**MALEC / Daranyi — Predication**

<some tabular material set as text>

**Toward Predication-Based Semantic Trawling for Folk Narrative Studies**

Malec, S., Cohen, T., Widdows, D. and Darányi, S.

There is growing interest in intangible cultural heritage for narrative generation (Peinado and Gervás, 2006), yet with no test collections of tales or myths available, the computational modeling of narratives is still in its infancy (Finlayson et al., 2014). On the other hand, folk narratives are a global phenomenon, amply documented by fieldwork. These on the one hand pin down the major tale types as canonical motif sequences (Uther, 2004), so that storyline variants can be grouped under such types as document classes. Also, there exist lists of indexing criteria, which can be used to describe the content of such narratives on different conceptual levels (Thompson, 1955–1958). Therefore, methods that can accommodate nested compositional structure are required to index such material.

At the same time, information societies increasingly exploit digitized content for knowledge discovery, the scalability of data often leading to unforeseen opportunities for data analytics (Virshup et al., 2013). As there are apparent structural similarities between the recombinative transmission mechanisms of hereditary material in biology vs. cultural memories inherent, e.g., in texts, departing from earlier work on narrative genomics (Darányi et al., 2012; Ofek et al., 2013), and by calling in significant theoretical and development progress in analogical information retrieval and inference in the biomedical domain (Cohen et al., 2012; 2014), here we report on a new integrated methodology for narrative analysis. The idea goes back to literature-based discovery (Swanson, 1986; Cohen et al., 2010), and implements Propp’s formalism to analyze Russian fairy tales as representations of sets of concept-relation-concept triplets, or predications, in high-dimensional space.

Predication-based semantic trawling (PBST) is designed to both aggregate material conformant with specific semantic markup in a knowledge base and to improve the robustness of the same knowledge base by a feedback loop. The idea goes back to the metaphor of a fishing net with variable mesh size, implemented as (1) a knowledge base with a learning component; (2) storing semantically marked-up source material as a test collection, where markup regulates content granularity, i.e., the mesh size; (3) using the knowledge base for information filtering (Brin, 1998; Agichtein and Gravano, 2000) by trawling external data; (4) iteratively expanding the test collection by trawling results; and (5) periodically analyzing incoming data with the results fed back to the learning base.

Our new methodology is designed with scalability, quality, and automation in mind. To this end, we consider the respective integration of available technological steps into a single processing workflow as the missing link to get from little to big data in folk narrative analysis (Malec et al., 2014). These technological steps come from a combination of biomedical text analysis with the Proppian analysis of fairy tales.

In biomedical text analysis, Predication-based Semantic Indexing (PSI) provides the means to efficiently search across tens of millions of concept-relation-concept triplets, known as semantic predications, extracted from the biomedical literature using a Natural Language Processing (NLP) system called SemRep (Rindflesch and Fiszman, 2003). SemRep uses the UMLS1 and MetaMap2 to map relevant expressions from free text to concepts in a controlled vocabulary, and extracts relationships between these concepts using underspecified syntactic parsing, a set of indicator rules, and constraints present in the UMLS semantic network. PSI derives high-dimensional vector representations of concepts from the predications they occur in, effectively circumventing the combinatorial explosion of possible pathways between concepts by converting the task of traversing individual predications into the task of measuring the similarity between composite concept vectors. Consequently, search time for single, double, or triple predicate paths is identical once the relevant concept vectors have been constructed. This method can also detect double and triple predicate pathways connecting example pairs of therapeutically related drugs and diseases, and use these inferred pathways to guide search for treatments for other diseases. Further, PSI has been used to mediate semantic search by utilizing high-dimensional vector representations to infer the nature of the relationship between query concepts and other concepts in relevant documents. Inference is accomplished in high-dimensional space using Expansion-by-Analogy, a novel analogical approach to pseudo-relevance feedback, in which the relationships between query concepts and other concepts in documents they occur in guide the query expansion process. The semantic vector–based approaches developed show improvements in performance over a baseline bag-of-concepts model, and these are most pronounced on queries that are not conducive to keyword-based search (Cohen et al., 2014). This approach can be used to create predication-based semantic space for folk narratives.

V. J. Propp’s theory that the canonical form of Russian fairy tales is a compulsory sequence of actions—called *functions* and selected from a list of 31 typical activities performed by typical actors—was based on a limited sample of cca 50 fairy tales from the Afanas’ev collection, itself comprising cca 600 stories, selected and compiled in the 19th century (Propp, 1968). Whereas the in-principle applicability of the scheme, with or without modifications, has been extensively debated ever since, researchers have started to look at the reproduction of Propp’s conclusions only recently (Bod et al., 2012). The scheme lends itself to semantic markup (Malec, 2010; Lendvai et al., 2010), with subject-verb-object (SVO) triples underlying Proppian functions suitable for predicate encoding of tale content, an observation first publicly observed among Western readers by Rumelhart (1975).

