Learning a Better Motif Index: Toward Automated Motif Extraction*

W. Victor H. Yarlott[1] and Mark A. Finlayson[2]

1 **Florida International University, School of Computing & Information Sciences, Miami, FL, USA**
wvyar@cs.fiu.edu

2 **Florida International University, School of Computing & Information Sciences, Miami, FL, USA**
markaf@fiu.edu

Abstract

Motifs are distinctive recurring elements found in folklore, and are used by folklorists to categorize and find tales across cultures and track the genetic relationships of tales over time. Motifs have significance beyond folklore as communicative devices found in news, literature, press releases, and propaganda that concisely imply a large constellation of culturally-relevant information. Until now, folklorists have only extracted motifs from narratives manually, and the conceptual structure of motifs has not been formally laid out. In this short paper we propose that it is possible to automate the extraction of both existing and new motifs from narratives using supervised learning techniques and thereby possible to learn a computational model of how folklorists determine motifs. Automatic extraction would enable the construction of a truly comprehensive motif index, which does not yet exist, as well as the automatic detection of motifs in cultural materials, opening up a new world of narrative information for analysis by anyone interested in narrative and culture. We outline an experimental design, and report on our efforts to produce a structured form of Thompson's motif index, as well as a development annotation of motifs in a small collection of Russian folklore. We propose several initial computational, supervised approaches, and describe several possible metrics of success. We describe lessons learned and difficulties encountered so far, and outline our plan going forward.

1998 ACM Subject Classification I.2.7 Natural Language Processing, J.5 Arts and Humanities

Keywords and phrases Text analysis, automated feature extraction, folklore, narrative, Russian folktales

Digital Object Identifier 10.4230/OASIcs.CMN.2016.7

1 Motifs as a Source Cultural Information

Motifs are distinct, recurring narrative elements found in folklore and, more generally, cultural materials. Motifs are interesting because they provide a compact source of cultural information: many motifs concisely communicate a related constellation of cultural ideas, associations, and assumptions. For example, "troll under a bridge" is an example of a motif common in the west. To members of many western cultures, this combination entails a number of related ideas that are by no means directly communicated by the surface meaning

* This research was made possible by an FIU's Presidential Fellowship and FIU's SCIS's Director's Fellowship, both awarded to W. Victor H. Yarlott. This work was also partially supported by National Institutes of Health (NIH) grant number 5R01GM105033-02.

7th Workshop on Computational Models of Narrative (CMN 2016).
Editors: Ben Miller, Antonio Lieto, Rémi Ronfard, Stephen G. Ware, and Mark A. Finlayson; Article No. 7; pp. 7:1–7:10

Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of the words: the bridge is along the critical path of the hero, and he must cross to achieve his goal; the troll often lives under the bridge, crawling out to waylay innocent passers-by; the troll charges a toll or exacts some other payment for crossing the bridge; the troll is a squatter, not the 'officially' sanctioned master of the bridge; the troll enforces his illegitimate claim through threat of physical violence; and the hero often ends up battling (and defeating) the troll instead of paying the toll.

Because of this density of information, motifs are often retained as a tale is passed between cultures and down generations, and folklorists have observed that a tale's specific composition of motifs can be used to trace the tale's lineage [35, Part 4, Chapter V]. This has led folklorists to construct motif indices that identify motifs and note their presence in specific tales (usually as represented in a particular folkloristic collection). The most well-known motif index is the Thompson motif index (TMI) [34], which references tales from over 614 collections, indexed to 46,248 motifs and sub-motifs, 41,796 of which have references to tales or tale types. In Thompson's index each motif is given a designating code; for example, "troll under a bridge" is referenced by the codes G304 and G475.2. In this case, "troll under a bridge" is represented by two motifs as Thompson generalizes trolls to ogres, a general class of monstrous beings; thus, the motifs are "troll as ogre" (G304) and "ogre attacks intruders on bridge" (G475.2).

While Thompson's index is the best known, there are many other motif indices targeting specific cultures and periods, for example, early Irish literature [8], traditional Polynesian narratives [27], or Japanese folk-literature [21]. In addition, the idea of motif was incorporated into another useful notion, the *tale type*, which seeks to classify whole tales based on their motifs. Antti Aarne constructed an index of tale types in 1910 [1], with translations and revisions by Thompson [34] and Uther [37] (the last being known as the *ATU* catalog).

