

# MagicTagger — Product Brief (v1.3)

26.11.2025

## 1) Executive Summary

A locally deployable (Docker) tool for folklorists that transforms a corpus of Russian tales of magic into a reproducible analytical environment. The application focuses on automatically identifying Top-3 ATU types for input text in Russian and highlighting evidential fragments with a one-expert decision protocol. The system surfaces anchor evidence and mini variant descriptors (tags) within a type, where we have variations in corpus. Built-in visualizations provide geographic mapping and year/decade timelines by type, character name and role frequencies. All results export as Linked Open Data.

The tool supports an external corpus inside a Docker volume, [FEB — Russian Folk Tales](#) by A. N. Afanasyev (comp. L. G. Barag, N. V. Novikov; eds. E. V. Pomerantseva, K. V. Chistov). This mirror is optional and separate from the primary archive corpus. Working mode can use Corpus A (Local Gold from the Estonian Folklore Archive) and, optionally, Corpus B (External Corpus) for retrieval-based suggestions (similar types, plot variations, anchor snippets).

### Global Metrics:

Metric	What we measure	What it validates	What the report will show	Threshold (target / min)
<b>ATUtype — Hit@3 (multi-label)</b>	Share of tales where <b>any</b> gold label appears in the model's <b>Top-3</b> . Ambiguous excluded.	Practical usefulness: the system does not miss the correct type in the short list.	Main Hit@3 on Test; breakdowns by ATU/region/decade; nearest-confusion matrix; <b>auto-adopt</b> share ( $\max\_score \geq 0.75$ ).	<b>0.80 / 0.75</b>
<b>NER — Macro-F1 (PERSON, BEING)</b>	<b>F1</b> for PERSON/BEING at the <b>entity level</b> (exact span + correct label); Macro-F1 = mean over the two person & being.	Readiness of names/beings tables for publication/search without mass manual cleanup; suitability for variant-tags.	Macro-F1; P/R/F1 per class; typical error list (over/under).	$\approx 0.80 / \geq 0.75$

## 2) Users and Jobs-to-be-Done

### Jobs of Russian Folklore Researchers

- **Make a quick first typing.**

When I have a new text, I want a short Top-3 ATU candidate list with anchor quotations and a simple decision protocol, so that I reduce classification time and have a defensible rationale.

- **Show evidence.**

When I fix a decision, I want to view and export anchor citations and decision notes, so that I can justify conclusions to colleagues.

- **Describe within-type differences.**

When I need to show variability, I want variant tags (from contrasts in the texts) with citations and CSV/RDF export, so that differences are concise and citable.

- **Maintain a unified index of characters.**

When I prepare indexes and filters, I want Names & Roles in a clear, publication-ready view (table/simple visuals) with alias normalization and CSV export, so that I don't waste time on manual cleanup.

- **Cross-check with trusted sources offline.**

When I arbitrate contentious texts, I want similar Russian variants from a local external archive with a source badge and citations, so that I speed up arbitration and literature review.

- **Prepare graphs tied to archival metadata.**

When I write a chapter, I want visuals tied to our fields (region/year, etc.), so that I minimize edits.

### Jobs of Researchers of Other Traditions

- **Present similarities/differences carefully.**

When I formulate the comparison, I want variant tags and roles as hypotheses with citations and a small-n badge, so that I honestly show proximity or divergence without overreach.

### Common Jobs

- **Grasp the corpus at a glance.**

When I prepare a report, I want type distribution, a regional map, and year/decade timelines, so that I can describe composition and dynamics.

- **Publish reproducibly.**  
I want a one-click FAIR export, so that I can deliver materials and archive them.
- **Work privately and offline.**  
When data are sensitive or can't be shared, I want local Docker processing with no default storage, so that I comply with project policy.

## 3) Problem Statement and Context

Researchers at the Estonian Folklore Archives and affiliated institutions produce tale-type catalogs: inventories of all tales in a collection with explicit identification of plot kernels and structural variation within each type. The current workflow remains largely manual.

