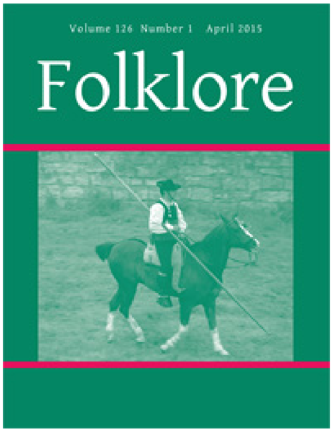


Folklore
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rfol20

MOMFER: A Search Engine of Thompson's Motif-Index of Folk Literature

Folgert Karsdorp, Marten van der Meulen, Theo Meder & Antal van den Bosch

Published online: 20 Apr 2015.

Folklore 126 (April 2015): 37–52
http://dx.doi.org/10.1080/0015587X.2015.1006954

RESEARCH ARTICLE

MOMFER: A Search Engine of Thompson's *Motif-Index of Folk Literature*[1]

Folgert Karsdorp, Marten van der Meulen, Theo Meder & Antal van den Bosch

Abstract

More than fifty years after the first edition of Thompson's seminal *Motif-Index of Folk Literature*, we present an online search engine tailored to fully disclose the index digitally. This search engine, called MOMFER, greatly enhances the searchability of the *Motif-Index* and provides exciting new ways to explore the collection. This is enabled by our use of modern techniques from both natural language processing and information retrieval. The key feature of the search tool is the way in which it allows users to search the *Motif-Index* for semantic concepts, such as 'mythical animals', 'mortality', or 'emotions'. This paper will explain the motivations for creating the search tool, explicate the production process, and show in a number of case studies how the search tool can be used to explore the index in innovative ways.

Introduction

In the introduction to his seminal *Motif-Index of Folk Literature* (TMI), Stith Thompson addressed the challenges of folklore research. The crucial problem of 'the need for a comprehensive classification of the materials in all kinds of traditional narrative' (Thompson 1955–58, 8) became all the more urgent in light of the ever-growing amount of tales that he observed were being collected. Furthermore, he felt that the then current type index by Antti Aarne (1928) was not sufficient for indexing folktales from outside Europe. Thompson took it upon himself to try to remedy this situation by classifying folktales on the basis of motifs, which led to the first edition of the TMI in the 1930s (Thompson 1932–36), and to a revised and enlarged second edition in the 1950s that nearly doubled its scope (Thompson 1955–58).

In folktale research, motifs are a key concept in the classification of folktales into tale types. In the authoritative folktale type catalogue, *The Types of International Folktales* (Uther 2004)—a project initiated by Aarne in 1910 in his *Verzeichnis der Märchentypen* (Catalogue of tale types) and translated and revised by Thompson (Aarne 1910, 1928, 1961)—motifs form the primary descriptive units, and their configuration defines the folktale type of a story. There is, however, a large body of criticism of the TMI as well as the different editions of the folktale type index, such as, for example, Dundes (1997) in relation to Aarne (1961), and Karsdorp et al. (2012) with respect to Uther (2004). This criticism has focused on the distribution and overlap of tale types and motifs. Also, Hasan El-Shamy (1980) addresses the lacunae in the collection of sources used. Despite the various problems with respect to the applicability of the Aarne/Thompson folktale index to non-European tales, El-Shamy states that 'these can be adequately addressed through its [the folktale index's] relatively open classification system, which allows for adding new items, particularly when compared to other systems' (1988, 158). Additionally, Alan Dundes states:

> It must be said at the outset that the six-volume *Motif-Index of Folk-Literature* and the Aarne-Thompson tale type index constitute two of the most valuable tools in the professional folklorist's arsenal of aids for analysis. This is so regardless of any legitimate criticisms of these two remarkable indices, the use of which serves to distinguish scholarly studies of folk narrative from those carried out by a host of amateurs and dilettantes. The identification of folk narratives through motif and/or tale type numbers has become an international *sine qua non* among bona fide folklorists. (Dundes 1997, 195)

While many new motif indices (for example, Baughman 1966; Ikeda 1971; Kirtley 1977; Childers 1977; Flowers 1980; Neuland 1981; El-Shamy 1995), as well as many new collections of folktales (for example, El-Shamy 1980; Meder 2000; Slone 2001; Crooke and Chaube 2002)[2], have since been put forward, relatively few attempts have been made to combine the two (although see, in recent years, Ben-Amos 2006; Slone 2001). We hypothesize that the lack of navigability of the TMI plays an important role in this. Also, while many of the researchers explicitly integrate their new indices into the TMI (for a brief and recent example, see Bell 2009), no complete and combined edition of all these indices is in existence. Because of the problems that printing a book of such magnitude would generate, it is understandable that such an edition is unavailable. Such considerations would evaporate for a digital edition, but such a work, however valuable a research tool it would be, does not exist at present. Given the TMI's unquestioned status as *the* source for the worldwide coverage of folk narrative, we believe that a search engine of the TMI would be of interest to many scholars, ranging from folklorists to narratologists and story generation researchers.

In this paper we highlight some of the navigational challenges one has to overcome in order to successfully make use of existing motif indices, taking the TMI as a case study. We agree with Harriet Goldberg that '[t]he utility of an index of folk-motifs clearly lies in its ability to present in an orderly framework those transitory flashes of recognition that we experience upon hearing a familiar story or a familiar narrative component in a new context' (Goldberg 1998, xiii).

To meet this objective, we present Meertens Online Motif FindER (MOMFER), a free online search engine designed to increase the accessibility of the TMI.[3] This tool provides various new and unique access points to the index, taking advantage of insights from techniques from the fields of natural language processing and information retrieval.[4] MOMFER enables several different search options: next to basic options such as single word, multi-word, and phrase searches, it is also possible to use faceted search, and, most interestingly, semantic search. We hope that these options will provide new tools for researchers and that they will lead to the asking and answering of questions that were hitherto unthought of or which were deemed unanswerable. Also, we hope that this tool will be used as a framework into which all other existing motif indices will be integrated.

The rest of the paper is structured as follows. First, we will briefly introduce the TMI to those readers who may be unfamiliar with it and discuss some of the important navigational challenges faced when using it. Next, we will describe the creation process and architecture of MOMFER. We will then proceed to highlight some of the more interesting search options by presenting several case studies, using the TMI as our corpus. Finally, we will present ideas on how MOMFER can serve as the first step towards an all-inclusive folklore motif-index.

