

Polish DynaWord: Continuous, Provenance-Aware Corpus Development for Polish

Kacper Wikeł Arkadiusz Słota Dawid Majewski Paweł Puzio
Bartosz Kobyliński Piotr Styła Katarzyna Majzel-Pośpiech

Fabryka AI (SlayerLab)

27 September 2026

Abstract

We present Polish DynaWord, a continuously maintained Polish pretraining corpus. It follows Dynaword, an approach in which a corpus grows through versioned releases and reviewed contributions instead of a single deposit, and applies it to heterogeneous national, European, academic, cultural-heritage, governmental, and web sources. We reconstruct the document and token totals of six stable releases exactly from per-source statistics. Release v0.2.5 contains 4,366,651 documents, 25.27 billion characters, and 9.677 billion tokens across 20 sources. Every document has a licence value, but for most sources that value reflects a collection-level licence rather than a check of each document; 95.8% of documents come from sources that record no author, and four sources carry record-level attribution. We also quantify corpus composition over time: legal and parliamentary material still constitutes 47.0% of tokens in v0.2.5. Deciding what may enter the corpus and deciding how to sample it for training are therefore separate: capping legal text at 15% of tokens while keeping all other material would leave 6.03B tokens. We position the resource relative to Dynaword and identify metadata corrections for the next tagged release.

1 Introduction

Pretraining corpora for large language models are often described by a size, a list of sources, and a licence label. Those summaries are insufficient for a corpus that changes over time. A user also needs to know which source versions belong to a release, whether rights metadata applies at collection or document level, what filtering was actually applied to the released text, and how inclusion decisions changed between releases.

Dynaword proposes treating language corpora as continuously developed resources rather than one-shot deposits [1]. Polish DynaWord applies that model to Polish. The project began with eleven reviewed sources; subsequent releases added nine more, growing from 2.49M to 4.37M documents (Tables 1 and 2). Section 3.3 describes the sources.

This paper is a resource report and longitudinal audit of six stable releases.

Intended uses, commercial and non-commercial, are selecting a Polish subset by licence for pretraining or continued pretraining of small models, and checking evaluation benchmarks for overlap with known training text (contamination audits). No source carries a NonCommercial licence, but the rights to HPLT’s underlying web pages and to some 1000 Novels texts are unresolved (Sections 4 and 6). We report no downstream model evaluation.

Our contributions are:

1. a per-source account of six releases whose document and token totals reconcile exactly with their source statistics;
2. an audit of the released files, finding that filled licence fields do not establish per-document licence evidence, alongside measurements of document length and tokens per word;
3. a domain-composition analysis finding persistent legal and parliamentary concentration, with legal-share caps illustrating why release admission and training-mixture

weighting are separate.

2 Related Work

Continuous and documented corpora. Dynaword defines the immediate methodological basis: versioned source repositories, contribution workflows, dataset cards, and shared validation tooling [1, 2]. Datasheets for Datasets and Data Statements provide broader frameworks for documenting dataset composition, collection, intended use, and linguistic context [3, 4]. The Data Provenance Initiative finds missing or miscategorised licences on widely used dataset hosting sites [18]. The Common Pile assembles public-domain and openly licensed text for pretraining [19].

Position of this work. Polish DynaWord is a small, Polish-focused, continuously maintained corpus. We ask whether a reader can inspect, for each source, why it was admitted, how its text was transformed, how it is attributed, and which releases contain it.

3 Resource and Release Reconstruction

3.1 Canonical representation

From v0.2.1, sources are meant to share one document schema, which we call the *canonical schema*:

```
id, text, source, added, created, token_count,  
license, author.
```

The schema deliberately keeps the main table, the one used for training, compact. It is stored in Apache Parquet, a column-oriented table format. Richer metadata—such as URLs, titles, contributors, licence URLs, and modification notices—can

live in *sidecars*: separate files with one row per document, joined to the main table by id. Each source also has a dataset card, its *source card*, which records how the source was acquired and its statistics. We count tokens with OpenAI’s `cl100k_base` tokenizer [9], a fast, model-independent proxy, and call the results *proxy tokens*; see Section 8 for tokenizer dependence.

