In search of an appropriate abstraction level for motif annotations

Folgert Karsdorp, Peter van Kranenburg, Theo Meder, Dolf Trieschnigg, Antal van den Bosch*

Meertens Institute, *Radboud University Nijmegen
Amsterdam, The Netherlands, *Nijmegen, The Netherlands
{folgert.karsdorp,peter.van.kranenburg,theo.meder,dolf.trieschnigg}@meertens.knaw.nl, a.vandenbosch@let.ru.nl

Abstract

We present ongoing research on the role of motifs in oral transmission of stories. We assume that motifs constitute the primary building blocks of stories. On the basis of a quantitative analysis we show that the level of motif annotation utilized in the Aarne-Thompson-Uther folktale type catalogue is well suited to analyze two genres of folktales in terms of motif sequences. However, for the other five genres in the catalogue the annotation level is not apt, because it is unable to bring to front the commonalities between stories.

1. Introduction

In oral culture artifacts such as stories are propagated through the community and passed on to successive generations. Stories contain all kinds of cultural ideas that are replicated via the process of storytelling. During replication, most elements of a story remain stable producing recognizable variants (lineages) of a cultural artifact. A story like *Little Red Riding Hood* can be told in various ways, i.e. can have many textual forms, but at a more abstract level, the essence of the story remains virtually untouched. How can this be?

Our ultimate goal is to create a model of oral transmission of folktales. We hypothesize that oral transmission of folktales happens through the replication of sequences of motifs. In this view, motifs constitute the primary vehicles of cultural heritage in oral transmission of stories. A prerequisite for building such a motif-based model of oral transmission of stories is to formalize tales as sequences of motifs. Because the manual annotation of motifs is a time-consuming and error prone job with respect to consistency, we wish to create a system for the automatic recognition of motifs. This motif detection system will enable us to analyze large amounts of available data.

The term *motif* immediately gives rise to the question of what exactly is a motif. Without wanting to settle the debate, we propose, as a working hypothesis, that motifs are the simplest meaning-bearing units contributing to the overall plot of a story (Jason, 2007; Van Boven and Dorleijn, 2003). Following Thompson (1946, 415) we add to this that motifs should have "a power to persist in tradition". Motifs are thus recurring elements found in different stories (variants or types). In this paper we investigate whether motifs form the primary building blocks for stories. With a more or less fixed set of motifs we can analyze a large number of stories. We hope to find evidence for the idea that the way in which motifs can be recombined to produce new stories can best be described with a motif-based story grammar.

In this paper, we investigate a small part of our motif definition, namely what we should conceive as the *simplest* units for the task at hand. That is, what level of description of motifs is appropriate for (1) modeling oral transmission of stories and (2) conceiving stories as sequences of motifs? We will do so by means of a quantitative analysis of the authoritative folktale type catalogue *The Types of International Folktales* by Aarne, Thompson and Uther (henceforth: ATU catalogue) (Uther, 2004).

The outline of the paper is as follows. We will start with a brief theoretical background about the term motif and the materials used in this study. We then continue with the analysis of the ATU catalogue in which we examine whether the description level of motifs in the catalogue is appropriate for modeling stories as sequences of recurring motifs. The last section offers our conclusions and directions for further research.

2. Different levels of abstraction

We assume that motifs are the simplest meaning-bearing units of a story that have a power to persist tradition. Now, what do we mean by 'simplest'? Are motifs the elaborate and abstract functions that Propp (1968) distinguishes, or the many thousands of small and hierarchically ordered content units in Stith Thompson's (1955 1958) *Motif-Index* (henceforth: TMI)?

Propp (1968) recognizes 31 plot units which he calls *functions*, common to a small subgroup of fairy tales. An example of a function is given under (1):

(1) ABSENTION: A member of a family leaves the security of the home environment.

One important aspect of Propp's theory is that the functions abstract away from specific characters (*dramatis personae*). So, 'a member' may be any kind of hero in the story or a member of the family that the hero will later need to rescue. In the TMI we find over 45.000 motifs hierarchically ordered in a tree structure. Many motifs are bound to particular folktale types. Under (2) we list some examples:

(2) Q426 Wolf cut open and filled with stones as punishment;
F911.3 Animal swallows man (not fatally);
F823.2 Glass shoes;
J346 Better be content with what you have, than try to get more and lose everything.