Thompson informally defines a motif as items "worthy of note because of something out of the ordinary, something of sufficiently striking character to become a part of tradition, oral or literary. Commonplace experiences, such as eating and sleeping, are not traditional in this sense. But they may become so by having attached to them something remarkable or worthy of remembering" [34, p. 19]. He notes that motifs generally fall into one of three subcategories: an event, a character, or a prop [35, pp. 415–416]. Here we give an example of each (with their associated Thompson's motif code):

A **hero rescuing a princess** (B11.11.4) is perhaps one of the most well-known event motifs in western culture. Ask a westerner the following question: "A princess has been kidnapped: who kidnapped her, who rescues her, and what does the rescuer need to do to effect a rescue?", and common answers will be "a dragon kidnapped her, the knight must rescue her, and he must kill the dragon." This motif may be the climax of the story, with a "happily ever after" ending just after the hero defeats the dragon, or it may just happen in the course of a story: in *Ivan Dogson and the White Polyanin*, a Russian folktale [2, Tale #139], Ivan slays three dragons, each with more heads than the last, rescuing a princess each time. The motif is prolific, found across the tales, literature, and movies of many cultures.

Old Man Coyote (A177.1) is a character motif: known in some Native American Indian tribes as Coyote, he is one of the most recognizable gods. In Native American Crow folklore, Old Man Coyote creates the earth and all the creatures on earth. He travels the world, teaching the animals how they should behave. Old Man Coyote, however, is far from a noble and elegant creator. He creates ridiculous costumes and tries to trick the Crow tribe into wearing them, only to be run off. He purposefully bungles rituals to produce food, such as transforming skin from his back to meat, in order to guilt his guests into performing the ritual correctly to get free food, later performing it correctly to discredit his former guests

when they tell others he erred. Anywhere Old Man Coyote is referenced, he calls to mind someone who has done great things, but is lazy and often far too clever for their own good, falling prey to their own cunning.

A **magic carpet** (D1155) is a prop that allows the hero to fly through the sky, and is familiar to anyone who has watched Disney's *Aladdin*. In *One Thousand and One Nights*, Prince Husain encounters a merchant selling a carpet for an outrageous price; the merchant says: "O my lord, thinkest thou I price this carpet at too high a value? ... Whoever sitteth on this carpet and willeth in thought to be taken up and set down upon other site will, in the twinkling of an eye, be borne thither, be that place nearhand or distant many a day's journey and difficult to reach" [6, p. 496]. Solomon, said to be the third king of Israel, was said to have a carpet 60 miles on each side that could transport him vast distances in a short amount of time. In Russian hero tales, magic carpets are common items that aid the hero in his quest.

While the above examples are drawn from folklore, motifs have importance beyond folktales: they occur in common tales, news stories, press releases, propaganda, novels, movies, plays, and anywhere that cultural materials are found. A powerful recent example is the use of the *Pharaoh* motif in modern middle eastern discourse. The Pharaoh appears in Qur'an, and comes into conflict with Moses and his attempts to free the Hebrews from Egyptian slavery. The Pharaoh is an arrogant and obstinate tyrant who defies the will of God and is punished for it. In modern Islamist extremist narratives, the Pharaoh is a symbol of struggles against anti-Islamic regimes and has been invoked against leaders such as Anwar Sadat of Egypt, Ariel Sharon of Israel, and George W. Bush, whom Osama bin Laden referred to as the "pharaoh of the century" [19].

Because of their prominence and ubiquity, the ability to automatically identify and extract motifs would open up a vast repository of important cultural information to computational analysis. Currently, motifs must be extracted by hand by trained cultural experts, and any indices manually constructed, with all the attendant delay, error, and incompleteness. Motifs are rarely identified explicitly in new stories, due to size of the indices and the amount of work involved: developing automated motif extraction would allow extensive, exhaustive identification of motifs across textual cultural materials, and allow us to apply all the power of statistical machine learning and related techniques on this new wealth of information.

In this work, we outline a supervised approach to solving this problem. We use supervised techniques in service of our goal of learning a model of how folklorists create and understand motifs: the ubiquity of motifs suggests that there may be some interesting cognitive processes at work and in modeling them we may get closer to understanding these underpinnings.

2 Motif Extraction: Problems

There are a number of barriers to overcome in automatically extracting motifs.