3.2 Reconciling release totals

Table 1 lists the totals of all six releases, and Table 2 the source additions and expansions at each step. The per-source statistics of v0.2.0 sum to 6,221,393,871 proxy tokens. For every later release, the document and token increments from source additions and expansions sum exactly to the change in the release total [10, 11]. HPLT enters twice: an initial contribution in v0.2.3 and a much larger expansion in v0.2.4.

Release	Sources	Documents	Tokens (B)
v0.2.0	11	2,490,773	6.221
v0.2.1	12	2,491,773	6.282
v0.2.2	13	2,579,963	6.363
v0.2.3	16	2,710,974	6.881
v0.2.4	17	4,246,429	9.598
v0.2.5	20	4,366,651	9.677

Table 1: Totals of the six stable releases, reconstructed from exact per-source statistics. Token totals are displayed in billions rounded to three decimals; the Hugging Face dataset card often rounds earlier releases to two decimals.

Release	Addition	Documents	Tokens
v0.2.1	1000 Novels	1,000	60,517,363
v0.2.2	gov.pl	88,190	80,666,129
v0.2.3	HPLT (initial)	110,590	512,383,778
	Global Voices	2,040	3,715,020
	NKJP1M	18,381	2,601,201
	Total added	131,011	518,699,999
v0.2.4	HPLT (expansion)	1,493,384	2,043,141,874
	Biblioteka Nauki	42,071	673,488,831
v0.2.5	samorzad.gov.pl	72,771	41,154,506
	Sejm API speeches	38,812	36,335,587
	Wiktionary examples	8,639	1,115,820
	Total added	120,222	78,605,913

Table 2: Source additions and expansions, with exact document and proxy-token counts. Within each release the source rows sum to the difference between the unrounded totals underlying Table 1; subtotal rows are shown for v0.2.3 and v0.2.5.

3.3 v0.2.5 inventory

Table 3 lists the 20 sources of v0.2.5 at revision `7ae4ac7`, covering all sources under the tagged release’s `data/` directory. The largest are HPLT v3 web text from the High Performance Language Technologies crawl [5], EUR-Lex (the EU’s legal database, here its Polish-language texts), the Polish Parliamentary Corpus (a collection of Polish parliamentary proceedings), Polish Wikisource (a library of source texts), Polish Wikipedia (an encyclopedia), and Biblioteka Nauki (an open-access repository of scholarly publications). The

Dziennik Ustaw source contains Poland’s Journal of Laws and Monitor Polski, two official gazettes. Wolne Lektury is a free online library of literature, ELTeC-pol is the Polish part of the European Literary Text Collection, and 1000 Novels is a collection of literary texts written in or translated into Polish [16]. The remaining Wikimedia sources are Polish Wikiquote (quotations), Wikivoyage (travel guides), Wikibooks (text-books), Wikinews (news articles), and Wiktionary examples (dictionary usage examples). The other sources are gov.pl (central-government pages), Global Voices (civic-media articles), NKJP1M (a one-million-word sample of the National Corpus of Polish [14]), samorzad.gov.pl (local-government articles), and Sejm API speeches (recent speeches from the lower house of the Polish parliament). Polish copyright law excludes normative acts and official documents and materials from copyright protection [13, Art. 4]; this is the basis we record for the parliamentary sources and Dziennik Ustaw. The record URL of Sejm API speeches (Table 3) identifies the sitting rather than the individual speech.

4 Audit of Released Files and Metadata

We recomputed document, character, and proxy-token counts from the 20 Parquet files at the `v0.2.5` tag; every per-source value matches the published source statistics exactly. Verification against the release manifest is described in Data and Code Availability. The licence and author findings below are taken from source cards and from the records of each contribution (pull requests and commits).

Document length and tokens per word. The released files contain 25,266,552,685 characters and 3.54B whitespace-delimited words. The 9,676,513,980 `cl100k_base` proxy tokens correspond to 0.383 tokens per character (2.61 characters per token) and an average of 2.73 tokens per word; by source, tokens per word range from 2.52 (ELTeC-pol) to 3.00 (Wikibooks). English rules of thumb for tokens per word do not transfer to Polish [8]. What counts as one document differs by orders of magnitude: the median Wiktionary example has 40 words, the median NKJP1M sample 55, the median Wikipedia article 114, and the median ELTeC-pol novel 67,007. Legal sources are heavy-tailed: EUR-Lex has a median of 544 words but a 90th percentile of 7,179; Dziennik Ustaw has a median of 969 and a 90th percentile of 12,354. Section 5 compares document and token shares by domain.