In modeling cultural evolution it is important to realize that the more abstract the level at which we identify motifs,

the harder it is to trace lineages with confidence (Dennett, 1995, 357). If the level of comparison is too abstract, we can only identify very general commonalities that are not distinctive enough. Therefore we must take into account the particular forms of expression with which motifs are realized. With this in mind, the rather abstract functions of Propp seem less appropriate for modeling oral transmission of folktales than the more concrete motifs in the TMI, at least if we would use Propp's functions exclusively. Another reason why Propp's functions seem less suitable is that they are only defined for one group of fairy tales, whereas we would like a system that can cope with all kinds of folktales, including genres such as traditional and contemporary legends and jokes.

Therefore, we will use the motifs from the TMI as a starting point. In the ATU catalogue, the motifs from the TMI play a key role in the classification of tales into a certain type. Every folktale type contains a short summary of the plot. In this summary we find a sequence of specific motifs that constitute the primary descriptive units of a tale type without an overarching level. An example of a story summary in the ATU catalogue:

> ATU 327A "**Hansel and Gretel**. A (poor) father (persuaded by the stepmother) abandons his children (a boy and a girl) in the forest [S321, S143]. Twice the children find their way back home, following scattered pebbles [R135]. On the third night, birds eat the scattered peas (breadcrumbs) [R135.1]. The children come upon a gingerbread house which belongs to a witch (ogress) [G401, F771.1.10, G412.1]. She takes them into her house. The boy is fattened [G82], while the girl must do housework. The witch asks the boy to show his finger in order to test how fat he is [G82.1], but he shows her a bone (stick) [G82.1.1]. When the witch wants to cook the boy, the sister deceives her by feigning ignorance and pushes her into the oven [G526, G512.3.2]. [...] The children escape, carrying the witch's treasure with them. Birds and beasts (angels) help them across water. They return home."

This folktale type is defined by the motif sequence:

(3) `([S321, S143] [R135] [R135.1] [G401, F771.1.10, G412.1] [G82] [G82.1] [G82.1.1] [G526, G512.3.2])`

We take the collection of folktale summaries in the ATU catalogue to be a corpus of stories with motif annotations. We will use this corpus to investigate the question whether we can use the concrete level of motif description utilized in this corpus for extracting a grammar of folktales consisting of sequences of recurring motifs.

As a first step, we aim to establish experimentally that motifs indeed represent the primary recurring building blocks of stories. Different combinations of motifs (some new and some old) give rise to new stories, some of which are variations of already existing types, others give rise to new types. If motifs represent the building blocks of a story – just as words form the basic elements of a sentence – they should recur in different stories. We can consider motifs to be recurring in two ways. First, motifs are 'recurring' if they are found in multiple *variants* of a particular folktale type. The version most widely known today of *Little Red Riding Hood* is based on the Brothers Grimm version. Charles Pernault gives a variant of the story in which little red riding hood is not rescued from the belly of the wolf. Still, both stories share a sufficient number of motifs to conceive them as variants of the same type. Second, motifs can also be said to be recurring if they occur in other story *types*. In both *Little Red Riding Hood* (ATU 333) and *The Wolf and the Kids* (ATU 123), the wolf is "cut open and filled with stones as punishment" which is motif Q426 in the TMI. The ATU catalogue only provides information about folktale types and not about variants. Therefore, 'recurring' in this paper only means recurring in different folktale types.

Genre	# tale types	# motifs
Animal tales	298	478 (1.6)
Tales of magic	223	1573 (7.1)
Realistic tales	200	666 (3.3)
Tales of the stupid ogre	124	184 (1.5)
Religious tales	140	397 (2.8)
Anecdotes and jokes	675	1069 (1.6)
Formula tales	47	80 (1.7)
Total	1707	4447 (2.6)

Table 1: Basic statistics about the contents of the ATU. For each genre the table shows the number of folktale types, the number of motifs and the average number of motifs per tale type.

3. Quantitative analysis of the ATU catalogue

3.1. Statistics

The ATU catalogue lists 2247 unique folktale types divided into seven genres. In our analysis we will only use the 1707 types that explicitly mention the motifs that belong to that type. These 1707 types contain 4447 motif instances and 3698 unique motifs. The fact that motifs are infrequently reused indicates a high *specificity* of the index language; the low average motif sequence length (2.6) indicates a low *exhaustivity* of indexing (Van Rijsbergen, 1979, 13). Table 1 presents some basic statistics per genre. We see that 'Anecdotes and Jokes' are overly well represented. Furthermore, we see that the average motif sequence length is much longer in 'Tales of Magic' than in any other genre.

