from **15 legacy CJVT/DZS Slovenian dictionaries** (14 bilingual + 1 monolingual encyclopedia), converted to OASIS **DMLex 1.0** and HuggingFace **Parquet**, then enriched. This document describes both datasets, the pipeline that produced them, their schemas, and **what to look for** when using them.

> **Status:** 805,279 entries · 1,019,685 senses · 2,174,880 translation pairs · 113,868 relations. All 15 dictionaries validate against the official DMLex XSD 1.1 + JSON Schema with **0 leaks, 0 parse errors, 0 residual glyph markers**. 52 unit tests pass.

---

1. The two datasets at a glance

	Collection 1 — Core	Collection 2 — Enriched
Path	`dist/core/` (+ canonical `lexidma/`, `parquet/`)	`dist/enriched/` (extends Core)
Provenance	**Intrinsic** — derived only from the project's own dictionaries	**Extrinsic** — Core + external resources
External tools	none	CLASSLA-Stanza, sloWNet/OMW, English WordNet (oewn)
Size	~2.1 GB	~184 MB (layers only; use *with* Core)
Reproducible offline	yes	needs the external resources (one-time download)
Contents	DMLex XML/JSON, derived tables, 12 LM task JSONLs	silver morphology + MSD task, synset/ILI links, imported antonyms, candidate scoring, sloWNet-enriched KNAUR DMLex

Use them together. Enriched is a thin layer of external-resource columns/files that *extends* Core; it does not duplicate it.

---

2. Source corpus

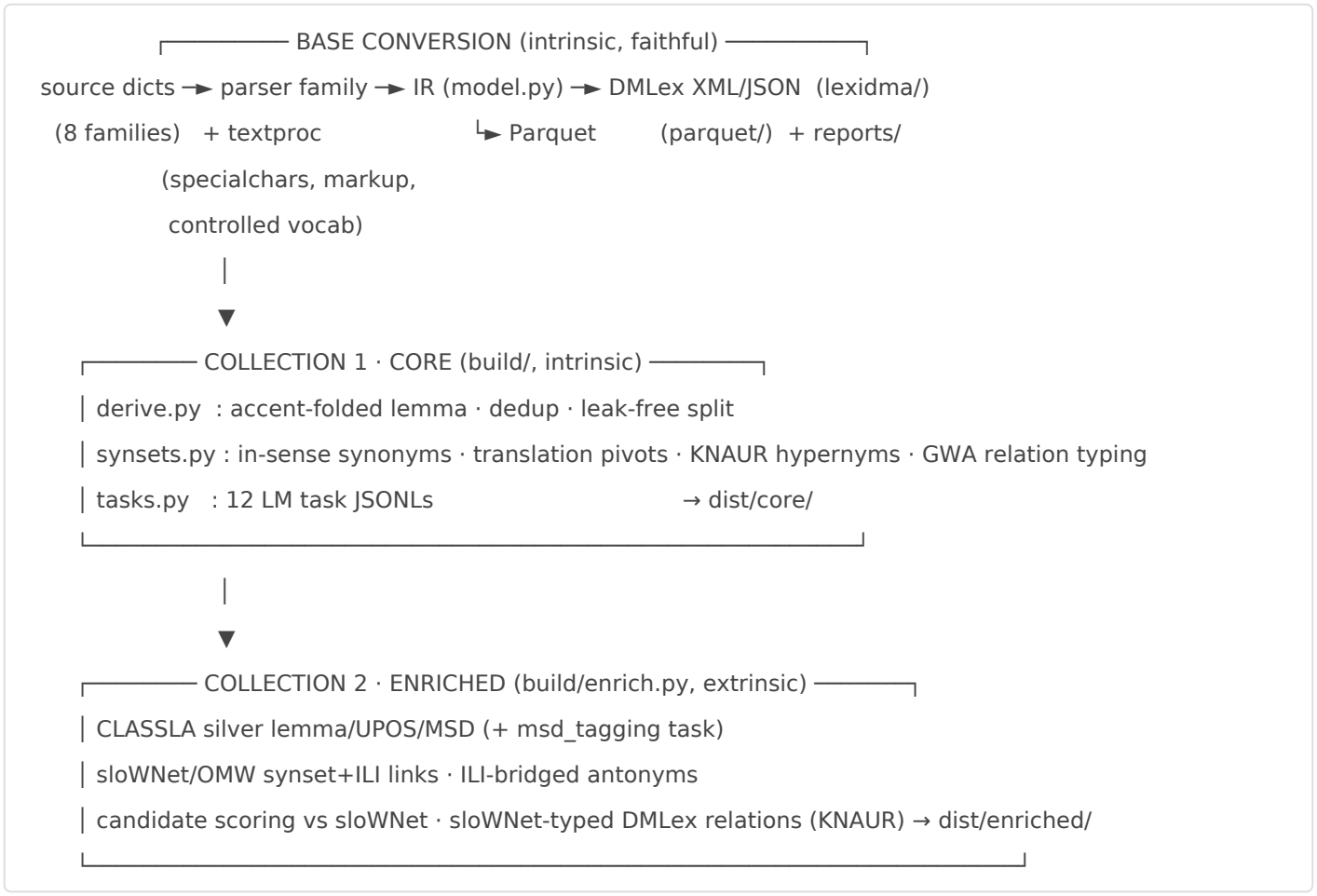
`sh` = Serbo-Croatian (legacy unified tag). KNAUR is monolingual (definitions + cross-references, no translations). DVRUSL is reconstructed from a Word `.doc` (OCR-grade).