First, current motif indices are **incomplete** and **inconsistent**. Many interesting narrative elements are not listed in the existing motif indices, and the motifs that are listed are not identified at a uniform level of abstraction. This means that extracting motifs from even well-studied materials (such as folklore) is not just a matter of looking for motifs listed in the index. Rather, the problem is two-fold: motif-like elements must be identified within a text and it must be determined whether they represent an existing motif, a specialization or variant of an existing motif, an entirely new motif, or a spurious false positive.

Second, there is a **lack of data**. No one, to our knowledge, has undertaken even the most basic annotation of text with motifs. Therefore we have no data to which machine

learning techniques can be immediately applied to quickly make progress. As is well known, generating manually annotated data is a time-consuming and labor-intensive process [20], making it difficult to learn what a motif is from the text of the narrative itself.

Third, and even more fundamentally, motifs are **ill-defined**. There is no formal definition of motifs and current definitions fall short of the specificity needed for computational work. Further, folklorists have not laid out the principles behind motif identification, nor do we understand the cognitive principles which would drive people to naturally identify motif-like information. We believe, however, that there are some underlying principles, as motifs are not only transmitted culture to culture, but often arise independently between cultures.

2.1 Defining Motifs

In this project we seek to overcome these barriers to demonstrate that automated motif identification, extraction, and annotation is feasible. The first step in this process is to tighten up our definition of motifs, ideally creating a formal model which describes exactly what a motif is. While we do not present a formal model here, we lay the groundwork by identifying the features of motifs such a model would need to address.

Thompson defined motifs as *something remarkable or out of the ordinary*: eating is not a motif, but eating from a magical table is. Even so, Thompson described his analysis as selecting anything he felt was of interest to future scholars, which suggests a somewhat less principled and more intuition-driven approach. From Thompson's discussion on motifs, a concise version of Thompson's definition might be: a motif is any remarkable or non-commonplace element in a story. We get a definition of "element" from *The Folktale*, where Thompson defines the classes most motifs fall into as actors, items, and single incidents [35, pp. 415–416]. Within this paper, we refer to these elements as characters, props, and events, respectively.

This simple first attempt at a definition has some problems. What does *remarkable or non-commonplace* mean? Practically, it means that the element that comprises a motif is not an unremarkable, everyday narrative element, such as eating or an ordinary table. In folklore, such commonplace elements are often excluded (or not consistently retained) as they are not interesting enough to be retained over generations of retellings. Motifs are maintained across many variations of the same tale because they carry culturally relevant, interesting information. Further, even if a commonplace element is inserted into a particular telling, these elements are not likely to show up consistently across tales with the same tale type, suggesting that it would be possible to smooth out remaining commonplace elements from a selection of tales.

Second, the definition does not address the appropriate level of abstraction for motifs. On the one hand, many motifs as listed in existing motif indices are highly specific: for example, a runner who keeps his leg tied up to prevent himself from running away (F681.1), as opposed to the individual presence of a *runner*, *tying a leg*, or *prevention of running away*. This suggests that motifs should tend toward more specific forms that are repeated across tales: for example, "eating oneself up" (F1035) would be preferred over "eating."

On the other hand, motifs often have closely-related variants that lead to the creation of more abstract entries. For example, in the Russian folktale *Bukhtan Bukhtanovich* [18, p. 168], a fox tricks the Tsar into thinking Bukhtan is very wealthy by pretending to have to measure Bukhtan's money using a large bucket. In Thompson's motif index, there is only an entry about a *cat* using such a trick (K1954.1). In this particular case, while we could imagine creating a new motif specific to the *fox*, other examples suggest a preference for generalizing the existing motif or grouping the motifs together under a category like *animal*

uses the measuring trick. Examples of both methods can be found in Thompson's index: *kindness rewarded* (Q40) is a single motif entry with ten sources cited and no submotifs, but *conception from eating animal* (T511.5) has four motifs as children (T511.5.1–T511.5.4).

One important note is that motifs are not necessarily constitutive elements – that is, the presence of absence of a motif is not definitional for the identity of a particular tale. Motifs, rather, impose a "family resemblance" relationship between different version of the same tale. For example, in the well-known tale *Cinderella* [7], found across many different cultures, several motifs commonly recur across retellings: three evil step-sisters, a fairy godmother, a glass slipper, and so forth. But the story will continue to be recognizable as *Cinderella* if the pumpkin carriage (F861.4.3 – *Carriage from pumpkin.*) is replaced by another means of transportation or does not appear at all. A story having all the motifs of other tales of the same type is sufficient, but not necessary, for it to be recognized as a member of that tale type. Fisseni and Lawrence [17] have shown results where, in some cases, modifying the motifs involved may result in a story very similar to the original in what they refer to as a "simple solution to the problem of integrating the proposed change" (p. 103). Ignoring these non-constitutive motifs smooths over details that may potentially contain cultural information and, thus, is not in our interests.