Exact duplicates. We hashed each document’s raw text (SHA-256, no normalization) across the 20 released Parquet files. Within sources, 47 documents repeat an earlier text. All are in HPLT and are copies of 23 distinct boilerplate texts, such as recruitment consent clauses and comment-policy notices. Across sources, 10 texts occur twice (20 documents): six Wikipedia articles and three formal parliamentary questions to the government (interpellations) also appear in HPLT, and one NKJP1M sample matches a Wikipedia article. This audit postdates the v0.2.5 tag, so these 57 surplus copies (0.0013% of documents) remain in the release; removing them belongs to the corpus-wide deduplication proposed in Section 7. See Section 8 for the scope of exact matching. Per-source counts

Source	Docs	Tokens (M)	Median words/doc	Domain	Added	Collection licence	Per-record provenance	Record URL
HPLT v3 Polish	1,603,974	2,555.5	346	web/news	v0.2.3/4	CC0 packaging; underlying web rights separate	none	dropped
EUR-Lex	243,060	2,378.1	544	legal	v0.2.0	CC BY 4.0	none	—
Parliamentary corpus [15]	324,622	1,646.8	388	legal	v0.2.0	official records; Art. 4 basis	none	—
Wikisource	632,005	801.9	200	literary	v0.2.0	CC BY-SA; card records 3.0	none	—
Wikipedia	1,171,897	707.2	114	reference	v0.2.0	CC BY-SA; card records 3.0	none	—
Biblioteka Nauki	42,071	673.5	4,252	academic	v0.2.4	mixed CC BY/BY-SA/PD	licence; author, year, title, publisher	—
Dziennik Ustaw	35,442	486.1	969	legal	v0.2.0	official records; Art. 4 basis	none	dropped
Wolne Lektury	6,141	103.0	155	literary	v0.2.0	CC BY-SA 4.0 / Free Art 1.3	none	—
gov.pl	88,190	80.7	254	government	v0.2.2	CC BY-SA 4.0	none	—
1000 Novels [16]	1,000	60.5	5,745	literary	v0.2.1	CC BY 4.0	none	—
						compilation; per-text rights mixed		
samorzad.gov.pl	72,771	41.2	132	government	v0.2.5	CC BY-SA 4.0	publishing institution; sidcar	sidcar
Sejm API speeches	38,812	36.3	207	legal	v0.2.5	official records; Art. 4 basis	speaker; sidcar	sidcar
Wikiquote	30,363	31.9	154	reference	v0.2.0	CC BY-SA; card records 3.0	none	—
ELTeC-pol [17]	100	21.5	67,007	literary	v0.2.0	CC BY 4.0	none	—
Wikivoyage	13,645	17.1	65	reference	v0.2.0	CC BY-SA; card records 3.0	none	—
Wikibooks	9,112	15.6	240	reference	v0.2.0	CC BY-SA; card records 3.0	none	—
Wikinews	24,386	12.1	129	web/news	v0.2.0	CC BY 2.5	none	—
Global Voices	2,040	3.7	624	web/news	v0.2.3	CC BY 3.0	author/translator; sidcar	sidcar
NKJP1M [14]	18,381	2.6	55	web/news	v0.2.3	CC BY	original text ID; author constant “NKJP”	—
Wiktionary examples	8,639	1.1	40	reference	v0.2.5	CC BY-SA; card records 3.0	none; author constant “Wikimedia contributors”	—
Total	4,366,651	9,676.5				20 sources		

Table 3: v0.2.5 source inventory at revision 7ae4ac7, from each source’s card, statistics file, and released Parquet file [10]. Tokens are c1100k_base proxy counts, rounded to 0.1M; calculations use exact source statistics. *Median words/doc*: median document length in whitespace-delimited words, unlike the Docs and Tokens columns, which are source totals; measured on the Parquet files at the tag; 10th and 90th percentiles and further length statistics are in *evidence/length_stats.csv* (Appendix A). *Domain*: the six domains analyzed in Section 5. *Added*: first stable release containing the source. *Collection licence*: source-level rights label, not per-record clearance. CC means Creative Commons; BY requires attribution; SA adds a ShareAlike condition; PD means public domain; CC0 is a public-domain dedication. *Card records 3.0* means the source card names version 3.0 only (Section 4). *Per-record provenance*: record-level fields beyond the source name; *none* means only the source-level licence and a source-wide created range. *Record URL*: *sidcar* means a sidcar file maps each id to the preferred source-page URL; *dropped* means the URL was available when the source was collected but is not kept in the canonical schema; — means no URL.

and examples are in *evidence/exact_duplicates.csv* (Appendix A).