Code	Languages	Family	Entries	Senses	Pairs	Relations
DVANSL	en→sl	block_blankline	68,761	75,208	264,454	0
DVFRSL	fr→sl	line	39,980	70,817	175,458	0
DVITSL	it→sl	block_blankline	60,282	98,444	197,449	0
DVRUSL	ru→sl	doc_runs	32,844	32,844	80,364	0
DVSHSL	sh→sl	block_blankline	92,302	92,302	150,228	5,563
DVSPSL	es→sl	dzs_utf16	36,460	66,722	114,691	414
DVSLAN	sl→en	block_blankline	42,088	42,227	159,625	0
DVSLFR	sl→fr	line	33,866	44,092	128,561	34
DVSLNE	sl→de	german	84,951	90,374	252,074	120
DVSLSH	sl→sh	block_blankline	72,690	87,140	152,068	2,694
DVSLSP	sl→es	dzs_utf16	33,676	66,456	93,562	349
DRSLAN	sl→en	dzs_nested	25,559	33,273	75,374	457
VSIS	sl→it	block_geslo	90,675	116,223	275,043	7,764
LAT_AZ	la→sl	block_geslo	11,521	20,315	55,929	1,852
KNAUR	sl (mono)	encyclopedia	79,624	83,248	0	94,621

Several pairs exist in **both directions** (DVSLFR/DVFRSL, DVSLAN/DVANSL, DVSLSH/DVSHSL, DVSLSP/DVSPSL, VSIS+DVITSL) — exploited for translation-pivot synonyms and reverse-dictionary tasks.

---

3. The pipeline



Stage 1 — Base conversion (`src/dictconv/`, command `dictconv convert all`). Each dictionary's publisher typesetting markup is tokenized (not XML-parsed — it is malformed), `{…}` escapes decoded to Unicode (side-aware accents; undecodable font/template codes removed and logged), qualifiers classified into a controlled vocabulary, and emitted as a source-agnostic intermediate representation. Serializers produce **DMLex 1.0 XML+JSON** (validated against the official OASIS XSD 1.1 / JSON Schema) and three **Parquet** artifacts. The conversion is *faithful*: it never invents content and flags rather than silently drops.

Stage 2 — Core (`dictconv build-core`). Adds derived, ML-oriented layers computable from our data alone: a normalized lemma/UPOS layer, exact/near-duplicate collapse, a **leak-free** train/dev/test split, synonym/pivot/hypernym candidate tables, and 12 instruction-style task JSONLs.

Stage 3 — Enriched (`dictconv enrich`). Adds layers that need outside resources, kept strictly separate from the intrinsic data (gold-vs-silver provenance is explicit).