We are interested in the way folktale types are related in terms of their motif sequences. If certain motifs recur in different folktale types, this provides evidence for the idea that motifs can be used as building blocks for (new) stories. Each folktale type T can be considered as an n-dimensional vector of attribute values:

$$T = (m_1, m_2, m_3, \ldots, m_n) \quad (4)$$

S_k	k
338	1
6	1
5	1
4	4
3	14
2	65
1	1170

Table 2: Frequency spectrum of subgraph size.

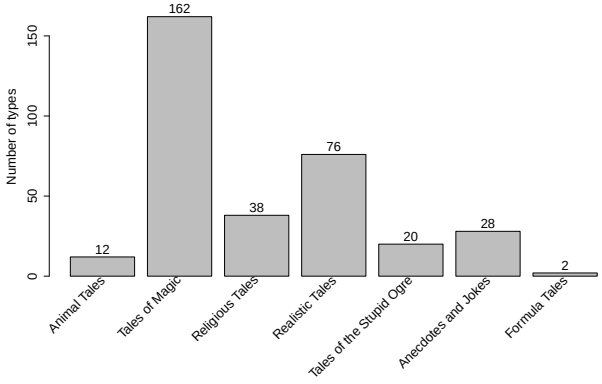

Figure 1: Frequency of genre types in main subgraph.

where m represents a motif and n the number of motifs of T in the ATU catalogue. We construct an undirected graph in which all 1707 folktale types from the ATU catalogue represent the nodes. The nodes are connected to each other if they have any motifs in common. The weight of the edges is defined by the number of overlapping motifs between two types, i.e. the length of their intersection or matching coefficient. Note that in this graph motif sequences are considered sets or 'bags of motifs'. In this paper we do not take into account the particular order of motifs.

The graph contains 1256 subgraphs, but not all subgraphs are equally large. Table 2 shows the frequency spectrum for the size of the different subgraphs. It summarizes the frequency distribution in terms of the number of subgraphs with size S_k per frequency class k. There is one subgraph that consists of 338 folktale types, one subgraph that contains 6 types, and there are 65 subgraphs with a node size of 2.

3.2. Motifs as building blocks

31 percent of the types in the ATU catalogue are connected to other types. The connected folktales form subgraphs of different sizes. One large subgraph containing 338 nodes stands out. In this subgraph, many folktales share motifs which supports the view of motifs as building blocks of stories.

It would be interesting to have insight into which folktale types and genres are represented in this subgraph. Figure 1 clearly shows that most of the types ($N = 162$) belong to the genre of 'Tales of Magic'. This indicates that folktale types in the category 'Tales of Magic' have the largest amount of homogeneity between them in terms of their motif material. In second place come the 'Realistic Tales'. This is interesting as it might reflect the idea that Tales of Magic and Realistic Tales are alike: they both describe adventures and heroes and differ only in the use of magic. Moreover, we see that all distinguished genres in the ATU catalogue are represented in the subgraph. This shows that there are no clear boundaries between the genres in terms of the motifs used. Overall we can state that the existence of overlapping motifs supports the idea that motifs can be recombined to produce new story types.

The subgraph can provide us with information about the relative importance of folktale types based on their positions in the structure of the complete subgraph. This is useful information for a model of oral transmission of folktales, because it adds insight into where many motifs are exchanged. Moreover, because many motifs are shared among many folktale types, these 'exchange centers' provide the strongest case for the view of stories as sequences of recurring motifs.

There are a few ATU types that can be expected to be central nodes, because, at least in narrative practice, they seem to supply other types with one or more of their own motifs. These 'central stations' in the land of Tales of Magic are: (1) *The Dragon-Slayer* (ATU 300), (2) *The Magic Flight* (ATU 313), (3) *Bird, Horse and Princess* (ATU 550) and (4) *Water of Life* (ATU 551). These types play an important role in Propp (1968) and should contain the prototypical motif sequences that Tales of Magic have. We will use the graph to test whether the hypothesized centrality of these types is justified.