Licence coverage versus licence specificity. The release metadata give a non-empty `license` value for all 4,366,651 rows. For most sources, that value is the source’s collection licence copied to every row, not a licence checked for each document. Biblioteka Nauki is the only source whose documents carry more than one licence value: 28,335 CC BY 4.0, 13,733 CC BY-SA 4.0, and three public-domain records. Those 42,071 rows are only 0.96% of the corpus. Global Voices uses a single CC BY 3.0 value, verified for each article. HPLT’s single CC0 value covers packaging, not underlying pages (Section 6) [5].

Attribution is sparse in the canonical field. The cards of fourteen sources—the twelve in v0.2.1, gov.pl, and HPLT—report that no document has an `author` value: 4,183,937 documents, or 95.82% of v0.2.5. NKJP1M fills `author` for all 18,381 rows, but with the constant corpus label “NKJP” rather than per-document authors; Wiktionary examples does

the same for its 8,639 rows with the constant “Wikimedia contributors”. Three sources have record-level authors in `author`: samorzad.gov.pl (72,771 rows), Sejm API speeches (38,812 rows, 1,149 distinct speakers), and Global Voices (2,040 rows). Together these are 113,623 rows (2.60%). Biblioteka Nauki recovered attribution for its eligible records, but when the source was added in v0.2.4 the attribution was stored in a source-specific `attribution` column rather than the canonical `author` field [10]. This is a schema inconsistency (Section 7), not evidence that attribution is missing at the original publisher. Thus five sources have record-level provenance fields (Table 3), four have record-level attribution including Biblioteka Nauki, and three have authors in the canonical `author` field.

Licence version and scope. Several source cards record CC BY-SA 3.0 for Wikimedia projects. Wikimedia changed the default licence for new edits on many projects to CC BY-SA 4.0 in June 2023, while older edits remain under 3.0 [12]. A single 3.0 label is therefore incomplete for dumps that contain edits made after the change (Section 7). The NKJP1M

card cites CC BY without a version. The registry uses the standardized machine-readable licence identifier (SPDX) for version 4.0, but we report the licence as unversioned CC BY because the original publisher states no version.

Literary collections. Collection-level licence labels do not establish the rights to every individual work. The Wolne Lektury source combines public-domain classics with separately cleared works under CC BY-SA 4.0 or the Free Art License. The 1000 Novels collection is licensed under CC BY 4.0 as a compiled resource, and that licence cannot cover third-party rights in its texts. A per-document screen of the released file (evidence/1000_novels_rights_screen.csv, Appendix A) finds 830 texts whose author and any translator died by 1955 and are therefore in the public domain in Poland [13, Arts. 36 and 39], 10 more marked public domain by Wolne Lektury, 17 still in copyright, and 143 whose translator or author could not be verified. Wolne Lektury publishes all 17 in-copyright texts under CC BY-SA 3.0, so v0.2.5’s single CC BY 4.0 label omits their ShareAlike condition. Table 3 describes only the released Wolne Lektury subset, not every work hosted by Wolne Lektury; the 17 rights findings concern the separately packaged 1000 Novels source.

5 Composition Over Time

We use the six coarse domains listed in Table 3: legal, reference, literary, academic, government, and web/news. The web/news domain also includes NKJP1M, a mixed-genre language sample. See Section 8 for the limits of this grouping.