Jason [23] makes an effort to more clearly define motifs, leveling similar complaints on the clarity of Thompson's definition of motifs to those in this paper. Jason provides a definition of motifs as narrative elements that meet the following criteria: they must be (1) the simplest unit of content that fill a primary formal slot of literary structure (a character or deed) and (2) context-free (not belonging to a certain plot). There are issues with this definition. Jason does not appear to define what simplest means beyond filling a slot of literary structure. Restricting motifs to characters or deeds ignores the importance of props within a story, such as magic carpets (D1155). And context-free motifs ignore the vast wealth of cultural knowledge that motifs contain: to encapsulate cultural knowledge, motifs necessarily arise from related tales (a tale type) within a culture.

To address the concerns raised by other definitions, we propose our own definition based on Thompson's original definition: "A motif is a set of closely-related variants of a non-commonplace, specific narrative element that is repeated across tales of the same type." In future work we intend to specify the components of this definition more formally; below, in discussions of our experimental procedure, we indicate how we might do this.

3 Experimental Design & Pilot Work

3.1 Goals

We have two general goals for our experimental work. While we do not achieve either of these goals in this paper, we do make substantive progress toward them, identifying key data, revealing hard problems, and sketching implementations of solutions.

The first goal is to develop a system to extract motifs from narrative text. Automatic extraction is the ability of a system to identify and extract motif-like elements from a raw-text document. We do not expect that the system be capable of assigning a proper, descriptive name for each motif, but it should be capable of grouping of individual occurrences of motifs together (that is, clustering tokens into types), as well as clustering motifs by topic.

The second, longer-term goal is to learn a model of what folklorists think a motif is: that is, how folklorists define motifs, how they determine what elements of a story comprise a motif, and how they extract motifs from narrative texts. We would expect our model to be capable of examining a narrative and identify the same motifs that folklorists identify.

Importantly, the motifs identified by our system in the first goal, and the motifs identified by folklorists in the second goal, will not necessarily be the same. As with all people, folklorists are fallible, prone to errors of commission, omission, and inconsistency. We will seek to expose the basic principles used by folklorists to uncover motifs, apply those principles uniformly, and show how the computational approach can add value to the manual approaches of folklorists by correcting errors, filling gaps, and enforcing consistency.

3.2 Experimental Procedure

We outline here an experimental procedure using supervised techniques that could be used to accomplish our first goal. While we have not implemented this whole procedure, we have made concrete progress on a number of steps as discussed in later subsections.

1. **Input of Raw Narrative Text.** Texts containing narratives (folklore or other cultural materials) are input to the system.
2. **Initial Processing via NLP Pipeline.** The texts are processed by a natural language processing (NLP) pipeline that performs common analyses such as tokenization, lemmatization, part of speech tagging, chunking, syntactic parsing, word sense disambiguation, latent semantic analysis, semantic role labeling, and event detection [15, 16, 24].
3. **Grouping by Term Distribution.** Using term frequency–inverse document frequency (tf–idf)[32, 33], the system will identify the most important terms in each narrative document. The system will sort the narratives into rough similarity groups based on these terms and annotate each document with the group to which it belongs. Another option for this step is to do topic modeling [4, 28], to cluster texts into groups by topic distribution. This step enables the system to smooth out commonplace events, as described in step 6.
4. **Candidate Identification.** For each text, the system will identify spans of text that could potentially be motifs. Spans of text will be identified that meet the general criteria of a motif, in that they are a narrative "element", often indicated by an event or a nominal representing a character or prop. Another strategy would be to look for common, important terms using tf-idf and then attempting to expand a window around these terms. This would allow the system to prefer *three-headed dragon* over *three-headed* or *dragon* if *three-headed dragon* appeared in multiple tales.
5. **Candidate Classification.** Then, within each group identified above, the system will assess the commonality of each identified span, classifying them as motif or not using either a rule-based system or a machine classifier trained on annotated data. A rule based approach, for example, could look for cut-off points in the tf–idf score distribution for a group. A machine learning approach could use features learned from texts manually annotated with the presence of motifs.
6. **Commonplace Event Elimination.** At this stage, candidate motifs can be compared across groups. When a candidate motif appears across multiple tale groups, it is biased against, as these are more likely to be commonplace events, such as eating or sleeping. Candidate motifs that fall below a threshold (either hard-coded or learned) will be eliminated. The remaining motifs candidates are graduated to identified motifs.
7. **Motif Alignment.** The system attempts to group motifs together into variant groups: groups of closely-related motifs that would be subtypes of a single motif in a motif index using semantic role labeling and the relations catalogued in WordNet [36]. For example, if *cat uses the measuring trick* and *fox uses the measuring trick* were both identified, they would be grouped together as a single variant. Narratives are annotated with both the specific motif and with the variant group that these motifs belong to.