Accessibility of the Motif-Index of Folk Literature

The TMI (Thompson 1955–58) contains over forty-five thousand motifs spread out over five volumes. The motifs are hierarchically ordered in a tree structure. There are twenty-three top-level, alphabetically labelled categories, whose contents range from mythological motifs (Category A) to motifs concerning traits of character (Category W) and beyond. Each top-level category is divided into various subcategories: for example, Category D (*Magic*) is divided into, inter alia, *Transformation* (D0–D699) and *Magic Objects* (D800–D1699), which in turn branch out into child motifs, and so on and so forth. At the leaves of the tree, or terminal nodes, we find the most specific instances of a particular motif (for an elaboration on the build-up of the index and search strategies, see Thompson 1955–58, Volume 1, 19–25; El-Shamy 1995, 16–17). Figure 1 is an exemplary representation of part of the hierarchical tree.

As Thompson states himself, the motifs in the first five volumes are grouped together as the result 'of a gradual evolution, not of any preconceived plan' (Thompson 1955–58, 19). This method (or lack thereof) has several potential negative effects. First, as can be seen from Figure 2, the distribution of the motifs over the main categories in the index is rather uneven, with categories A–K containing seventy-eight per cent of all motifs in the index. Second, the index lists seventy-two motifs twice under different headwords (i.e. 144 unique motifs). For example, the motif *Transformation: utensil to person* has been indexed both as D436.1 and D434.1. Other examples include *One day and one night: object borrowed for a day and a night retained* indexed under K2314.1 and K232.2, and *Well shines at*

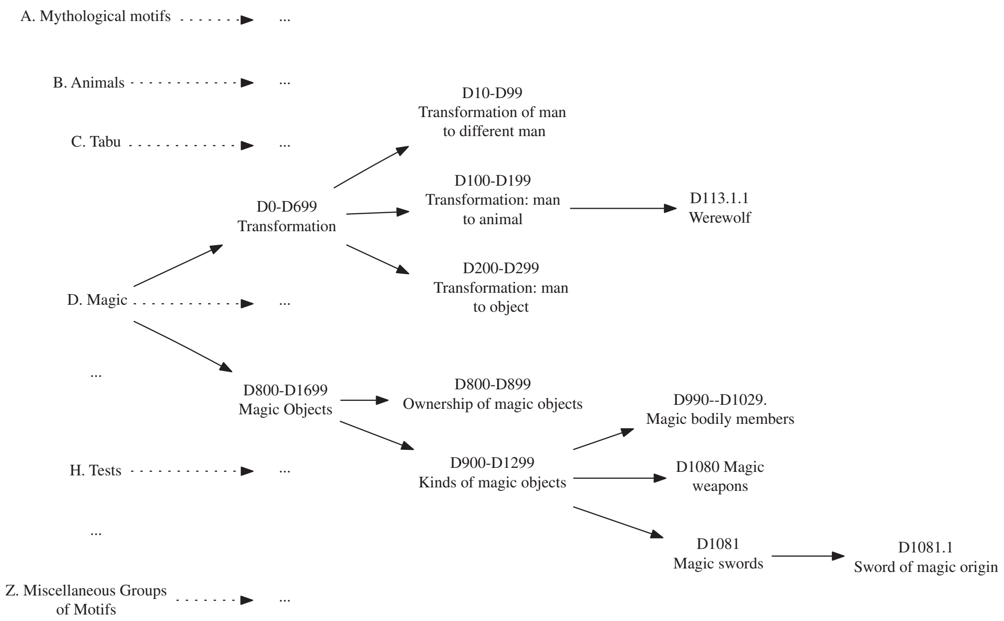

A. Mythological motifs
...
B. Animals
...
C. Tabu
...
D. Magic
...
...
H. Tests
...
...
Z. Miscellaneous Groups of Motifs
...
D0-D699 Transformation
D800-D1699 Magic Objects
D10-D99 Transformation of man to different man
D100-D199 Transformation: man to animal
D200-D299 Transformation: man to object
D800-D899 Ownership of magic objects
D900-D1299 Kinds of magic objects
D113.1.1 Werewolf
D990--D1029. Magic bodily members
D1080 Magic weapons
D1081 Magic swords
D1081.1 Sword of magic origin

Figure 1. Partial view of Thompson's *Motif-Index of Folk Literature*.

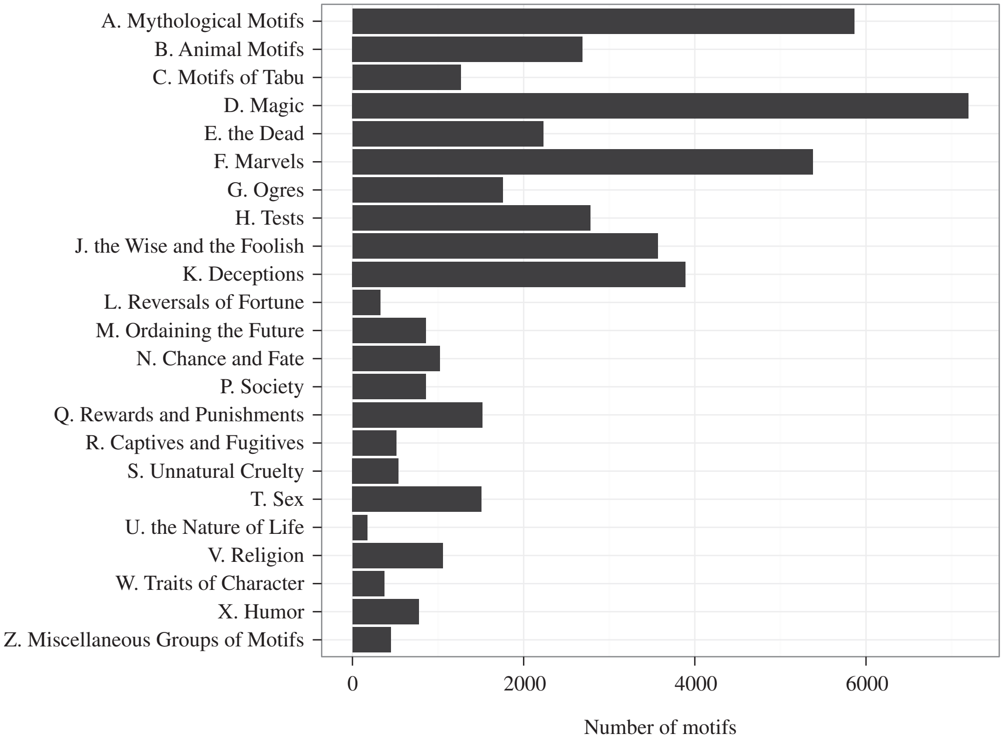


Figure 2. Distribution of motifs over main branches of the *Motif-Index of Folk Literature*.

night, which is categorized at two completely separate branches of the tree: *Magic* (D1645.9) and *Marvels* (F718.5). These cases represent the extreme endpoint on the scale of a more general problem that, as observed by Dundes (1997), many motifs from different branches show a high degree of semantic similarity.