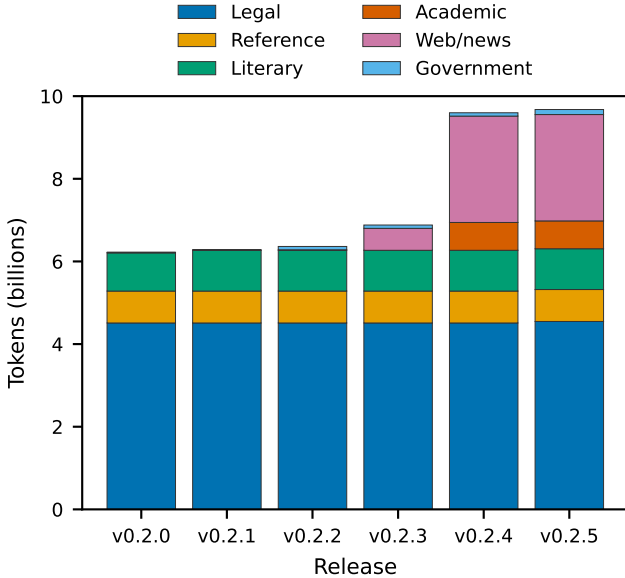


Figure 1: Token composition across reconstructed releases, counted with c1100k_base as proxy tokens. v0.2.3 uses the initial 512.4M-token HPLT contribution; v0.2.4 uses the expanded 2.556B-token source. The legal domain’s token share falls from 72.5% in v0.2.0 to 47.0% in v0.2.5.

To illustrate, we keep every non-legal token and downsample only the legal/parliamentary group until it makes up at most a target share of the resulting mixture (the cap). Table 4 shows the result. At a 15% cap, the corpus retains 6.03B tokens (62.4% of v0.2.5) but only 0.905B legal tokens, 19.9%

Legal cap	Legal kept (B tok)	Total kept (B tok)	Share of v0.2.5 (%)
10%	0.570	5.699	58.9
15%	0.905	6.034	62.4
20%	1.282	6.411	66.3

Table 4: Counterfactual downsampling of only the four legal/parliamentary sources while preserving all other v0.2.5 tokens.

of the original legal material. Temperature-based weighting adjusts how strongly source size affects sampling probabilities [7]. Such weighting or learned mixture optimization [6] can trade diversity against volume. In either case, the training mixture need not mirror the proportions of the corpus.

6 Source Admission Policy

Admission conditions by source type. For sources with a confirmed uniform content licence, such as Wikimedia projects and government publications, collection-level licence evidence suffices. Academic and cultural aggregators can contain differently licensed records: Biblioteka Nauki therefore requires per-record rights evidence before admission. Web sources without a confirmed content licence, such as HPLT, need review beyond the package licence: permissive packaging such as CC0 permits redistribution of the package but does not establish rights to the underlying pages.

Statuses and exclusions. For v0.2.5, the corpus maintainer classifies each candidate source as *admitted*, *held for review*, or *excluded*, using the source’s documentation and rights evidence from the original publisher. OpenSubtitles was excluded because rights to the individual subtitle records could not be established. The raw SpeakLeash [20] redistribution of Europeana’s aggregated European cultural-heritage material was excluded because mixed-rights records lacked reliable per-record rights metadata. A separate direct-ingestion probe preserved rights and creator metadata but was not included in a stable release.

Decision logs. The manuscript package contains a *decision ledger* with one structured record for each of the 20 released sources and the two documented excluded candidates (OpenSubtitles and the raw SpeakLeash Europeana redistribution). Appendix B prints the source-by-source decisions; the accompanying YAML records preserve the evidence inspected, rationale, and release scope in full.

7 Release Engineering and Corrections

The audit identifies corrections for the next tagged release.

Immutable release identity. Future stable releases should ship manifests equivalent to the one described in Data and Code Availability so that changes in release contents remain auditable. The v0.2.5 tag was itself moved twice to correct documentation, without changing any Parquet file (Data and Code Availability). A manifest makes such moves checkable,

but later documentation fixes should go into a new tag, so that a tag names one revision.

1000 Novels. Up to commit `f71cc09`, the `v0.2.5` dataset card listed 1000 Novels, included since `v0.2.1` (Table 3), and also kept it in the table of excluded candidates with a warning that the novels may include in-copyright works. The per-text screen confirms that warning. At `7ae4ac7`, the current target of the tag, the exclusion row is replaced by a rights note in the source’s datasheet that reports the screen’s counts. The Parquet file is unchanged, so every row still carries `CC-BY-4.0`. The next release should relabel the 17 in-copyright texts `CC BY-SA 3.0` with attribution, or remove them, and should withhold the 143 unverified texts until their rights are established (Section 4).

Schema and licence corrections. Biblioteka Nauki’s source-specific `attribution` column should be moved into the canonical `author` field or a sidecar, with the mapping documented. Wikimedia source cards should no longer present `CC BY-SA 3.0` as the sole licence for post-2023 dumps; they need a notice that both versions apply, or per-revision attribution, rather than replacing every `3.0` value with `4.0` [12].