8. **Evaluation.** We will evaluate the performance of the system by comparison with manually annotated motif data, using agreement statistics such as the F_1 measure or Fleiss' kappa. Throughout the project we will consult with folklorists to check our work. Specifically, we plan to consult with them after developing a model, when annotating stories, when determining unique motifs, and as we continue to develop the system. This is a crucial step in achieving our second goal of understanding how folklorists identify and extract motifs from narratives and enabling an automated system to do the same.

3.3 Structured Parsing of Thompson's Index

Several keys steps of the experimental design (e.g., steps 4, 5, and 8) require annotated data in the form of a motif index applied to actual text. Therefore we have been working on generating a structured, electronic version of the most comprehensive motif index available, Thompson's motif index [34]. An electronic form of such a resource would allow us to map motifs to the individual stories that contain them, and provide a starting point for manual annotation. Such a resource would allow us, for example, to identify a dense subnetwork of motifs and narratives that would be used as a pilot corpus for testing and training. Thompson's motif index is one of the most widely-used sources for motif information and having it in a structured, easily queryable form will be a great benefit to the community at large.

We have uncovered numerous challenges in this apparently simple task: first, there is no high-quality digitized version of Thompson's motif index. One commonly cited online source, hosted at `Ruthenia.ru` [31], a joint effort between Moscow-based publisher OGI and the Department of Russian Literature at Tartu University to provide sources for Russian language research, has inconsistent HTML and numerous OCR errors that makes parsing of the index difficult. The MOMFER effort to parse the motif index with the intention of creating a search engine [26], provides code for parsing the HTML motif index hosted at Ruthenia.ru, but is incomplete and does not accurately parse large parts of the index.

Even if we had a pristine digitization of Thompon's index, the text suffers from inconsistent formatting, abbreviations, and typographical conventions throughout. Delimiting bibliographic sources is inconsistent: semi-colons are usually a delimiter, but Thompson also uses commas and periods to delimit motifs. This is particularly troublesome, as commas are used to separate multiple references within a single collection ("Grimm Nos. 3, 35, 81 ...") and periods are used to abbreviate and end entries.

Through this effort, other issues with Thompson's motif index have come to light: many of his references to "tales" are simply cross-references to other collections (such as Cross' index of Irish literature [8] or Boberg's index of Icelandic literature [5], among many others). Thus the index does not provide in many cases a direct link between motifs and tales: many stories are cited for only a single motif, despite containing more. Many of the cited stories and collections are hard to find or may no longer be accessible. Due to these issues, the motif index will likely not provide a solution to the initial problem we identified: the need for a corpus with many related motifs.

Despite all the challenges and issues with Thompson's motif index, we are developing a sequence of regular expressions to handle parsing of the index into structured form.

3.4 Preliminary Analysis

Even with a motif index, comprehensive or not, to train and test a motif extraction system we would need the motifs annotated on actual text. To that end, we have completed an initial

development annotation of the motifs in fourteen tales[1] using Thompson's motif index as a resource. Thompson's motif and tale type index contain information taken from Andreev's tale-type index for Russian tales [3, 22], making Thompson's index a suitable reference. The tales we annotated are English translations of Aleksandr Afanase'ev's collection of Russian folklore [2], chosen in part because of the large body of work already related to them due to their prominence in Vladimir Propp's Morphology of the Folktale [30].

This preliminary annotation serves to inform us as we develop an annotation scheme for motifs, a necessary step in developing a corpus of stories. In the near future, we plan to consult a folklorist regarding our motif analysis, formalize our annotation scheme, and fully annotate a substantive set of stories Story Workbench, a narrative annotation tool [15, 16]. These stories will form a gold standard pilot corpus for motif annotation.