Volume Six of the TMI consists of an alphabetical index to the other five volumes. This index provides a different entrance point to the motif index using an alphabetically sorted list of important terms that appear in the other five volumes. Looking under the term *mice*, for example, we find motifs such as the following (where the headword is left out):

- [mice] army saves kingdom from invasion K632.1;
- [mice] consecrate bishop (lie) X1226.1;
- [mice] and hogs let loose put elephant cavalry to flight K2351.3.

The examples are ordered on different levels. The first level is an alphabetical ordering of the second word of the motif, where the headword is the first word. For instance, as the example above shows, a motif starting with *army* comes before one starting with *consecrate*. However, as can be seen in the same example, this structure is not completely consistent: Thompson ignored function words such as *and*, *with*, and *of*, so that after *consecrate* comes *and hogs*.

After this first set of motifs, a second set of motifs is listed that do not start with the headword but do contain it:

- army of m. B268.6;
- bargain with king of m. M244.1.

While this approach may at first glance seem quite unproblematic, it soon becomes troublesome when entries have a larger number of associated motifs. For example, the word *horse* has over five pages with motifs listed (including *horse's* and *horses*). Here it quickly becomes cumbersome to read through all the motifs in the hopes of finding the appropriate one. Also, several stratagems have been undertaken by Thompson to save space: these include referring readers to a certain category (as for headword *devil*: '*G303ff. See, in addition to the following, the extensive list of motives assembled at G303') and listing certain sets of motifs together without details pertaining to the content of the motifs (as for headword *cat*: 'and witches D1766.4, G211.1.7, G224.11.12, G241.1.4, G225.3, G243.2.2, *G252, G262.1.1, G262.3.2'). Thompson left room for new motifs —although his reasons for doing so were in some cases, such as the erotic and scatological dimensions, suspect (Legman 1962)—and this room has been gratefully filled up by subsequent scholars. New motif indices are made compatible with the TMI by integrating newfound motifs into the hierarchical structure of the TMI. For example, Ernest W. Baughman (1966) posits the motif *Transformation: man meets and copulates with female snake* as D191.2, as an addition to Thompson's D191: *Transformation: man to serpent (snake)*. While this shows that a complete motif index is (at least theoretically) possible, the compilation of such an index far exceeds printing possibilities. Even Thompson himself was aware of this, and left out certain parts of folklore because 'To have included these would have doubled the size of the index' (1955–58, 11), and an online version integrating all subsequent motif indices has as yet not been undertaken. Nevertheless, such a complete work is something to be desired, as it would provide a more comprehensive starting point for comparing different narrative traditions, which was, ultimately, the goal of the TMI (Thompson 1955–58, 9–10).

Today, with most of our digital search results only a few clicks away, the time-consuming labour of manually searching through the TMI, hoping to stumble upon the entry we are looking for, appears rather outdated. In response to this, several initiatives have been undertaken to make a version of the TMI available in digital format, both offline and online.[5] The first example of this kind, as far as we are aware, is the CD-ROM edition of the TMI as published by Indiana University Press in 1993 (Thompson 1993). While this version allows users to enjoy some of the benefits of digital searches (e.g. Boolean searches), it is not readily available, and it is safe to say that it is outdated, since its system requirements are 'DOS 2.0 or higher' (Smith 1994).[6] By presenting the TMI in a digital format, users are able to search through the index using the search facilities provided by their web browser (such as the omnipresent 'Find' function). Although this seems to be quite an improvement over searching through the paper index, in reality the improvements are small: queries remain limited to single expressions and/or phrases. There is one online search engine available at www.storysearch.symbolicstudies.org. Unfortunately, this search engine is limited in its applications: searches seem to be confined to one word, and documentation on other options is unavailable. Also, it does not seem to be intended as a search tool, more as a source of inspiration for storytellers.

In conclusion, the benefits of the currently available digital and online versions are rather small in comparison with the paper version.

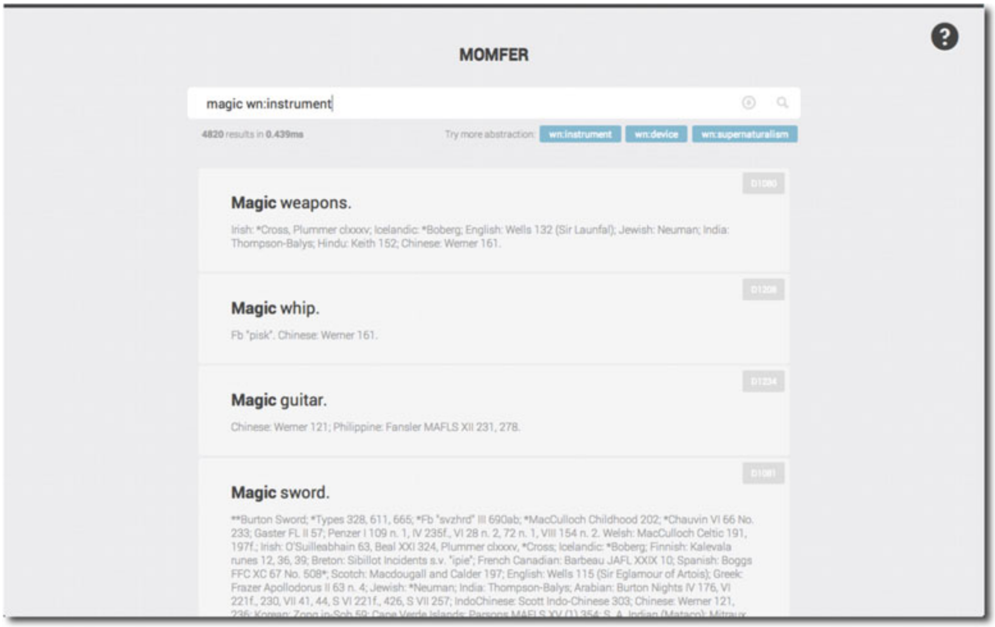
The Motif Index Search Tool

In this section we will introduce MOMFER. First, we will describe the architecture of the search tool, including the sources we used and some of the preprocessing steps we performed. We then continue with an in-depth description of the most salient search features implemented in the tool. Figure 3 is a representation of the web interface of MOMFER.