Preserve filter counters. Each source’s build records how many documents each filter removed. The schema-normalization step that prepares a release recomputes statistics from the rows that remain and can reset those counts to zero. Build-time counts should instead be kept in immutable per-source files, and the release recount written separately, so that later audits can distinguish “what the source pipeline dropped” from “what the final Parquet contains.”

Corpus-wide processing. Future releases should assess corpus-wide deduplication and personal-data removal; neither is established by the source-specific processing or exact-match audit (Section 8).

8 Limitations

This resource report contains no downstream model evaluation. Proxy-token counts depend on `cl100k_base`; other tokenizers will yield different totals. Raw-text exact matching in Section 4 is only a lower bound on duplication: copies differing in whitespace, markup, or edition are missed, so no exact overlap between Wikisource and Wolne Lektury does not rule out shared works. Source-level domain groupings in Section 5 are coarse and do not classify individual documents. The legal-cap results are downsampling counterfactuals, not trained or evaluated mixtures; they assume neither replacement nor temperature sampling. The shared legacy builder keeps a document only if Polish diacritics make up at least 0.5% of its alphabetic characters. This deterministic heuristic needs no external model, but it is not a validated language identifier. Independently contributed sources use source-specific checks. Release `v0.2.5` did not undergo a corpus-wide personal-data removal pass.

9 Conclusion

Six reconciled releases illustrate continuous corpus development through source additions, attribution sidecars, and metadata corrections. Release `v0.2.5` contains 4.37M documents and 9.677B proxy tokens, and every change in document and token totals between stable releases can be reconstructed from per-source statistics. Legal/parliamentary text is 47.0% of `v0.2.5` tokens. An explicit 15% legal cap would retain 6.03B tokens, illustrating the difference between the corpus that may be released and the mixture used for training.

Every row has a licence value, but for most sources that value reflects collection-level terms rather than a document-level check; 95.8% of documents come from sources whose cards report no author values. The corrections in Section 7 concern release manifests, per-text rights in 1000 Novels, attribution and Wikimedia licence labels, filter counters, corpus-wide deduplication, and personal-data removal.

Data and Code Availability

The `v0.2.5` release analyzed in this paper is available at <https://huggingface.co/datasets/SlayerLab/polish-dynaword/tree/v0.2.5> (commit `7ae4ac7`). Its annotated dataset tag identifies that revision. The tag has been moved twice: from `02bcb0b` to `f71cc09`, which corrected the card, changelog and source registry, and then to `7ae4ac7`, which added the 1000 Novels rights note (Section 7). Neither move changed a Parquet file; the tag message records both earlier targets. The manifest `evidence/v0_2_5_manifest.csv` records paths, byte sizes, and SHA-256 digests for all 70 files under `data/` at that revision: 20 source Parquet files with their cards, statistics files, build scripts, and three attribution sidecars. Readers can verify downloaded files against these values. The accompanying manuscript package contains the evidence files and scripts listed in Appendix A, and figures. No training text is redistributed with the manuscript.

Author Contributions

Roles cover the `v0.2.5` release and this paper.

Kacper Wikeł Project maintenance, curation, source and licence review, release engineering, documentation, validation, review and merging of contributions; supervision; writing – original draft; writing – review and editing.

Arkadiusz Słota Added the HPLT source (`v0.2.3`) and expanded it (`v0.2.4`), including its cleaning, phone-number removal, duplicate removal within HPLT, validation and documentation.

Dawid Majewski Added the Global Voices (`v0.2.3`) and `samorzad.gov.pl` (`v0.2.5`) sources with per-document attribution files.

Paweł Puzio Added the `gov.pl` source (`v0.2.2`); tools that compute release totals and generate documentation; contributor guidelines; tagging of `v0.2.5` and correction of its dataset card from 18 to 20 released sources, including Sejm API speeches and Wiktionary examples.

Bartosz Kobylński Added the 1000 Novels (`v0.2.1`) and Biblioteka Nauki (`v0.2.4`) sources, including recovery of

per-document licence and attribution; preparation for rebuilding Europeana.

Piotr Styła Added the Sejm API speeches source (v0.2.5).

Katarzyna Majzel-Pośpiech Added the NKJP1M (v0.2.3) and Wiktionary examples (v0.2.5) sources.