The following are typical examples of predication from the biomedical informatics domain:

*Concept\_1 Relation Concept\_2*

Mammalian Oviducts LOCATION\_OF Sterility

Thymol turbidity test DIAGNOSES Disease

Epididymis LOCATION\_OF Obstruction

In comparison, predicates based on Russian fairy tales look like the following:

*Concept\_1 Relation Concept\_2*

Baba Yaga IS\_A donor

Golden apple IS\_A gift

Baba Yaga LIVES\_IN hut on chicken legs

Donor GIVES gift (direct object)

Donor GIVES\_TO protagonist (indirect object)

In the Proppian sample corpus, e.g., Baba Yaga gives a magic apple to Ivan Simpleton, who therefore is the protagonist,3 exemplifying the ‘donor’ function. The following tale segment helps to identify this function (Afanas’ev 96: ‘Morozko/Jack Frost’):

*The poor little thing remained there shivering and softly repeating her prayers.*

*Jack Frost came leaping and jumping and casting glances at the lovely maiden.*

*‘Maiden, maiden, I am Jack Frost the Ruby-nosed!’ he said.*

*‘Welcome, Jack Frost! God must have sent you to save my sinful soul.’*

*Jack Frost was about to crack her body and freeze her to death, but he was touched by her wise words, pitied her, and tossed her a fur coat.*4

Here the ‘maiden’, a stepdaughter, is ordered by her stepmother to be left out to the elements in the forest, a plot element representative of Cinderella-like tales (ATU Type 510A) called ‘mat’ padcheritsa’ or ‘Zolushka’ tales in Russian folklore tradition. Jack Frost is clearly the donor. The verb ‘tossed’ is mapped to the ‘GIVES\_TO’ relation for its predicate**.**

Steps in the processing workflow are as follows:

• Apply PoS tagging,5 anaphora resolution,6 and semantic role labeling with SemLink7 (Palmer et al., 2010).

• Normalize concepts and predicates and identify triplets (Rusu et al., 2007).

\* Decompose tales sentence by sentence, noting context with respect to Proppian function sequence.

\* Shallow parsing of texts using Stanford parser or OpenNLP to identify SVO.

\* Extract triples.

• Situate these normalized concept and predicate triplets as metadata within the marked-up corpus as the Proppian SemRep (PSR) output.

\* Through a kind of hermeneutic circle, create an analogy of the UMLS in biomedical domain, except with domain knowledge particular to folk narratives, e.g., a controlled vocabulary for predicates (GIVES\_TO, ACCEPTS, PUNISHES, etc.), concepts (hero).

\* Index predications into PSI-space and index corpus with permuted random indexing (PRI), then visualize results by Epiphanet (Schvaneveldt, 1990) and the query/d3.js interface.

\* Extract triples from the Morphology itself to create a knowledge base for purposes of reasoning about the Russian fairy tale domain.

The PSR tool allows for the sequential encoding of marked-up text by circular holographic reduced representation (De Vine and Bruza, 2010), the binary spatter code (Kanerva, 1996), or PRI (Sahlgren et al., 2008) to map them in predication space. For this, different open-source software components are available, prominently the SemanticVectors package.8 The output of PSR enables the analogical retrieval of predicates, finding missing pieces of evidence to reason about dramatic personae and the plot logic of magic tales.

The core of the tale set for the initial knowledge base are a subset of more than 45 tales in English translation to reflect the subset of Afanas’ev’s tales in Propp’s schema in appendix III. The second sweep will cover a larger subset of the Afanas’ev collection, namely that which Propp himself stated at the outset as his concern9 before the semantic trawling of web resources may commence.

We have presented here a very rough model and a proof of concept. There will likely be room to address some aspects of the complex issue of the dynamic learning of narrative macro-structures, which tend to be distinctive of particular traditions and populations (Colby, 1973). Finally, our contribution ought to be seen in the light as ‘bootstrapping’ an artificial system to facilitate analogical reasoning about a limited subset of themes, structures, and encoding devices of traditional narrative.

**Notes**

1. http://www.nlm.nih.gov/research/umls/.

2. http://metamap.nlm.nih.gov/.

3. This is a special case where we need an indirect object, Ivan Simpleton.

4. https://github.com/kingfish777/ProppModel/blob/master/000\_096\_Morozko.xml.

5. http://nlp.stanford.edu/software/lex-parser.shtml.

6. http://cswww.essex.ac.uk/Research/nle/GuiTAR/gtarNew.html.

7. http://verbs.colorado.edu/semlink/. Semlink can create mappings between PropBank, VerbNet, WordNet, and FrameNet.

8. https://code.google.com/p/semanticvectors/.

9. Tales #93-270 from Afanas’ev, inclusive of the tales already mentioned.

**References**

**Agichtein, E. and Gravano, L.** (2000). Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*. ACM, pp. 85–94.

**Bod, R., Fisseni, B., Kurji, A. and Löwe, B.** (2012). Objectivity and Reproducibility of Proppian Narrative Annotations. In Finlayson, M. (ed.), *Third Workshop on Computational Models of Narrative*, pp. 17–21, http://narrative.csail.mit.edu/cmn12/proceedings.pdf.