Preprocessing Steps

As mentioned above, there are a few digitized versions of the TMI in existence. For this paper we constructed a full-text version of the motif index using the edition made available online at http://www.ruthenia.ru/folklore/thompson/. This digital edition is based on the revised and enlarged edition of the TMI (Thompson 1955–58). We subjected this edition to a range of preprocessing steps, which we will describe in order. After stripping all of the HTML tags from the digital index, we used the Stanford CoreNLP toolkit (version 3.3.1) (Toutanova et al. 2003) to tokenize all motifs into formally bounded sentences. We used the same toolkit to lemmatize and tag all words by part of speech (nouns, verbs, etc.) and to apply named entity recognition.

Although earlier writers have classified the motifs in the TMI as having a tripartite structure (notably El-Shamy 1995), for the purposes of this paper we interpret each motif in the TMI as consisting of four parts: a unique key, denoting the place in the hierarchy



MOMFER

magic wn:instrument

4820 results in 0.439ms

Try more abstraction:

D1080

Magic weapons.

Irish: *Cross, Plummer clxxxv; Icelandic: *Boberg; English: Wells 132 (Sir Launfal); Jewish: Neuman; India: Thompson-Balys; Hindu: Keith 152; Chinese: Werner 161.

D1208

Magic whip.

Fb "pisk". Chinese: Werner 161.

D1234

Magic guitar.

Chinese: Werner 121; Philippine: Fansler MAFLS XII 231, 278.

D1081

Magic sword.

**Burton Sword; *Types 328, 611, 665; *Fb "svzhrd" III 690ab; *MacCulloch Childhood 202; *Chauvin VI 66 No. 233; Gaster FL II 57; Penzer I 109 n. 1, IV 235f., VI 28 n. 2, 72 n. 1, VIII 154 n. 2. Welsh: MacCulloch Celtic 191, 197f.; Irish: O'Suilleabhain 63, Beal XXI 324, Plummer clxxxv, *Cross; Icelandic: *Boberg; Finnish: Kalevala runes 12, 36, 39; Breton: Sibillot Incidents s.v. "ipie"; French Canadian: Barbeau JAFL XXIX 10; Spanish: Boggs FFC XC 67 No. 508*; Scotch: Macdougall and Calder 197; English: Wells 115 (Sir Eglamour of Artois); Greek: Frazer Apollodorus II 63 n. 4; Jewish: *Neuman; India: Thompson-Balys; Arabian: Burton Nights IV 176, VI 221f., 230, VII 41, 44, S VI 221f., 426, S VII 257; IndoChinese: Scott Indo-Chinese 303; Chinese: Werner 121, 236; Korean: Zong in-Soh 59; Cape Verde Islands: Parsons MAFLS XV (1) 354; S. A. Indian (Mataco): Mitraux

Figure 3. Web interface of the Meertens Online Motif FindER.

(consisting of a upper-level letter and a lower-level number key); a primary description; a secondary description (which is optional); and bibliographical information.

> H659.1.1. *What is oldest?* God. BP II 358.

After preprocessing, a motif like this is represented as follows:

id	primary description	additional description	references
H659.1.1.	*What is oldest?*	God.	BP II 358

Here 'id' represents the unique key as given by Thompson; 'primary description' and 'additional description' provide a verbal characterization of the motif, where the second part is usually an explication of the first part; 'references' points to any references and bibliographic information (e.g. collector, collection, and/or location) available for this motif. With the help of a number of programming scripts, we extracted these fields for all motifs in the index.[7]

Semantic Expansion

One of the key features of our search engine is the way in which motifs are semantically expanded to match more generic descriptions. If a user issues a query such as *colour animal*, we want the system to not only return direct hits (motifs in which both words occur), but also motifs containing instances of these words with either a higher or a lower specificity, such as B731.2 *Green horse*. In order to do so, we matched all lemmatized nouns and adjectives in the motif descriptions against the semantic lexicon WordNet (Fellbaum 2005). WordNet is a large semantic lexicon of English in which words of various syntactic categories are grouped into sets of related words, or 'synsets'.[8] The relations between different synsets are encoded by means of super-subordinate relations (or hyperonyms). Thus, a word like *furniture* is linked to both more specific examples (or hyponyms) such as *bed* and *chair* and more general concepts (or hyperonyms) such as *artefact*, and ultimately *entity*, which forms the root node of the network (see Figure 4). While it is possible to search for such far-removed parents, MOMFER by default returns only the two immediate parental relations, primarily for efficiency reasons and also because a complete expansion would return all motifs in every query, only in different orderings.

By linking motifs with words in other motifs based on hyperonymic relations, the search engine provides a completely new way of accessing the TMI. This enables one to find many related motifs that are otherwise difficult to detect. To give a small example, say we are interested in the various occultists present in the TMI. The umbrella term *occultist* does not occur at all in the index, yet we know the index is packed with motifs about witches, wizards, sorcerers, druids, magicians, enchanters, and so forth. In WordNet all these words are connected through the hyperonym *occultist*. By connecting motifs based on their hyperonymic relations with other motifs, we allow users to extract all types of occultists using the single search query *occultist*.[9]

Indexing Schema

Figure 5 provides a detailed view of our indexing schema. All motifs are preprocessed and fed to an indexer. This indexer extracts the various fields described above and stores

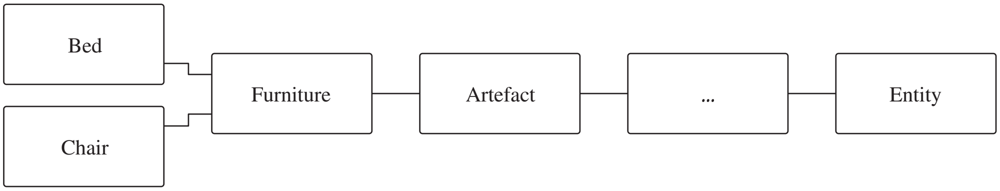


Figure 4. Example of hyperonymic and hyponymic relations in WordNet.

the result in an index. When a user issues a search query, the index retrieves the top results which, after scoring and sorting, are presented to the user.[10]

The system assumes that not all matches are equally important. If, for example, a user enters the query *book*, motifs that contain the word *book* in the reference field are intuitively less relevant than those in which it is part of the primary description. To reflect this intuition, the system weighs the matches in the different fields in the following descending order: (1) primary description, (2) additional information, (3) WordNet expansion, and (4) bibliographical information and location (references). Thus, a query for the word *book* will first return motifs where the word is found in the primary description and/or additional information, before it turns to motifs matching the WordNet expansion.

Query System

MOMFER provides an expressive and rich query system, allowing users to retrieve their results efficiently. In this section we will describe the most important search features implemented in the tool.