### Pain Points

- **Time-intensive and labor-heavy.** Classifying texts, extracting kernels, and documenting variants across dozens of tales takes months.
- **Limited reproducibility.** Decisions depend on expert memory; transparent evidence chains (anchor quotations, versioned decision rules) are rarely captured.
- **No rapid exploratory analysis.** Teams cannot quickly visualize which types exist, where/when they occur (region, year/decade), or how variant proportions change.
- **Static publication formats.** Results are hard to share as FAIR/LOD data and costly to revise.
- **No ad-hoc validation tools.** It's cumbersome to "sanity-check" an external text's type quickly, compare against known variants, or export reproducible evidence for review.

### 3.1 Research Questions and Expected Contributions

This tool is designed not only as an infrastructure component, but as a basis for empirical research on Russian magic tales in the Estonian Folklore Archives. We focus on these research questions:

- **Archival regimes and their imprint on the corpus**  
How do the repertoires of ATU types, motifs and structural features of Russian magic tales differ across the two fieldwork waves: pre-war ERA expeditions and post-war RKM and TRÜ student expeditions in the 1960s?
- **Borderland repertoires**  
How do magic tales recorded in Russian-speaking villages in Estonia differ from those recorded in Pskov Oblast in terms of ATU repertoires, motifs and character roles? Do the Estonian borderland villages function as a “conservative reservoir” of older plots, or rather as a zone of hybridisation?
- **Time, war and plot dynamics**  
How do the frequencies of ATU types, character configurations (hero/helper/adversary/donor) and the “temperature” of the supernatural (e.g. frightening vs. comic vs. moralising plots) change between the pre-war 1930s, and the post-war and late-Soviet 1960s?

The tool will be evaluated not only by accuracy metrics, but also by its ability to support these questions through maps, timelines, and variant/role summaries.

## 4) User Value Proposition

### For Russian tale researchers:

- **Decide faster with evidence.** Top-3 ATU + narrative anchors with a one-expert decision protocol (auto-adopt for confident cases; ambiguous flow for review).
- **Describe variation.** Mini variant descriptors (tags) within a type, backed by citations; exportable to CSV.
- **See the corpus at a glance.** Regional maps and year/decade timelines tuned to archival metadata.
- **Index characters consistently.** Names & Roles in a publication-ready view (table and/or simple visuals), with alias normalization and CSV export.
- **Be audit-ready and reproducible.** Decision notes, cited evidence, and versioned rules/models recorded in PROV; FAIR exports.

### For researchers of other traditions:

- **Transparent, portable outputs.** FAIR bundles with sources/URLs, timestamps, and model/rule versions; no cross-corpus full-text ingestion (optional local read-only mirror only).

## 5) Feature Roadmap

### 1. Automatic classification by Aarne–Thompson–Uther (ATU) types

When user put in the special form external tale in Russian, the system suggests three most likely ATU types and shows evidence why it thinks so. It also supports cases where one tale has several correct types. If the system is unsure, the tale gets a “needs review” status, but you still see the Top-3 candidates.

For each candidate ATU type, the system computes SCORE (calibrated model confidence, from 0 to 1) and ANCHOR (evidence strength from rule-based motifs, from 0 to 1).

Priority	Label	SCORE condition	ANCHOR condition	Extra condition ( $\Delta = \text{SCORE}_1 - \text{SCORE}_2$ )
1	auto-adopt (primary)	SCORE $\geq 0.75$	ANCHOR $\geq 0.60$	$\Delta \geq 0.15$
2	needs review (A)	SCORE $\geq 0.50$	ANCHOR $\geq 0.40$	—
3	needs review (B)	SCORE $\geq 0.35$	ANCHOR $\geq 0.70$	—
4	weak	otherwise	otherwise	—

### 2. Highlighting evidential fragments (“anchors”)

For each suggested type, the system shows 2–4 short quotations from the tale that serve as evidence.

**How we accept the work:** At least one relevant quotation for each type; if confidence is low, an honest label “indicative, needs review” is shown.

### 3. Difference tagging

Produces tags in English that describe how tales differ within one type (for example, key artifact, different participants etc.).

**How we accept the work:** Tags are visible in the interface and can be exported both as tables and as machine-readable files.

### 4. Comparison with an external corpus

The system shows a list of similar tales with a clear source label (internal or external corpus) and short quotations that explain the similarity.

**How we accept the work:** The block displays a visible “external source” label with the source name (for example, “FEB”); everything works offline; nothing is saved.