---

4. Base conversion artifacts

`lexidma/<CODE>.xml` and `.json` — OASIS DMLex 1.0

Faithful lexicographic record. Camel-case elements; text of every object except `headword` in a nested `<text>`; crosslingual module (`headwordTranslation`, `exampleTranslation`); the linking module (

`relation` + **`relationType`** **definitions that carry a Global WordNet** **`sameAs`** **URI**, e.g. `see → https://globalwordnet.github.io/schemas/wn#also`). Validated with a real **XSD 1.1** processor (all identity constraints + assertions, except 3 cardinality-defective ones the published schema cannot satisfy). KNAUR uses the monolingual schema.

`parquet/<CODE>.entries.parquet` — one row per entry (nested)

```
dict_code, entry_id, headword,
lemma,            # accent-folded join key (Sloleks/Gigafida/CLASSLA/sloWNet keying)
accented_form,    # original tonal/accented display form (null if == headword)
homograph_number, source_lang, target_lang,
meta_lang,        # editorial metalanguage = "sl"
parts_of_speech[str], upos,    # UPOS from the entry's first POS (intrinsic static map)
frequency_band,   # DRSLAN corpus band 0..3 (null elsewhere)
labels[str], collocates[str],  # DRSLAN <KO> collocates
pronunciations[{text, scheme}],
inflected_forms[{text, tag}],
senses[{ sense_id, indicator, labels[str], definitions[str],
        headword_translations[{text, lang_code, parts_of_speech[str], labels[str]}],
        headword_explanations[{text, lang_code}],
        examples[{text, labels[str], translations[{text, lang_code, labels[str]}]}] }],
has_content,      # False => no senses / all senses empty (filter before LM use)
source_ref, raw   # provenance
```

`parquet/<CODE>.pairs.parquet` — one row per (source,target) unit (flat)

```
dict_code, source_lang, target_lang, entry_id, sense_id, homograph_number,
pair_type (headword|example), source_text, target_text,
source_lemma,    # accent-folded entry headword (dedup + leak-free split key)
part_of_speech, labels[str], domain, register
```

`parquet/<CODE>.relations.parquet` — the full cross-reference graph

```
dict_code, source_lang, target_lang, relation_index, type, description,
members[{ref, headword, role, target_id}],
serialized       # True => >=2 members resolved => present in DMLex XML/JSON
```

This is the **lossless** home of the cross-ref graph: it keeps cross-references whose target never resolved to an entry id (which the DMLex XML/JSON must drop).

`reports/<CODE>.report.json` + `reports/_summary.json`

Per-dictionary stats, validation results, the controlled-value inventory, flagged-token counts, and the aggregate summary + artifact manifest (sha256 of every output).

---

5. Collection 1 — Core (`dist/core/`)

5.1 Derived tables (`dist/core/derived/`)

File	Rows	What it is
`lemmas.parquet`	805,279	one row per entry: lemma, accented_form, upos, frequency_band, cluster_id, split
`pairs_dedup.parquet`	2,078,214	de-duplicated translation pairs + `occurrence_count`, `canonical_id`, split
`synonym_sets.parquet`	378,783	in-sense (target-language) near-synonym sets + gloss
`synonym_pairs.parquet`	2,131,996	in-sense + pivot synonym pairs (evidence, confidence_tier)
`pivot_synonyms.parquet`	152,655	Slovene synonym candidates from translation pivots (GOLD 67,296 / SILVER 85,359)
`hypernym_candidates.parquet`	4,169	KNAUR genus-differentia hypernym candidates (confidence)
`relations_typed.parquet`	113,868	the cross-ref graph, GWA-typed (`gwa_relType`)

> Pivot-synonym yield: 152,655 SILVER+GOLD pairs (≥2 agreeing pivots) materialized; **632,231** single-pivot BRONZE pairs were counted but **not** materialized (large, low precision); 339,415 distinct pivots used.

5.2 LM tasks (`dist/core/tasks/*.jsonl`)

Each row: `{id, task, split, input:{…}, output:{…}, metadata:{…}}`. Split is `train/dev/test` (≈90/5/5), leak-free (§5.3). Marker policy `drop` (undecodable-glyph rows are cleaned/omitted).

