



Junior conference on Data Science and Engineering

Automated Motif Extraction from Folktales Using Large Language Models

JDSE 2025

Saba Shahsavari, Alessa Mayer, and Fabian M. Suchanek

September 24, 2025

Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France
{saba.shahsavari, alessa.mayer}@ip-paris.fr

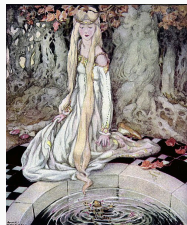
Do you know a tale, where an animal is magically transformed into a human?

Do you know a tale, where an animal is magically transformed into a human?

- Beauty and the Beast
- The Frog Prince



Beauty and the Beast



The Frog Prince

Do you know a tale, where an animal is magically transformed into a human?

- The Beauty and the Beast
- The Frog Prince
- The Legend of the White Snake (Chinese fairy tale)
- The Enchanted Pig (Romanian fairy tale)



The Legend of the White Snake



The Enchanted Pig

- Motifs are recurring narratives in stories.
 - Problem: Manual motif extraction is slow, subjective, and not scalable.
- We propose an LLM-based motif extraction method that can automatically detect motifs in folktales.

- **Motif:** smallest definite element of a tale.
- **Tale type:** recurring, self-sufficient plot or group of motifs.¹

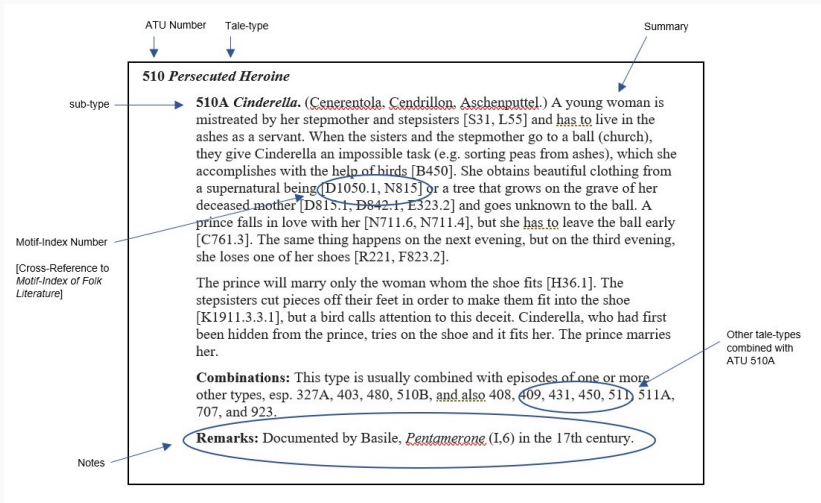
¹Definitions taken from the Harvard Library (guides.library.harvard.edu)

Related Work

- Early collections of folktales: "Contes de ma mère l'Oye" (Perrault, 1697) and "Magasin des enfants" (Leprince de Beaumont, 1798)
- "Motif-Index of Folk-Literature" (Thompson, 1955–1958)
 - Catalogue of narrative elements.
- "Morphology of the Folktale" (Propp, 1968)
 - Focus on tale structures.

- **Aarne-Thompson-Uther Index (ATU Index)**
 - Original index by Antti Aarne (1910), expanded by Stith Thompson (1961) and Hans-Jörg Uther (2011).
 - Classification system central to folkloristics.
 - Groups tales based on their motifs.

Related Work: Folktale classification



ATU Index 510A

- Automatic classification of tale genres (Nguyen et al., 2012).
 - Automatic classification of story types (Nguyen et al., 2013).
 - Detecting tropes in short social media texts (Flaccavento et al. 2025).
- Lack of automated in-depth analysis using motifs of stories

Dataset

- 1 331 folktales translated into English.
- 133 ATU tale types (e.g., Cinderella, Tales of Magic, Animal Helper).
- Average length: 979 tokens per story.

Methodology

Methodology: Pipeline

1. LLM extracts motifs from folktale.
2. Cluster motifs to reduce synonyms.
3. Represent story as motif embeddings.
4. Classify into ATU tale type.

Methodology: LLM extracts motifs from folktale

- Gemini-2-Flash for motif extraction from raw text.
- Prompt: short, general phrases; discourage irrelevant content; slight story-specific details for clarity; motifs given as examples.

- m_i embedding of motif i using Sentence-T5-large embeddings.
- Agglomerative clustering for similar motifs.
- Motif m_i replaced by representative embedding $C(m_i)$ closest to cluster centroid c_k :

$$C(m_i) = c_j, \text{ where } j = \arg \max_k \cos(m_i, c_k).$$

- Reduced 1,222 motifs to 216 clusters.

Methodology: Story as motif embedding

- Each story: mean vector of its motifs.

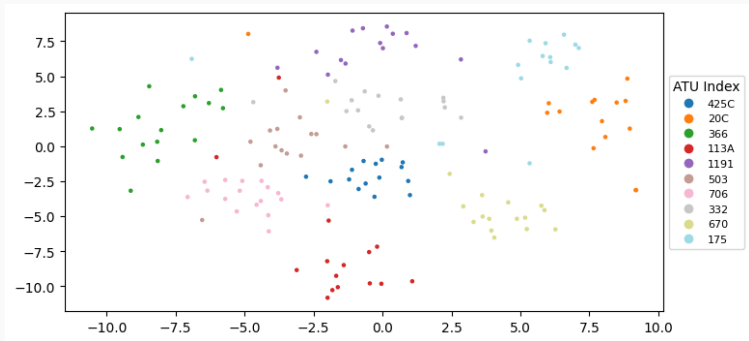


Figure 1: Visualization of stories of 10 ATU Indices. Each dot represents a story (calculated as the average of its motifs), and each color represents a different ATU category.