Single word search is the simplest form of search, where users search for a single word. The index retrieves all motifs containing the search word and ranks them according to how informative the word is for that particular motif. If, for example, a user searches for the word *devil*, the motif G303 ('Devil') will be ranked highest because it contains no other words than the search term. Subsequently, if a motif mentions a word more than once, it will lead to a high ranking. For example, motif G303.9.4.7.1 (Devil and girl. 'Are you lonely?'; Girl: 'No, devil, with God and angels') has *devil* in both the primary description and in the additional information, leading to a high ranking. As explained

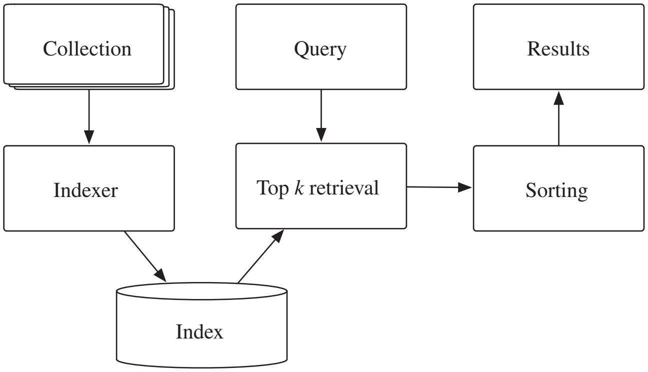


Figure 5. Search engine schema.

above, the system will also look for the two immediate WordNet expansions of the search term and their daughter terms.

Another search feature is multi-word search. By default all queries are represented as Boolean 'OR' statements. This means that when a user searches for two words, for example *cat* and *transformation*, the index will search for all motifs containing either *cat* or *transformation*. Motifs that contain both search words are considered to be better matches and will be ranked higher than those containing only a single word. To force the system to retrieve exclusively motifs containing both search terms, users need to add the Boolean operator 'AND', as in *cat* AND *transformation*. Naturally, this can be extended to as many search terms as the user requires.

The Boolean operator 'AND' requires two or more juxtaposed terms to co-occur in the same motif. Phrasal search can be seen as a further restriction of searching with 'AND' where the index attempts to match a list of two or more contiguous words. This functionality is provided by enclosing two terms between double quotation marks, as in "black cat".

Finally, we describe field-specific search. By default, the index searches for matches of a particular word in all available fields (description, additional description, WordNet expansions, and references) using the field weights as described above. Users can override this default behaviour by making explicit in which field they want to search. For example, to search for all motifs that mention a reference to Baughman, one could issue the query *references:Baughman*, where the search term is preceded by the expression *references:*. Similarly, one could search for all motifs that contain words with the concept 'colour' as one of its hyperonymic parents, but not the word *colour* itself, using *wn:colour*. Users can search the different fields individually using the following keywords, each followed by a colon:

- *motif* (to search for motif IDs, e.g. *motif:A100*);
- *description* (to search for words solely in motif descriptions, e.g. *description:magic*);
- *additional* (to search for words solely in the additional descriptions of motifs, e.g. *additional:dragon*);
- *wn* (to search for concepts available in the WordNet expansions, e.g. *wn:instrument*);
- *references* (to search for words that are part of the references (e.g. sources or countries of origin) in the index, e.g. *references:Thompson*);
- *location* (to search for motifs on the basis of the country of the sources in which a motif appears, e.g. *location:india*).

Case Studies

While the goal of this paper is expressly to introduce the search engine rather than to present actual research, we nevertheless selected some case studies to highlight the different ways in which MOMFER can be utilized to ask new questions and to enclose different material. All of these case studies are necessarily tentative, as are the hypotheses and conclusions based on them.

Monster Sighting

Monsters are an important part of folklore, often fulfilling the role of antagonist. Problematically, however, they are not found in a single category in the TMI, but rather are spread out under different subheadings, such as *Mythological Animals* (B0–B99), *Marvelous Creatures* (F200–F699), and *Kinds of Ogres* (G10–G399). Furthermore, monsters can reside in less obvious habitats, such as Category C (*Tabus*, e.g. C311.1.4 *Tabu: looking at werewolf*), Category H (*Tests*, e.g. H1174.2 *Task: overcoming dragon*), and Category V (*Religion*, e.g. V236.1 *Fallen angels become fairies (dwarfs, trolls)*). The semantic query expansion implemented in MOMFER provides a convenient solution for such search problems. After a filtering step, the search query *wn:monster* results in 352 motifs containing one or more monsters. Dragons and serpents dominate this result list with 199 and 191 mentions respectively. They are followed by a long tail distribution of monsters such as griffins, werewolves, chimeras, unicorns, and so on and so forth.

Thompson intended his TMI to cover motifs from all over the world. Despite the fact that the references only 'give some preliminary guidance in finding examples of the item concerned' (Thompson 1955–58, 24), leading to some notable geographical preferences (India) and geographical gaps (e.g. the Arab world; see for commentary and resolution El-Shamy 1980, 1995), it can be argued that Thompson achieved his goal quite well. The TMI mentions over five hundred different locations ranging from small villages in California to islands around the Coral Sea. This information is unique and can provide us with interesting suggestions about the geographical preferences of certain motifs. We are interested in the geographical distribution of the monsters found in the previous paragraph. Which nations prefer werewolves and where do all those dragons hide out?

For each of the 352 motifs mentioning a monster (according to the classification by WordNet), we extracted all the locations in the references field. The locations were aggregated by country; that is, place names, provinces, and so forth were replaced by the name of their corresponding countries. The geographical distribution is visualized in Figure 6. The colour gradient represents the frequency with which monsters are found in a particular country, ranging from white (zero monsters) to black (most monsters). We see a strong preference for sources containing monsters from Ireland, Iceland, India, and, most notably, China. We acknowledge that given Thompson's relatively unbalanced data collection, the significance of these findings may be relatively trivial and incomplete, but the results serve as an illustration of the ease with which similar information could be extracted from the TMI and hint at certain trends that could be investigated more thoroughly.

Black, White, and Red: Colour Term Appearance in the TMI

Colour naming has long been of interest to researchers in such diverse fields as psychology, linguistics, and anthropology. It gained renewed interest with the publication of Brent Berlin and Paul Kay's groundbreaking study *Basic Color Terms: Their Universality and Evolution* (1969). Their main hypothesis was that 'color categorization is not random and the foci of basic color terms are similar in all languages' (Berlin and Kay 1969, 10). This universalist hypothesis, which states that the addition of colours to a language follows a more or less implicational hierarchy (all languages distinguish the concepts *black* and *white*, languages with three colours add *red* to their arsenal, but no language has only *red* and *white*), has withstood the test of time in key respects, despite the vast amount of critique from more culture-relativistic oriented researchers.

