RNC contexts

We implemented a two-stage pipeline:

1. manual collection of idiom examples from the Russian National Corpus (RNC) parallel subcorpus, followed by programmatic aggregation; and
2. integration of these examples with the curated idiom data from already annotated Dubrovin's English collection

Stage 1: Manual collection and aggregation of RNC idiom examples

Idiom examples were collected manually from the Russian National Corpus (RNC), specifically its parallel Russian–English subcorpus.

Each idiom was queried using two types of searches:

1. Exact search for fixed idiomatic expressions.
2. Lemma + grammatical tag search to capture inflected or syntactically varied realizations of idioms.

For each idiom, search results containing contextual bilingual examples were manually exported directly from the RNC interface into CSV format. Each exported file contained:

- the Russian sentence with the idiom in full context,
- the aligned English translation,
- metadata such as author, title, publication venue, publication year, and linguistic labels,
- the internal RNC example identifier.

Each file was saved manually and named according to the idiom’s unique ID, which established a 1-to-1 correspondence between idioms and example files.

The automated aggregation script subsequently processed these manually downloaded files as follows:

1. The script scanned the folder containing all manually downloaded files and detected all CSV files.
2. Filenames were interpreted as idiom IDs (e.g., *47.csv* → `idiom_id = 47`)
3. Each CSV was loaded and converted into a standardized format. All fields were treated as text to ensure uniformity across files.
4. The script copied key fields from RNC into a normalized schema, including:
    - Russian example (`Full context`),
    - English parallel example (`Para context 1`),
    - author, title, publication information,
    - linguistic annotations (sphere, topic, medium),
    - translation metadata (translator, original and translation languages),
    - and the RNC internal example identifier.
5. Each example received a unique `example_id` ensuring consistent tracking across all idioms.
6. After processing all files, the script concatenated the per-idiom DataFrames into a single aggregate dataset, saved as *english_examples.csv*.

Stage 2: Integration with the existing Dubrovin's collection

To link corpus examples to the main idiom inventory, we merged the aggregated RNC examples with a main English version of Dubrovin's dataset, which contained for each idiom:

- its Russian lemma
- transliteration
- literal meaning
- semantic meaning
- English equivalents
- and transparency ratings

The merging procedure included:

1. Loading both datasets.
2. Ensuring consistent ID types by converting `id` and `idiom_id` to integers.
3. Left-join by idiom ID.
4. Schema cleaning and column selection.
5. Export of the final dataset.

.