- Idea: Validate motif extraction via ATU index classification.
- Story v assigned to ATU type k with closest mean embedding t_k (training set):

$$\hat{y} = \arg \max_k \cos(v, t_k).$$

Results

Classification Results

- Motif-based ATU-index classification: **88.6% accuracy**.
 - Baseline (direct story embeddings): 72.7%.
- Accuracy increased by more than 15%.
- Promising motif extraction method.

Findings: Recurring Motifs

- Motifs across different ATU indices.
- e.g.
 - “Deception to trick another” appears in 117 tale types (out of 133).
 - “Animal offers aid” appears in 73 tale types (out of 133).
- Confirms universal storytelling strategies.

Findings: Regional Patterns

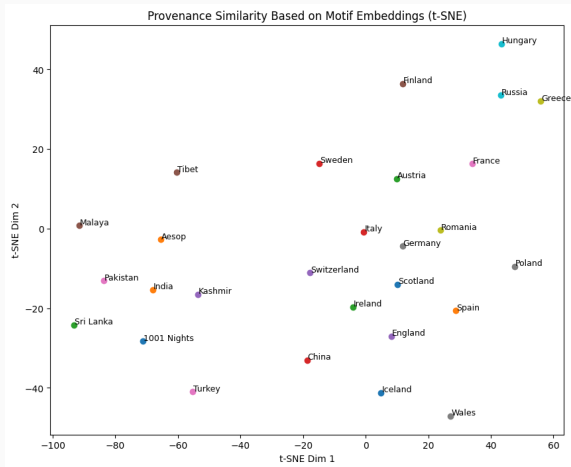


Figure 2: Two-dimensional t-SNE projection. Each point represents a geographical region, based on the mean embedding of motifs from stories attributed to that region.

Findings: Regional Patterns

- Mean motif embeddings computed for each region.
- Close regions share similar motifs.
- Region-specific motifs, e.g. “house as protection” in English tales.
- Aesop’s Fables aligned with Indian tales, not Greek/European.
 - Consistent with scholarly findings.

Conclusion and Future Work

Conclusion

- Scalable and reproducible LLM-based framework for automated motif extraction and analysis from folktales.
- Validated by previous manual folktale analysis.
- High accuracy in tale type classification.
- Extraction of universal motifs.
- Allows to study regional patterns.

- Extend to multilingual datasets.
- Add temporal metadata to study story evolution.
- Applications such as cultural studies.



Thank you!

Any questions?

-  Dundes, A.: The Motif-Index and the Tale Type Index: A Critique. In: Journal of Folklore Research, pp. 195–202 (1997)
-  Aarne, Antti (1910). Verzeichnis der Märchentypen. Vol. 3. FF Communications. Helsinki: Suomalaisen Tiedeakatemian Toimituksia.
-  Uther, H.J. (2011). The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson. FF communications. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.
-  Thompson, Stith (1961). The Types of Folktale: A Classification and Bibliography. 2nd Revision. Helsinki: Academia Scientiarum Fennica.

-  Propp, Vladimir (1968). Morphology of the Folktale. Trans. by Laurence Scott. Austin: University of Texas Press.
-  Perrault, Charles (1697). Histoires ou contes du temps passé, avec des moralités: Contes de ma mère l'Oye. Paris: Barbin.
-  Leprince de Beaumont, Jeanne-Marie (1798). Magasin des enfans, ou Dialogues d'une sage gouvernante avec ses élèves de la première distinction. Vol. 2. Lyon: Veuve Rusand.
-  Nguyen, D., Trieschnigg, D., Theune, M.: Folktale Classification Using Learning to Rank. In: Advances in information retrieval. Proceedings, pp. 195–206 (2013). 10.1007/978-3-642-36973-5_17

-  Nguyen, Dong, Trieschnigg, Dolf, Meder, Theo, and Theune, Mari"et (Sept. 2012). "Automatic classification of folk narrative genres". In: 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012, pp. 378–382.
-  Flaccavento, Alessandra, Peskine, Youri, Papotti, Paolo, Torlone, Riccardo, and Troncy, Raphael (Jan. 2025). "Automated Detection of Tropes In Short Texts". In: Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 5936–5951.
-  Ashliman, D.L.: Folktexts: A Library of Folktales, Folk Legends, Fairy Tales, and Mythology. Accessed: 2025-05-17.
<https://sites.pitt.edu/~dlim1/folktexts.html>

-  Hagedorn, J., Daranyi, S.: Bearing a Bag-of-Tales: An Open Corpus of Annotated Folktales for Reproducible Research. In: Journal of Open Humanities Data 8.16, pp. 1–10 (2022). 10.5334/johd.78.
<https://doi.org/10.5334/johd.78>
-  Berezkin, Y., Duvakin, E.: Buried in a Head: African and Asian Parallels to Aesop's Fable. In: Folklore 127.1, pp. 91–102 (2016).
<http://www.jstor.org/stable/24774365>

Picture "ATU Index 510A": https://guides.library.harvard.edu/folk_and_myth/indices
Other pictures: wikipedia.org

Given the following folktale, extract key motifs as a list of short, general phrases (e.g., 'magical helper aids,' 'journey undertaken,' 'reward earned'). Avoid using synonyms with backslashes (e.g., 'scarcity strikes,' not 'famine/scarcity'). Add slight detail for clarity, but keep phrases general and applicable across stories, avoiding specific names or objects.

Folktale: {story}

Motifs:

Figure 3: The final prompt used in the experiments.