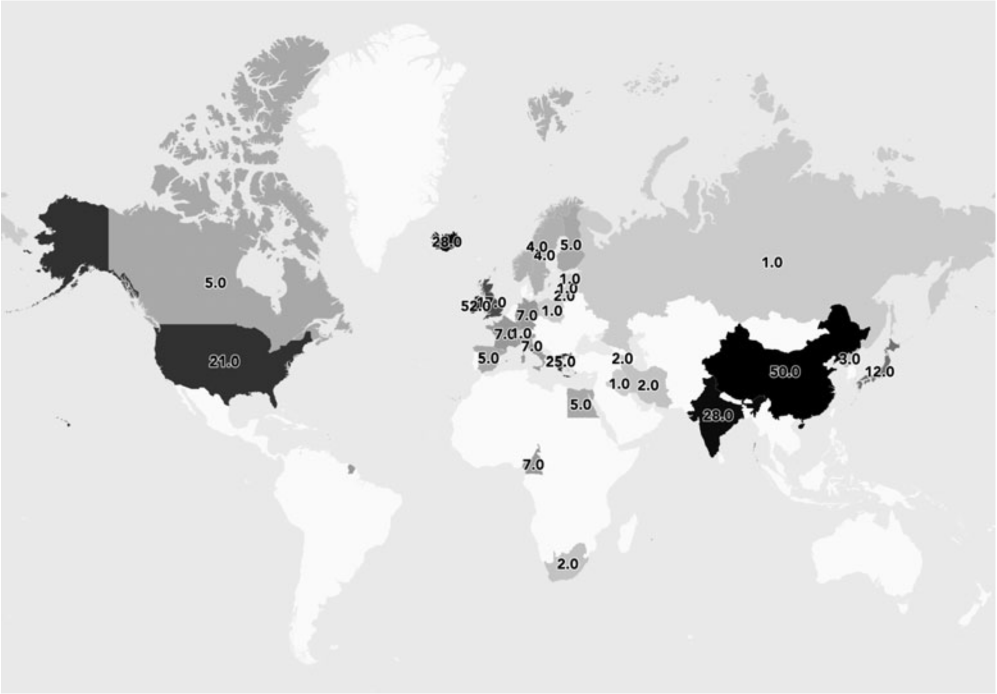


Figure 6. Geographical distribution of motifs in Thompson's *Motif-Index of Folk Literature* containing the WordNet concept 'monster'.

The discoveries of Berlin and Kay also gave impetus to folkloristic research. Ralph Bolton and Diane Crisp (1979), for example, demonstrated that there is a positive correlation between the relative salience of colour categories in folktales and the evolutionary development as proposed by Berlin and Kay. In other words: the higher the colour in the hierarchy, the more frequently it occurred in folktales. In a recent study, Jessica Hemming (2012) proposes an evolutionary source underlying the resonance of the tricolour in symbolic contexts all over the world.

It is to be expected that the importance of colour terms will find its reflection in the TMI as well. An example of a motif containing a colour name is the motif Z65.1.1 *Red as blood, white as snow, (and black as raven)*, which in Western folklore is most closely associated with the tale of Snow White (classified as ATU 709 in Uther 2004). This motif neatly reflects the importance of the tricolour *black*, *white*, and *red*. To (tentatively) investigate whether the TMI as a whole reflects the hierarchy of colour terms as proposed by Berlin and Kay, we used MOMFER to search for all motifs that mention a colour name using the field-specific search query *wn:color*. This results in 581 motifs containing either chromatic (i.e. blue, green, etc.) or achromatic (i.e. black, white, and grey) colours. We visualize the distribution of the colour terms in these motifs in Figure 7. The size of the plates represents the frequency with which each colour occurs in the TMI. The clear significance of the tricolour *black*, *white,* and *red* is in accordance with findings in previous studies. Again, it must be noted that our investigation serves only to illustrate how MOMFER can serve as a tool to extract (new) information from the

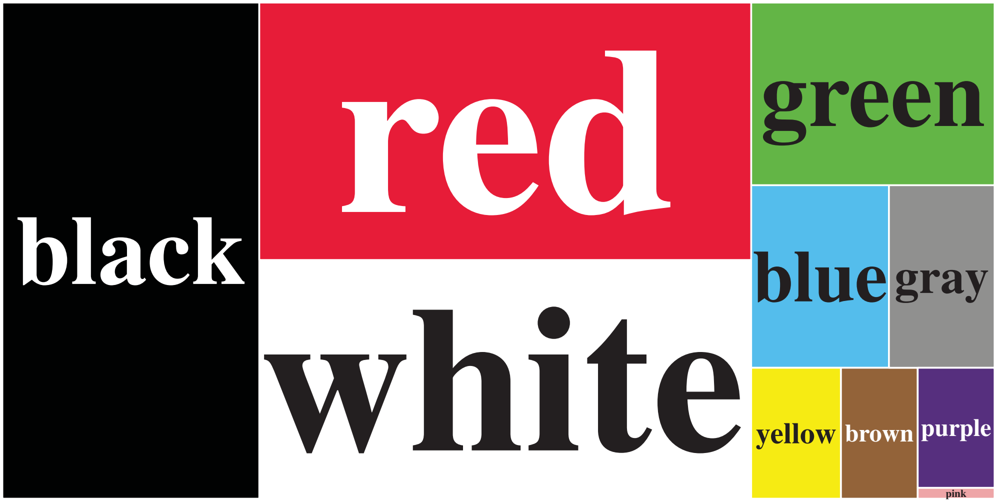


Figure 7. Visualization of colours in the *Motif-Index of Folk Literature*. Plates are sized according to the frequency of their corresponding colour.

TMI. However, it is noteworthy that even in a limited corpus such as the TMI the three colours *black*, *white*, and *red* are indeed the most common.

Marvellous Men, Deceptive Women

Thompson's TMI has been subjected to a number of gender studies. Torborg Lundell in particular points to a range of gender biases present in the index and makes a strong case for some textual revisions to it (Lundell 1983, 1989; for a different view, see El-Shamy 1990, 83–85):

> The Motif-Index in general (1) overlooks gender identity in its labelling of motifs, thus lumping male and female actions or characters under the same, male-identified heading or (2) disregards female activity or (3) focuses on male activity at the cost of female. (Lundell 1989, 150)

Leaving any intrinsic commentary aside, we decided to investigate quantitatively how prominent these gender biases are. We extracted all motifs from the TMI mentioning either a female character or a male character. In this example, the advantage of the semantic word expansion of MOMFER becomes apparent. In the previous case study it would be possible manually to extract all motifs containing a colour term (although this would be quite a time-consuming exercise), because the list of possible primary colours is finite and quite small. Extracting all mentions of male and female characters, however, is barely possible by hand, because of the endless list of partial synonyms in both categories (e.g. lumberjack, carpenter, etc.). All these words are captured by the semantic expansion mechanism of MOMFER, however, yielding the results we wanted.