4 Related Work

On the folklore side, there are many motif indices, with the Aarne-Thompson Motif-index of Folk-literature [34] being the primary resource. There are numerous other indices, most primarily focused on a specific culture. Thompson also has substantial discussion on motifs and the compilation of indices in his book *The Folktale* [35]. While Thompson's motif index is perhaps the primary source of motif information used today, it has been criticized because of overlapping motif subcategories, censorship (primarily of obscenity), and missing motifs [14].

Additionally, much work has been done identifying tale-types: recent work by Hans-Jörg Uther expands and improves upon the Aarne-Thompson tale classification system, resulting in the ATU classification system [37].

Darányi [9] has called attention to the need for research into the automation of extraction and annotation of motifs in folklore. Further work by Darányi and Forro [10] has determined that motifs may not be the highest level of abstraction in narrative, Darányi *et al.* [11] have made substantial headway towards using motifs as sequences of "narrative DNA", and Ofek, *et al.* [29] have demonstrated learning tale types based on these sequences. Declerck *et al.* [13] have also done work on converting electronic representations of TMI and ATU to a format that enables multilingual, content-level indexing of folktale texts, building upon past work [12]. Currently, this work appears to be focused on the descriptions of motifs and tale types, without reference to the stories.

With regard to analyzing motif annotation schemes, Karsdorp *et al.* [25] present an analysis of the degree of abstraction present in the ATU catalog and the methods used to note what motifs belong to a given tale type. They find the ATU annotation insufficient for analyzing recurring motifs across types, in that it the ATU scheme fails to capture commonalities across closely related types.

5 Contributions

The information content and ubiquity of motifs makes them an important consideration for anyone working with culturally-influenced narratives. In this paper we have motivated a deeper computational look at motifs, and outlined an experimental plan for developing

[1] The fourteen tales analyzed were: Bukhtan Bukhtanovich (#163); Kozma Quickrich (#164); Shabarsha the Laborer (#151); The Serpent and the Gypsy (#149); Burenushka, the Little Red Cow (#101); Wee Little Havroshechka (#100); Ivan Popyalov (#135); Ivan the Bull's Son (#137); Ivan the Peasant's Son and the Thumb-sized Man (#138); The Flying Ship (#144); Ivan the Cow's Son (#136); Nikita the Tanner (#148); The Magic Swan Geese(#113); The Crystal Mountain (#162).

a system that automatically extracts motifs from text. Importantly we have identified numerous barriers: the lack of annotated data, the incompleteness of existing motif indices, the need for a formal model of motifs, and the lack of clarity into the cognitive processes that lead to the generation and identification of motifs in narrative.

We have already made progress toward addressing these problems. We identified several features of motifs that will form the foundation of a formal, computational model, a necessary step towards motif extraction; we have made significant progress in understanding how to parse Thompson's motif index into a structured resource; we have analyzed fourteen stories for motifs, which has revealed several important questions with regard to how motifs should be appropriately reliably annotated.

Motifs are key features in culturally-relevant narratives, and we seek to enable access to this vast resource and open up a new dimension in computational narrative analysis.