Task	Rows	input → output

`translation`	3,006,506	`{source_text, source_lang, target_lang, part_of_speech, labels, domain, register}` → `{target_text}` (both directions)
`example_translation`	574,961	example phrase → its translation
`definition`	145,633	`{headword, lang, indicator}` → `{definition}` (KNAUR)
`reverse_dictionary`	145,633	`{definition, lang}` → `{headword}`
`wsd`	22,036	`{word, context, lang}` → `{sense_gloss, sense_id}` (polysemous only; bare-number glosses dropped)
`example_usage`	85,155	`{headword, lang}` → `{example}` (monolingual usage sentences)
`morphology`	332,393	`{headword, lang}` → `{form, tag}` (dictionary inflected forms)
`pronunciation`	157,578	`{headword, lang}` → `{transcription, scheme}`
`synonyms_of`	358,573	`{word, lang}` → `{synonyms[]}` (a real **set**; 60% have >1)
`hypernym_of`	4,169	`{word, lang}` → `{hypernym_candidate, confidence}`
`relation`	113,373	a relation's first member → `{relation_type, members[…]}`
`relation_classify`	113,327	`{a, b, lang}` → `{relation_type}` (unordered-pair split)

5.3 Leak-free split (important)

- **Translation / sense / synonym tasks** key on the **folded Slovene lemma**, so a lemma and its reverse-direction twin (`hiša` in sl→fr and `maison→hiša` in fr→sl) are always in the **same** split. Verified: **0 of 172,081** Slovene headword lemmas straddle splits. (e.g. translation split: train 2,711,054 / dev 144,596 / test 150,856.)
- **Morphology & pronunciation** (about the headword *form*) split by **headword form**; `relation_classify` by the **unordered member pair** — so foreign homographs don't straddle either.
- The legacy `dict_code:entry_id` key (now superseded) leaked ~26 % of multi-dict lemmas.

5.4 Cleaning applied (Core)

Dedup before split (with `occurrence_count`); degenerate targets dropped (punct/digit-only, single-char, `src==tgt`); unbalanced parentheses balanced; PUA sentinels + control chars stripped; undecodable glyph markers removed (`marker_policy=drop`). `manifest.json` carries a content hash (`e76f2766…`) and per-task split counts; `dataset_card.md` is the in-tree card.

---

6. Collection 2 — Enriched (`dist/enriched/`)

File	Rows	What it is
`silver_morphology.parquet`	208,715	CLASSLA lemma / UPOS / **JOS-MULTEXT-East MSD** / feats per Slovene lemma; `morph_provenance="silver_tool"`
`tasks/msd_tagging.jsonl`	208,715	`{lemma, lang}` → `{upos, msd, feats}` (the morphology/POS task; **silver**)
`synset_links.parquet`	64,413	Slovene lemma → sloWNet/OMW `synset_id` + **ILI** (join key to Princeton WN / OMW)
`antonyms.parquet`	6,107	imported Slovene antonyms (ILI-bridged through the English WordNet)
`scored_synonyms.parquet`	364,334	every Core synonym candidate + `wordnet_confirmed` + `source_count`
`scored_hypernyms.parquet`	396	checkable hypernym candidates + `wordnet_confirmed`
`lexidma/KNAUR.{xml,json}`	97,418 rel	KNAUR re-serialized as DMLex with sloWNet **antonym (142)** + **synonym (2,655)** relations, ILI/synset in `relation/description`, GWA-typed

6.1 Candidate scoring vs sloWNet (measured precision — lower bounds; sloWNet is incomplete)

- **Synonyms:** 364,334 checkable, **17.0 %** confirmed — in-sense 14.6 %, pivot 27.1 %, **pivot-GOLD 38.5 %**.
- **Hypernyms:** 396 checkable, **47.5 %** confirmed (ILI-bridged through the English WordNet).
- Use `wordnet_confirmed=True` (and/or `confidence_tier=GOLD`) to extract a higher-precision subset.

6.2 External resources & how to reproduce

sloWNet/antonyms run in the main `.venv` (`pip install -e '.[enrich]'` → `wn`). **CLASSLA needs Python ≤ 3.13** (its pinned numpy fails to build on 3.14), so run the silver morphology from a 3.12 env:

```
uv venv --python 3.12 .venv-enrich
uv pip install -p .venv-enrich/bin/python -e '.[enrich]'
python -m wn download omw-sl ; python -m wn download oewn:2021
.venv-enrich/bin/python -c "import classla; classla.download('sl')"
.venv-enrich/bin/python -m dictconv.cli enrich --in dist/core --out dist/enriched --sample-limit 0
```

7. What to look for (usage guidance & caveats)

Filter before training

- `has_content` — drop entries with no usable content (≈1.2 % of entries) for entry-level tasks.
- `marker_policy` — task JSONLs are already built with `drop`; never train on the `keep` variant (it would teach the model to emit placeholder glyphs). Corpus markers are currently **0**.
- **Degenerate rows** — already removed from the tasks; if you build your own from `parquet/`, apply the same filters (punct/digit-only, `src==tgt`, unbalanced parens).
- **Use the provided split.** Re-shuffling by row re-introduces lemma leakage; the cluster split is the point. Hold out **whole lemma clusters**, not rows.