Acknowledgments

We thank the Dynaword authors and maintainers, SpeakLeash, the original publishers and public institutions, and the maintainers of the underlying corpora and APIs.

References

- [1] Kenneth Enevoldsen et al. 2026. *Dynaword: From One-shot to Continuously Developed Datasets*. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3782–3793. doi:10.63317/4x9cdkge22vb. <https://aclanthology.org/2026.lrec-1.301/>.
- [2] Danish Foundation Models. 2026. *Dynaword tooling*. <https://github.com/danish-foundation-models/dynaword>.
- [3] Timnit Gebru et al. 2021. *Datasheets for Datasets*. Communications of the ACM 64(12):86–92. doi:10.1145/3458723.
- [4] Emily M. Bender and Batya Friedman. 2018. *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*. TACL 6:587–604. doi:10.1162/tacl_a_00041.
- [5] Stephan Oepen et al. 2026. *HPLT 3.0: Very Large-Scale Multilingual Resources for LLMs and MT. Mono- and Bilingual Data, Multilingual Evaluation, and Pre-Trained Models*. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1409–1434. doi:10.63317/25xbdofco9od. <https://aclanthology.org/2026.lrec-1.110/>.
- [6] Sang Michael Xie et al. 2023. *DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining*. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
- [7] Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual Language Model Pretraining*. In Advances in Neural Information Processing Systems 32 (NeurIPS 2019).
- [8] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. *Language Model Tokenizers Introduce Unfairness Between Languages*. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
- [9] OpenAI. 2022. *tiktoken*. <https://github.com/openai/tiktoken>.
- [10] Polish DynaWord contributors. 2026. *Polish DynaWord dataset card and source documentation, v0.2.5*. Revision 7ae4ac7. <https://huggingface.co/datasets/SlayerLab/polish-dynaword/tree/7ae4ac7>.
- [11] Polish DynaWord contributors. 2026. *Polish DynaWord release history and source statistics*. <https://huggingface.co/datasets/SlayerLab/polish-dynaword/blob/7ae4ac7/CHANGELOG.md>.
- [12] Creative Commons. 2023. *Wikipedia Moves to CC 4.0 Licenses*. 29 June 2023. <https://creativecommons.org/2023/06/29/wikipedia-moves-to-cc-4-0-licenses/>.
- [13] Republic of Poland. 1994. *Act of 4 February 1994 on Copyright and Related Rights*. Consolidated text, Dz.U. 2025 poz. 24. <https://eli.gov.pl/eli/DU/2025/24/og1>.
- [14] Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.). 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.
- [15] Maciej Ogrodniczuk. 2018. *Polish Parliamentary Corpus*. In Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19.
- [16] Maciej Eder et al. 2016. *1000 Novels Corpus*. CLARIN-PL Repository. <http://hdl.handle.net/11321/312>.
- [17] Christof Schöch, Roxana Patras, Tomaz Erjavec, and Diana Santos. 2021. *Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives*. Modern Languages Open 1(25). doi:10.3828/mlo.v0i0.364.
- [18] Shayne Longpre et al. 2024. *A Large-Scale Audit of Dataset Licensing and Attribution in AI*. Nature Machine Intelligence 6(8):975–987. doi:10.1038/s42256-024-00878-8.
- [19] Nikhil Kandpal et al. 2025. *The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text*. In Advances in Neural Information Processing Systems 38 (NeurIPS 2025, Datasets and Benchmarks track). doi:10.52202/085713-1918. https://proceedings.neurips.cc/paper_files/paper/2025/hash/52acc050138d6f40dad6f12f91a4ce22-Abstract-Datasets_and_Benchmarks_Track.html.
- [20] SpeakLeash. n.d. *SpeakLeash (Spichlerz)*. <https://speakleash.org/>.