References

1. Antti Amatus Aarne. *Verzeichnis der Märchentypen*. Suomalainen tiedeakatemia, 1910.
2. Aleksandr Nikolaevich Afanas'ev. *Narodnye Russkie Skazki*. Moscow: Gos. Izd-vo Khudozh Lit-ry., 1957.
3. Nikolaĭ Petrovich Andreev, Antti Aarne, and Heda Jason. *Index of Tale-plots According to the System of Aarne*. Rand Corporation, 1968.
4. David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
5. Inger Margrethe Boberg. *Motif-index of early Icelandic literature*. Munksgaard, 1966.
6. Richard Francis Burton. *The Arabian nights*. Barnes & Noble, 2009.
7. Marian Roalfe Cox. *Cinderella: Three hundred and Forty-Five Variants of Cinderella, Catskin, and Cap o'Rushes*, volume 31. Folklore Society, 1893.
8. Tom Peete Cross. *Motif-index of early Irish literature*. Indiana University, 1952.
9. Sándor Darányi. Examples of Formulaity in Narratives and Scientific Communication. In *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, pages 29–35, 2010. URL: `http://ilk.uvt.nl/amicus/amicus_ws2010_proceedings.html`.
10. Sándor Darányi and László Forró. Detecting Multiple Motif Co-occurrences in the Aarne-Thompson-Uther Tale Type Catalog: A Preliminary Survey. *Anales de Documentación*, 15(1), 2012. URL: `http://revistas.um.es/analesdoc/article/view/analesdoc.15.1.134691/131801`.
11. Sándor Darányi, Peter Wittek, and László Forró. Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways. In Mark A. Finlayson, editor, *Third Workshop on Computational Models of Narrative (CMN)*, pages 2–10, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).
12. Thierry Declerck and Piroska Lendvai. Linguistic and semantic representation of the thompson's motif-index of folk-literature. In *Research and Advanced Technology for Digital Libraries*, pages 151–158. Springer, 2011.
13. Thierry Declerck, Piroska Lendvai, and Sándor Darányi. Multilingual and Semantic Extension of Folk Tale Categories. In *Proceedings of the 2012 Digital Humanities Conference (DH 2012)*, 2012. URL: `http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/multilingual-and-semantic-extension-of-folk-tale-catalogues/`.
14. Alan Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, pages 195–202, 1997.
15. Mark A. Finlayson. The story workbench: An extensible semi-automatic text annotation tool. In *Intelligent Narrative Technologies*, 2011.

16 Mark Alan Finlayson. Collecting semantics in the wild: The story workbench. In *AAAI Fall Symposium: Naturally-Inspired Artificial Intelligence*, pages 46–53, 2008.

17 Bernhard Fisseni and Faith Lawrence. A Paradigm for Eliciting Story Variation. In *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 100–105. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.

18 Norbert Guterman. *Russian Fairy Tales*. Pantheon Books, 1973.

19 Jeffry R. Halverson, Steven R. Corman, and H. L. Goodall Jr. *Master narratives of Islamist extremism*. Palgrave Macmillan, 2011.

20 Nancy Ide and James Pustejovsky, editors. *Handbook of Linguistic Annotation*. Springer, 2016. Forthcoming.

21 Hiroko Ikeda. *A type and motif index of Japanese folk-literature*. Orient Cultural Service, 1971.

22 Heda Jason. *NP Andreev,'Index of Tale-Plots According to the System of Aarne': A Partial Translation*. Rand Corporation, 1968.

23 Heda Jason. About 'motifs','motives','motuses','-etic/s','-emic/s', and 'allo/s-', and how they fit together. an experiment in definitions and in terminology. *Fabula*, 48(1-2):85–99, 2007.

24 Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

25 FB Karsdorp, P Kranenburg, Theo Meder, Dolf Trieschnigg, and A Bosch. In search of an appropriate abstraction level for motif annotations. In *Proceedings of the 2012 Workshop on Computational Models of Narrative*, 2012.

26 Folgert Karsdorp, Marten van der Meulen, Theo Meder, and Antal van den Bosch. Momfer: A search engine of thompson's motif-index of folk literature. *Folklore*, 126(1):37–52, 2015.

27 Bacil F. Kirtley. *A motif-index of traditional Polynesian narratives*. University of Hawai'i Press, 1971.

28 Jon D. Mcauliffe and David M. Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

29 Nir Ofek, Sándor Darányi, and Lior Rokach. Linking Motif Sequences with Tale Types by Machine Learning. In Mark A. Finlayson, Bernhard Fisseni, Benedikt Löwe, and Jan Christoph Meister, editors, *Proceedings of the 4th Workshop on Computational Models of Narrative (CMN'13)*, volume 32, pages 166–182, Hamburg, Germany, 2013. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/OASIcs.CMN.2013.166.

30 Vladimir Propp. *Morphology of the Folktale*, volume 9. University of Texas Press, 1968.

31 Ruthenia. S. Thompson. Motif-index of folk-literature. http://www.ruthenia.ru/folklore/thompson/. Accessed: 2016-03-09.

32 Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

33 Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

34 Stith Thompson. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books and local legends*, volume 4. Indiana University Press, 1960.

35 Stith Thompson. *The folktale*. University of California Press, 1977.

36 Princeton University. About wordnet, 2010. Retrieved on May 9, 2016 from: http://wordnet.princeton.edu.

37 Hans-Jörg Uther. *The types of international folktales: a classification and bibliography, based on the system of Antti Aarne and Stith Thompson*. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, 2004.