Candidates are candidates, not gold

- `synonym_*`, `pivot_synonyms`, `hypernym_candidates` are **induced** and noisy. Gate with the enriched scoring: `scored_synonyms.wordnet_confirmed` / `pivot confidence_tier=GOLD` (38.5 % precision) for synonyms; `scored_hypernyms.wordnet_confirmed` for hypernyms. The 47.5 % hypernym figure is measured on a small checkable slice and is optimistic for the full pool (genus heads are not lemmatized — ~25 % are oblique forms; lemmatize with CLASSLA before use).
- **Antonyms are imported, not mined** (synonyms/antonyms are translationally indistinguishable).
- **Precision is measured only on vocabulary sloWNet already has (~11 %).** The extension value (the ~89 % of members not yet in sloWNet) is **unproven** — commission a small human eval before treating those as silver.

Gold vs silver

- The **gold** lemma layer (`lemmas.parquet`) is 100 % coverage, accent-folded, NFC-clean.
- The **silver** morphology (`silver_morphology` / `msd_tagging`) is CLASSLA tool output (`morph_provenance="silver_tool"`). Keep it filterable; measure MSD accuracy on a hand-tagged sample before training a morphological analyzer on it. The dictionary's own `inflected_forms` are mostly *ending fragments*, not full words.

Per-dictionary quality

- **DVRUSL** (Russian) is OCR-grade (Word `.doc` reconstruction). Cleanups were applied (brace→paren, `|`/bullet stripping, space collapse) but residual noise is inherent — down-weight or exclude for high-precision work.
- **Definitions / reverse-dictionary** come **only from KNAUR** (monolingual Slovene). There are no definitions from the 14 bilingual dicts.

- **Conditioning labels are sparse** (POS on ~18 % of translation rows, domain ~4 %, register ~4 %).

Provenance / audit

- Every removed escape token is logged in `data/reference/removed_markers.tsv`.
- `reports/_summary.json` carries the per-file sha256 manifest; each collection's `manifest.json` carries a `content_hash` and split counts. Pin these with any eval run.

---

8. Loading

```
from datasets import load_dataset

# flat translation pairs (all dicts)
pairs = load_dataset("parquet", data_files="parquet/*.pairs.parquet", split="train")

# nested per-entry records (one dictionary)
entries = load_dataset("parquet", data_files="parquet/VSIS.entries.parquet", split="train")
entries = entries.filter(lambda r: r["has_content"])          # drop empty entries

# a Core LM task, by split
import json
train = [json.loads(l) for l in open("dist/core/tasks/translation.jsonl") if json.loads(l)["split"]=="train"]

# wordnet-confirmed synonyms only (Enriched gate)
import pyarrow.parquet as pq
syn = pq.read_table("dist/enriched/scored_synonyms.parquet").to_pylist()
gold = [r for r in syn if r["wordnet_confirmed"]]

# DMLex (faithful lexicographic view)
import json; res = json.load(open("lexidma/VSIS.json"))       # OASIS DMLex 1.0
```

9. Reproduce end-to-end

```
pip install -e .                # base deps (pyarrow, lxml, jsonschema, xmlschema)
dictconv convert all --write-summary     # base: lexidma/, parquet/, reports/
dictconv build-core                    # Collection 1 -> dist/core/
dictconv enrich --sample-limit 0         # Collection 2 -> dist/enriched/  (see §6.2 for CLASSLA env)
dictconv audit --write-summary           # readiness audit + artifact manifest
```

```
pytest -q                    # 52 tests
```

Revision #6
Created 23 June 2026 08:24:15 by Simon Krek
Updated 23 June 2026 22:54:52 by Simon Krek