A Reproduction Notes

The supplementary files are evidence/artifact_audit.csv (exact document, character, and proxy-token counts), evidence/v0_2_5_manifest.csv (release file hashes), evidence/length_stats.csv (length percentiles), evidence/exact_duplicates.csv (duplicate counts and examples), evidence/sources_v0_2_5.csv (source inventory), evidence/release_history.csv (release additions), evidence/release_domain_composition.csv (six-domain matrix and entropy inputs), evidence/legal_cap_counterfactual.csv (10/15/20% mixture calculations), and evidence/1000_novels_rights_screen.csv (per-text rights screen). Source-documentation notes and contributor reconciliation are in evidence/EVIDENCE_LEDGER.md and evidence/CONTRIBUTOR_RECONCILIATION.md. Source decisions live in repo_patch_suggestions/decisions/under_repo_patch_suggestions/decision-record.schema.yaml. The scripts scripts/snapshot_stats.py, scripts/audit_release.py, scripts/analyze_metadata.py, and scripts/novels_rights_screen.py reproduce the snapshot, release, metadata, and novel-rights audits. The snapshot script regenerates the manifest and length statistics from a local dataset clone at the v0.2.5 tag and fails if any recomputed count differs from the artifact audit. The rights-screen script regenerates the per-text rights CSV from the released 1000 Novels Parquet file and the CLARIN-PL file listing, and fails if a term-based classification contradicts a Wolne Lektury licence footer in the text.

For a legal cap c that preserves all non-legal tokens $R = T - L$, the maximum retained total is

$$T_c = \frac{R}{1 - c}, \quad L_c = cT_c. \quad (1)$$

B Source Decision Ledger

This is a readable summary of the 22 structured records in `repo_patch_suggestions/decisions/`, following `repo_patch_suggestions/decision-record.schema.yaml` 1. These are manuscript-package reconstructions proposed for the dataset repository, not records already adopted there. “From” identifies the first stable release containing an admitted source; all admitted sources remain in v0.2.5 (artifact revision 7ae4ac7). Evidence links, decision dates, artifact hashes, and detailed qualifications are in the individual YAML records.

- **10000_novels** — admitted from v0.2.1 on the collection CC BY 4.0 licence. The per-text screen flags 17 CC BY-SA 3.0 PL texts and 143 unresolved texts requiring correction or review before another release.
- **biblioteka_nauki** — admitted from v0.2.4 after per-document licence recovery from the publisher; only CC BY, CC BY-SA, and public-domain records were kept.
- **dziennik_ustaw** — admitted from v0.2.0; Polish official acts rely on the copyright exclusion in Article 4, not a CC licence.
- **eltec_pol** — admitted from v0.2.0 on the collection’s documented CC BY 4.0 licence.
- **eurlex** — admitted from v0.2.0 on documented reuse conditions for EU documents; the card’s CC BY 4.0 label does not establish a per-act CC grant.
- **global_voices** — admitted from v0.2.3; retained articles have CC BY 3.0 evidence and an attribution sidecar with per-record URLs.
- **govpl** — admitted from v0.2.2 on the documented gov.pl text licence; this is not per-record clearance.
- **european_hplt_v3_pl** — admitted from v0.2.3 and expanded in v0.2.4; CC0 covers the dataset package, not necessarily the underlying web pages.
- **nkjp1m** — admitted from v0.2.3 on the documented CC BY corpus licence; the upstream licence does not specify a version.
- **parliamentary** — admitted from v0.2.0; official-document exclusion under Polish law and the corpus redistribution terms are separate grounds.
- **samorzad_gov_pl** — admitted from v0.2.5 on the platform’s CC BY-SA 4.0 text terms; attachments may have separate conditions.
- **sejm_api** — admitted from v0.2.5 on official-document and public-sector reuse grounds, with per-speech attribution and URLs.
- **wikibooks** — admitted from v0.2.0 on documented Wiki-media terms; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wikinews** — admitted from v0.2.0 on the documented CC BY 2.5 content licence.
- **wikipedia** — admitted from v0.2.0 on documented Wiki-media terms; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wikiquote** — admitted from v0.2.0 on documented Wiki-media terms; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wikisource** — admitted from v0.2.0 on documented Wiki-media terms; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wikivoyage** — admitted from v0.2.0 on documented Wiki-media terms; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wiktionary_examples** — admitted in v0.2.5 on the tagged inventory and registry; post-2023 revisions may have mixed CC BY-SA 3.0/4.0 history.
- **wolne_lektury** — admitted from v0.2.0 on documented edition-level CC BY-SA 4.0 / Free Art License terms, distinct from per-text clearance.
- **opensubtitles** — excluded from stable releases: rights to individual subtitle records could not be established.
- **europena_aggregate** — excluded from stable releases: the raw SpeakLeash redistribution lacked reliable per-record rights metadata; the separate direct-ingestion probe was not released.