We issued two queries: *wn:male* and *wn:female*, resulting in 4178 motifs containing a male character and 3318 motifs mentioning a female character. These numbers may already indicate some gender biases, since there is no a priori reason to assume that men are more common in stories than women. The gender biases become even more

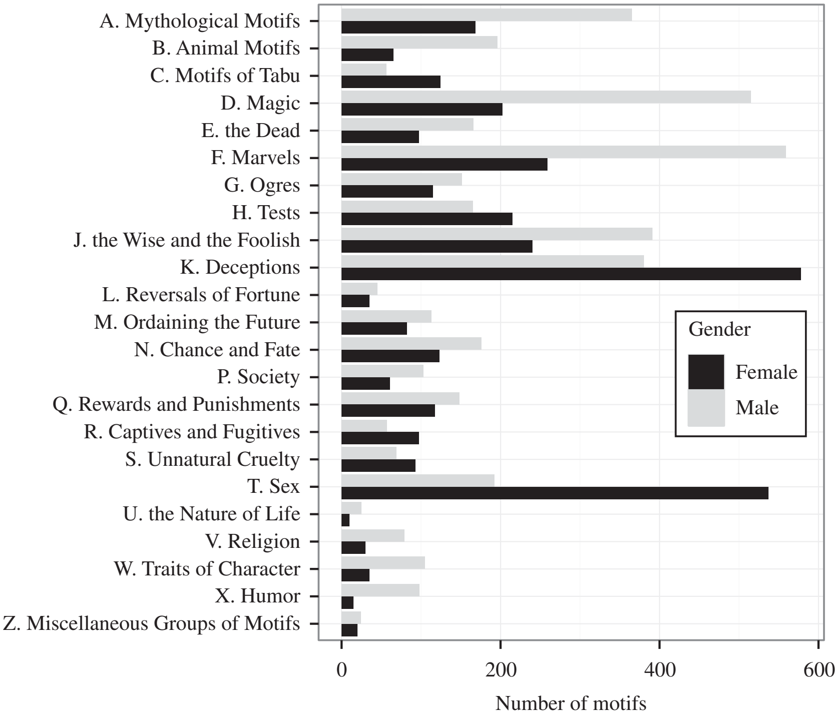


Figure 8. Distribution of male versus female entities in Thompson's *Motif-Index of Folk Literature*.

prominent when we look at the distribution of male versus female characters over the main categories in the TMI. Figure 8 shows that male characters dominate most motif categories. They are most prominent in *Mythological Motifs* (Category A), *Magic* (Category D), *Marvels* (Category F), and *the Wise and the Foolish* (Category J). Women, on the other hand, dominate the categories *Deception* (Category K) and *Sex* (Category T). These results confirm previous investigations and amplify the need for caution for modern scholars when exploring motifs mentioning male and female personas in the index.

Conclusion

In this paper we presented MOMFER, a new, fast research tool accompanying Thompson's TMI. We explicated the production process, documented the query system, and showed in three small case studies how the tool can be put to use to explore the index in new ways.

A complete collection of all motif indices would be an amazing research tool. MOMFER is only a first step towards such an index: with its flat architecture, it was built explicitly to enable the inclusion of other motif indices. As Heda Jason notes, proposals for new motifs should be made 'only after it has been found that no other motif index has listed the specific content element in question' (Jason 2000, 61). An integrated tool such as MOMFER could serve as a warrant for such requirements. We sincerely hope that researchers will join us in disclosing other indices so that we may finally have, more than fifty years after Thompson's great index, a complete index of all folklore motifs.

Notes

[1] We would like to thank Basten Stokhuyzen (www.bstn.nl) for designing the web interface. Also, we would like to thank *Folklore*'s two reviewers for their helpful comments. The work on which this article is based has been supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Tunes & Tales project.

[2] For further examples of both motif indices as well as folktale collections, see Uther (1996).

[3] Available at http://www.momfer.ml

[4] Cf. Declerck and Lendvai (2011) for a similar approach that attempts to enrich the motif index with semantic information.

[5] For an online example, see http://www.ruthenia.ru/folklore/thompson/

[6] Unfortunately, we were unable to acquire a copy of this edition, which is why we have to refrain from a more detailed comparison. However, the fact that no research institution or university library in the Netherlands has a copy available is a case in point. We base ourselves on Smith (1994), which is a review of this digital edition.

[7] The parsed index as well as the source code of MOMFER is available online at https://github.com/fbkarsdorp/tmi

[8] While there has been some criticism of the usefulness and limitations of WordNet (for example, Wilks, Slator, and Guthrie 1996; Lenat, Miller, and Yokoi 1995), our case studies show that, for the purposes of this research tool, synsets and WordNet are in fact valuable methods that improve the retrieval results.

[9] When linking the words from the TMI to the entries in WordNet, we did not perform any semantic disambiguation. We linked each word to the most common sense in WordNet. In a later stage, this heuristic could be further refined.

[10] We make use of the programming library Whoosh, a fast Python-based search engine library. See http://whoosh.readthedocs.org/

References Cited

Aarne, Antti. *Verzeichnis Der Märchentypen* [Catalogue of folktale types]. Folklore Fellows Communications 3. Helsinki: Academia Scientarium Fennica, 1910.

———. *The Types of the Folktale*. Folklore Fellows Communications 74. Helsinki: Academia Scientarium Fennica, 1928.

———. *The Types of the Folktale: A Classification and Bibliography*. Translated and enlarged by Stith Thompson. 2nd ed. Helsinki: Suomalainen Tiedeakatemia/Folklore Fellows Communications, 1961.

Baughman, Ernest W. *Type and Motif-Index of the Folktales of England and North America*. Indiana University Folklore Series 20. The Hague: Mouton, 1966.

Bell, Sita. 'Anti-Semitic Folklore Motif Index'. Master's thesis, Utah State University, 2009. http://digitalcommons.usu.edu/etd/299

Ben-Amos, Dan. *Folktales of the Jews*. 3 vols. Philadelphia: Jewish Publication Society, 2006.

Berlin, Brent, and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press, 1969.

Bolton, Ralph, and Diane Crisp. 'Color Terms in Folk Tales: A Cross-Cultural Study'. *Cross-Cultural Research* 14 (1979): 231–53.