## **5. Names and roles**

The system automatically extracts personal names and mythological beings, normalizes spelling variants, and assigns basic roles in English (hero, helper, adversary, donor). Provides a summary view and CSV export.

**How we accept the work:** On a small manual validation set, accuracy target is around 80%, minimum 75%).

## **6. Knowledge graph construction**

Converts the annotated data into subject–predicate–object statements (RDF triples) and exports the resulting graph so it can be loaded into a graph database for queries.

**How we accept the work:** The exported graph validates against the provided shapes/constraints; sample queries return the expected results.

## **7. Map and timeline**

Shows distribution by regions and by years/decades based on your metadata. Includes a switch to hide collector & narrator names if needed.

**How we accept the work:** Aggregations are precomputed.

## **8. Ephemeral processing with no saving by default**

All operations run in memory; nothing is saved automatically. Reports and data are downloaded only when you explicitly ask for it.

## **9. Decision traces and data provenance**

On request, exports a package showing when results were produced, by which versions of rules/models, and which quotations were used.

**How we accept the work:** All provenance fields are present in the package; nothing is stored on any server.

## **10. Export data packages**

Creates a local archive (ZIP) with data in tables and machine-readable formats (including a manifest and a structural validation report).

**How we accept the work:** The package passes structure checks; the archive has a date stamp and a version tag.

## **11. One-off processing of external texts (only in Russian)**

You can paste any text, get Top-3 types with quotations, and download a short report together with the data. All of this runs locally, without saving by default.

## 6) Data Scope and Evaluation Sets

The composition of the primary corpus and all inclusion/exclusion decisions are documented in detail in “`docs/archive_notes/sampling_log.md`” and ``docs/archive_notes/sampling pipeline.jpg``. Those documents describes which archival series were screened, how “tales of magic” (ATU 300–749) were operationally defined, and how the near-census corpus of 117 texts was derived.

Here we only summarise how these curated texts are organised into datasets for training and evaluation.

### **Corpus A — Local Gold (primary corpus from Estonian Folklore Archive)**

- 117 curated Russian tales of magic with ATU labels (multi-label where applicable) and full provenance.
- Split **once** into **Dev A (70%)** and **Test A (30%)** with no collector overlaps or duplicates.
- All dashboards (distribution, regional map, year/decade timeline) are computed only from Corpus A.

### **Corpus B — External Reference (local copy from the [FEB Library](#))**

- Curated local mirror of external edition to increase training diversity (more variants, types, roles).
- Included in training together with Corpus A to improve generalization on external input.
- Never mixed into official metrics; in UI and exports every item from this set carries a source badge (e.g., “Source: FEB”).
- Used at runtime for retrieval-based suggestions (similar tales, anchor snippets) — always flagged by source.

### **Test Set C — Hold-out for official evaluation**

- A separately curated set not included in A or B, taken from other collections.
- Used only for official performance reporting (ATU type classification).

### **Operational rules**

- **Training:** model is trained on **A (Dev) + B**; calibration is performed on **A (Dev)**.

- **Evaluation:** official metrics are reported on **Test A** and **Test C** only; **B** is used for robustness checks and demo cases with manual validation.
- **Runtime behavior:** when suggesting similar tales for an external input, items from B are shown with explicit source labels; collection overviews and all public-facing visuals reflect Corpus A only.

## 7) UI & UX Flows

- **Left navigation rail**
  - **Home** — project overview, entry to core workflows (“Start exploring”, “Try classification”).
  - **Explore**
    - *Overview* — high-level stats and featured maps/charts.
    - *Types* — browse ATU / custom types with links to examples.
    - *Maps & Timeline* — spatial-temporal exploration of tales and collectors.
    - *Names & Roles* — people, their roles, and relationships.
  - **Classify**
    - Paste or upload a text to:
      - check its **type** (ATU suggestions),
      - see **same-type texts** in our corpus,
      - analyze **plot variations within a type** via Variants / Evidence panels.
- **Top bar**
  - **Project name** — clickable logo/title, returns to Home.
  - **Quick Export** — one-click access to per-item JSON-LD, batch Turtle, CSV for current selection or view.
  - **Help** — onboarding tour, documentation, and links to methodology / licenses.

## 8) Privacy & Licensing

- **Code:** AGPL-3.0-only.
- **Derived transcriptions & metadata:** CC BY 4.0.