**Brin, S.** (1998). Extracting Patterns and Relations from the World-Wide Web. In *Proceedings of the 1998 International Workshop on the Web and Databases (WebDB ’98),* Valencia, Spain, 27–28 March 1998.

**Cohen, T., Schvaneveldt, R. and Widdows, D.** (2010) Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. *Journal of Biomedical Informatics,* **43**(2): 240–56.

**Cohen, T., Widdows, D. and Rindflesch, T. C.** (2014). Expansion-by-Analogy: A Vector Symbolic Approach to Semantic Search. In *2014 International Conference on Quantum Interaction (QI).* Springer, Berlin.

**Cohen, T., Widdows, D., Schvaneveldt, R. W., and Rindflesch, T. C.** (2012). Discovery at a Distance: Farther Journeys in Predication Space. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference*. New York: IEEE Press, pp. 218–25.

**Colby, B. N.** (1973). A Partial Grammar of Eskimo Folktales. *American Anthropologist,* **75**: 645–62.

**Darányi, S., Wittek, P. and Forró, L.** (2012). [Toward Sequencing ‘Narrative DNA’: Tale Types, Motif Strings and Memetic Pathways](http://bada.hb.se/handle/2320/11830). *Proceedings of CMN-12*, Istanbul, 26–27 May 2012, pp. 2–10, http://narrative.csail.mit.edu/ws12/proceedings.pdf.

**De Vine, L. and Bruza, P.** (2010). Semantic Oscillations: Encoding Context and Structure in Complex Valued Holographic Vectors. In *AAAI Fall Symposium: Quantum Informatics for Cognitive, Social, and Semantic Processes*, pp. 48–55.

**Finlayson, M. A., Meister, J. C. and Bruneau, E. G.** (eds). (2014). *Proceedings of 5th Workshop on Computational Models of Narrative*, 31 July–2 August 2014, Quebec City, Canada.

**Kanerva, P. (**1996). Binary Spatter-Coding of Ordered K-Tuples. *Artificial Neural Networks—ICANN,* **96**: 869–73.

**Lendvai, P., Declerck, T., Darányi, S. and Malec, S.** (2010). Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales. In *Conference for Digital Humanities 2010*, King’s College, London, 7–10 July 2010.

**Malec, S.** (2010). AutoPropp: Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales. In Darányi, S. and Lendvai, P. (eds), *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, University of Szeged, pp. 65–74.

**Malec, S., Darányi, S., Cohen, T. and Widdows, D.** (2014). Closing the Methodological Gap: From Little to Big Data in Folk Narrative Studies. Paper read at the *First International Workshop of Big Data in Cultural Heritage, International Conference on Cultural Heritage EUROMED-14*, Lemessos, Cyprus, 3–8 November 2014.

**Ofek, N., Darányi, S. and Rokach, L.** (2013). Linking Motif Sequences to Tale Types by Machine Learning. In Finlayson, M. A., Fisseni, B., Löwe, B. and Meister, J. C. (eds)., *Workshop on Computational Models of Narrative 2013,* OpenAccess Series in Informatics, Schloss Dagstuhl—Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, pp. 166–82.

**Palmer, M., Gildea, D. and Xue, N.** (2010). Semantic Role Labeling. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

**Peinado, F. and Gervás, P.** (2006). Evaluation of Automatic Generation of Basic Stories. *New Generation Computing,* **24**: 289–302.

**Propp, V. J.** (1968). *Morphology of the Folktale.* University of Texas Press, Austin.

**Rindflesch, T. C. and Fiszman, M.** (2003). The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. *Journal of Biomedical Informatics,* **36**(6) (December): 462–77.

**Rumelhart, D.** (1975). Notes on a Schema for Stories in Representation and Understanding. In Bobrow, D. G. and Collins, A. (eds), *Studies in Cognitive Science*. New York: Academic Press.

**Rusu, D., Dali, L., Fortuna, B., Grobelnik, M. and Mladenic, D.** (2007). Triplet Extraction from Sentences. In *10th International Multiconference on Information Society-IS,* pp. 8–12.

**Sahlgren, M., Holst, M. and Kanerva, P.** (2008). Permutations as a Means to Encode Order in Word Space. In *30th Annual Meeting of the Cognitive Science Society*, Washington, DC, 23–26 July 2008.

**Schvaneveldt, R. W.** (ed.). (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization.* Ablex, Norwood, NJ.

**Swanson, D. R.** (1986). Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine,* **30**(1): 7–18.

**Thompson, S.** (1955–1958). *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Medieval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends.* Indiana University Press, Bloomington.

**Uther, H. J.** (2004). *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson.* Academia Scientiarum Fennica, Helsinki.

**Virshup, A. M., Contreras-Garcia, J., Wipf, P., Yang, W. and Beretan, D. N.** (2013). Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *Journal of the American Chemical Society,* **135**: 7296–303.