Childers, James Wesley. *Tales from Spanish Picaresque Novels: A Motif Index*. Albany, NY: State University of New York Press, 1977.

Crooke, William, and Pandit Pam Gharib Chaube. *Folktales from Northern India*. Santa Barbara, CA: ABC-CLIO, 2002.

Declerck, Thierry, and Piroska Lendvai. 'Linguistic and Semantic Representation of the Thompson's Motif-Index of Folk-Literature'. In *Research and Advanced Technology for Digital Libraries*, edited by S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, 151–58. Lecture Notes in Computer Science 6966. Berlin: Springer, 2011.

Dundes, Alan. 'The Motif-Index and the Tale Type Index: A Critique'. *Journal of Folklore Research* 34, no. 6 (1997): 195–202.

El-Shamy, Hasan. *Folktales of Egypt*. Chicago: University of Chicago Press, 1980.

———. 'A Type Index for Tales of the Arab World'. *Fabula* 29, nos 1/2 (1988): 150–63.

———. 'Oral Traditional Tales and the Thousand Nights and a Night: The Demographic Factor'. In *The Telling of Stories: Approaches to a Traditional Craft*, edited by Morton Nøjgaard, J. de Mylius, I. Piø, and Bengt Holbeck, 63–117. Odense: Odense University Press, 1990.

———. *Folk Traditions of the Arab World: A Guide to Motif Classification*. Bloomington: Indiana University Press, 1995.

Fellbaum, Christiane. 'WordNet(s)'. In *Encyclopedia of Language and Linguistics*. 2nd ed., edited by Keith Brown (editor-in-chief), 665–70. Oxford: Elsevier, 2005.

Flowers, Helen Leneva. *A Classification of the Folk Tale of the West Indies by Types and Motifs*. New York: Arno, 1980.

Goldberg, H. *Motif-Index of Medieval Spanish Folk Narratives*. Tempe, AZ: Medieval & Renaissance Texts & Studies, 1998.

Hemming, Jessica. 'Red, White, and Black in Symbolic Thought: The Tricolour Folk Motif, Colour Naming, and Trichromatic Vision'. *Folklore* 123, no. 3 (2012): 310–29.

Ikeda, Hireko. *A Type and Motif Index of Japanese Folk-Literature*. Folklore Fellows Communications 209. Helsinki: Academia Scientarium Fennica, 1971.

Jason, Heda. *Motif, Type and Genre: A Manual for Compilation of Indices and a Bibliography of Indices and Indexing*. Helsinki: Suomalainen Tiedeakatemia, 2000.

Karsdorp, Folgert, Peter van Kranenburg, Theo Meder, and Antal van den Bosch. 'In Search of an Appropriate Abstraction Level for Motif Annotations'. In *Proceedings of the 2012 Computational Models of Narrative Workshop*, edited by Mark Finlayson, 22–26. Istanbul: 2012.

Kirtley, Bacil F. *A Motif Index of Traditional Polynesian Narratives*. Honolulu: University of Hawai'i Press, 1977.

Legman, Gershon. 'Toward a Motif-Index of Erotic Humor'. *The Journal of American Folklore* 75, no. 297 (1962): 227–48.

Lenat, Doug, George Miller, and Toshio Yokoi. 'CYC, WordNet, and EDR: Critiques and Responses'. *Communications of the ACM* 38, no. 11 (1995): 33–38.

Lundell, Torborg. 'Folktale Heroines and the Type and Motif Indexes'. *Folklore* 94, no. 2 (1983): 240–46.

———. 'Gender-Related Biases in the Type and Motif Indexes of Aarne and Thompson'. In *Fairy Tales and Society: Illusion, Allusion, and Paradigm*, edited by Ruth B. Bottigheimer, 149–63. Philadelphia: University of Pennsylvania Press, 1989.

Meder, Theo, ed. *De Magische Vlucht. Nederlandse Volksverhalen uit de Collectie van het Meertens Instituut* [The magic flight. Dutch folktales from the collection of the Meertens Institute]. Amsterdam: Uitgeverij Bert Bakker, 2000.

Neuland, Lena. *Motif-Index of Latvian Folktales and Legends*. Folklore Fellows Communications 229. Helsinki: Academia Scientarium Fennica, 1981.

Slone, Thomas H. *One Thousand One Papua New Guinean Nights. Folktales from Wantok Newspaper*. Vol. 1: *Tales from 1972–1985*. Oakland, CA: Masalai, 2001.

Smith, Allen. 'JAL Guide to Software, Courseware and CD-ROM: Stith Thompson's Motif-Index of Folk Literature'. *Journal of Academic Librarianship* 20, no. 4 (1994): 255.

Thompson, Stith. *Motif-Index of Folk-Literature. A Classification of Narrative Elements in Folk-Tales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-Books, and Local Legends*. Bloomington: Indiana University Press, 1932–36.

———. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jestbooks, and Local Legends*. Rev. and enl. ed. Bloomington: Indiana University Press, 1955–58.

———. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jestbooks, and Local Legends*. Bloomington: Indiana University Press, 1993. CD-ROM.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 'Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network'. In *Proceedings of HLT-NAACL 2003*, 252–59. Stroudsburg, PA: Association for Computational Linguistics, 2003. http://nlp.standford.edu/kristina/papers/tagging.pdf

Uther, Hans-Jörg. (1996) 'Type- and Motif-Indices 1980–1995: An Inventory'. *Asian Folklore Studies* 55, no. 2 (1996): 299–317.

———. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*. 3 vols. Folklore Fellows Communications 284, 285, and 286. Helsinki: Academia Scientarium Fennica, 2004.

Wilks, Yorick, Brian M. Slator, and Louise Guthrie. *Electronic Words: Dictionaries, Computers, and Meanings*. Cambridge, MA: MIT Press, 1996.

Biographical Notes

Folgert Karsdorp is a PhD candidate at the Meertens Institute in Amsterdam, The Netherlands, where he is involved in the Tunes & Tales project. He is affiliated with Radboud University and the eHumanities Group of the Royal Netherlands Academy of Arts and Sciences (KNAW).

Marten van der Meulen is a research assistant for the Tunes & Tales project at the Meertens Institute in Amsterdam. He is currently finishing his Research Master's at Leiden University.

Theo Meder is a senior folk narrative researcher at the Meertens Institute in Amsterdam who deals with research and documentation of Dutch language and culture. He is supervisor of the Dutch Folktale Database (www.verhalenbank.nl) and (co-) leader of several computational humanities projects.

Antal van den Bosch is professor of language and speech technology at Radboud University, Nijmegen, The Netherlands. He develops and studies computational models that learn to understand and generate natural language, leading to applications in machine translation, text mining, and computational humanities.