Until the archive decides on name disclosure, narrators and collectors from roughly 30% of texts (closed part of collection) are anonymised with stable IDs; the name mapping resides at the archive. Users are kindly requested to avoid re-identification and to coordinate any requests for identified data with the archive.

- **Scans:** Estonian Folklore Archive policies.
- **External Corpus B:** optional, read-only, clearly labeled in UI.

The authoritative policy is versioned in `docs/permissions&licensing.md` and bundled with the Docker image; all exports embed *dc:rights* and *dcterms:license* accordingly.

## 9) Risks and Mitigation

### 1. Data sparsity by type

- **Risk:** Only 1–4 texts per type → hard to study within-type variation. Very few types have  $\geq 2$  texts to compare.
- **Mitigation:** Attach an external corpus to observe more variants and train the model on richer data so it handles external texts more reliably.

### 2. Narrow national types (EE/SUS/VRD) without ATU equivalents

- **Risk:** Some texts are assigned with national types can distort the catalog and make texts non-comparable across traditions.
- **Mitigation:** Display alternative IDs (EE/SUS/VRD) and provide the closest ATU type or family. Add a badge and include SKOS mappings in exports (exact/close/broad/narrow/related).

### 3. Ambiguity and multi-label cases

- **Risk:** Disputed texts and multiple valid types per tale may confuse the model.
- **Mitigation:** Use Top-3 + multi-label rule; auto-adopt only when SCORE/ANCHOR/ $\Delta$  criteria are met; otherwise mark needs review (one expert).

### 4. Dialect/orthography/OCR noise

- **Risk:** Search quality drops, anchors miss, names drift.
- **Mitigation:** Cleaning/lemmatization, normalization of spelling variants, dialectal lexicons; a before/after report on search queries.

## **7. Reliability of anchor rules**

- **Risk:** Overfitting to narrow phrasings, many false positives.
- **Mitigation:** Maintain a playbook with anti-triggers, syntactic conditions, and weights; active review of contentious hits; rule versioning in PROV.

## **8. Names and roles**

- **Risk:** NER confuses personal names/mythologemes; roles are assigned too rigidly.
- **Mitigation:** Names & Roles lite (table + charts), alias normalization, treat role as a hypothesis with citation; target Macro-F1  $\approx 0.80$  (min 0.75) on a small validation set.

## **11. Misinterpretation of the external mirror**

- **Risk:** Users may assume Corpus B is part of “our” archive.
- **Mitigation:** Single combined list but A items first; B items carry a clear **source badge**.

## **13. UI scale for ~449 types**

- **Risk:** Users get lost among many types.
- **Mitigation:** A single Types workspace (master–detail), virtualized list, jump-to search and facets.

## **14. Duplicates and republications**

- **Risk:** The same performance appears in multiple editions → inflated frequencies.
- **Mitigation:** Review ATU coverage; exclude external texts that 100% match archive tales (if any are found).

## **18. Versioning and reproducibility**

- **Risk:** Results shift after rule/model edits; hard to reproduce.
- **Mitigation:** Version rules/playbook, models, and configs; export ZIP with PROV and checksums; maintain a changelog.

# **10) Future Developments**

## **1. Cross-Tradition Typing & Roles (EN input)**

User upload English translations from other traditions and gets ATU Top-3 (or “non-magic tale”) with anchors; extract Names & Roles in EN; compare to Russian corpus. Accelerates literature review and searching intercultural connections.

## **2. Tale Similarity Graph (pilot types)**

Interactive graph where nodes = tales, color = variant-tag bundle, edges = signature similarity (anchors + lexical features). Fast visual sense-making of neighborhoods, outliers, and boundary cases inside a type.

## **3. Magic Numbers**

Detection of formulaic counts **3/7/12** with citations (anchor snippets) and per-type summaries. Quick cultural signals for write-ups.

## **4. Batch Classify (Queue Mode)**

Drop a folder of RU tales. Speeds work up.

## **5. Easy Attachment of Third-Party Russian Corpora**

Local, read-only import of external collections inside Docker. External data stay isolated from Archive A metrics (used for retrieval/evidence only). Quickly broadens variant and type coverage for better suggestions and model robustness, while preserving privacy, provenance, and reproducibility without relying on external APIs.