One way to approach a node's centrality is by looking at its degree. We are interested in those cases where a folktale type is connected to many other folktales and where many unique motifs are shared. Therefore, we define the degree of a folktale as the number of unique motifs shared with its connected nodes. More formally, we take the length of the union of the intersections between a type T and the types $(T_1, T_2 \dots T_n)$ connected to T:

$$degree(T_j) = |\bigcup_{i=1}^{n}(T_j \cap T_i)| \tag{5}$$

Table 3 lists the top 10 nodes in the network. Two out of four tale types that were expected to be central nodes are present in this list: *Bird, Horse and Princess* and *Water of Life*. Compared to the average degree of 2.9, the other two types also score relatively high with a degree of 7 for the *Dragon-Slayer* and 9 for the *Magic Flight*. *The Animal as Bridegroom* has by far the highest degree. This type list motifs common to many other types, such as B620.1 ("Daughter promised to animal suitor") and D2006.1.1 ("Forgotten fiancee reawakens husband") which is also present in the *Magic Flight*. It remains to be seen whether this type fulfilled an exemplary role in oral transmission or whether it is

ATU type		degree
425A	*The Animal as Bridegroom*	18
403	*The Black and the White Bride*	14
425B	*Son of the Witch*	13
875	*The Clever Farmgirl*	12
560	*The Magic Ring*	12
550	*Bird, Horse and Princess*	12
531	*The Clever Horse*	12
400	*The Man on a Quest for his Lost Wife*	11
551	*Water of Life*	10
480	*The Kind and the Unkind Girls*	10

Table 3: Top ten folktale types in terms of their degree.

more likely to be the result of extensive borrowing of motifs from other stories. In any event, these types strongly suggest that, at least in the case of 'Tales of Magic' and to a lesser extent 'Realistic Tales', stories can be created by intertwining motifs from other stores.

3.3. Motifs are no building blocks

We should not, however, jump to any conclusions. Darányi and Forró (2011), for instance, have shown on the basis of cluster analyses that for a small part of the ATU catalogue (the genre of 'Tales of Magic') we can find partially overlapping types. However, for the ATU catalogue as a whole, many motif sets are mutually exclusive and the overlap between folktale types in terms of their motif material is rather sparse. The frequency spectrum in Table 2 shows that 1170 out of 1707 folktale types share no motifs with other types. For these types, the occurrence of a single motif is enough evidence to unambiguously distinguish a certain type. Within the group of tales that are connected to each other, there are 36 types that have completely equal sets of motifs. Again, this brings into question whether we can distinguish multiple folktale types on the basis of their motif material. Finally, there are many types ($N = 983$) that consist of a single motif. Here it becomes unclear what the difference is between a motif and a tale type (Dundes, 1997, 197). For all these cases, it has no added value to define folktale types in terms of sequences or sets of motifs.

The numerous folktale types in the ATU catalogue that consist of only unique motifs constitute a problem for a motif-based story grammar. We want to conceive motifs as the basic elements with which new stories (variants and types) can be produced. However, most motifs in the ATU catalogue are exclusively associated with single tale types. The ATU catalogue does not provide a positive clue that these motifs can recur. This does not exclude that they could recur, for instance in variants of the tale type.

Still, the amount of motifs unique to single tale types will hinder the generalization capabilities of any system induced from this data. To underscore this point more forcefully, let us explain this prediction in more detail.

On the basis of the frequency spectrum in Table 2, we can estimate the probability of finding a folktale type with solely non-overlapping motifs. This can be accomplished by dividing the number of tale types with unique motif material by the total number of tale types. We have $N = 1707$ folktale types and $n_1 = 1170$ types that have completely unique motif material. The probability P of finding a story with only non-overlapping and thus new motifs is $P = 0.69$, which is rather high.

From the viewpoint that the ATU catalogue is a motif-based classification system, all this indicates an important shortcoming of the system. The long tail distribution of types in the system shows that we are dealing with a collection of unrelated exemplars with little predictive power. The system falls short, because the description level of motifs is too specific to describe tales in terms of sequences of recurring motifs.

4. Concluding remarks

We investigated whether the description level of the ATU catalogue types is appropriate for a model of stories as sequences of recurring motifs. We have shown that many ATU types in the genres of 'Tales of Magic' and 'Realistic Tales' share motifs which makes it possible to describe motifs as building blocks to create (new) stories. However, for a model that aims to describe all kinds of folktales in terms of motif sequences, the description level of the ATU catalogue is inappropriate. The degree of specificity and consequently the lack of co-occurring motifs makes it hard to generalize over different stories in terms of their motif sequences. In the ATU catalogue, the majority of motifs cannot be conceived as building blocks for (new) stories. Therefore, we should aim to discover a way of formalization that is more appropriate; one that on the one hand enables us to discover enough commonalities between stories, while on the other ensures that enough distinctive features remain.

Future research should be directed towards a system where motifs exist at different levels of abstraction. In this multi-layered system, low-level motifs are compatible with those found in the TMI and will consist of particular phrases, often as concrete as strings of contiguous words ("the big bad wolf"). At a more abstract level, we will look for non-contiguous co-occurrences of higher-level linguistic elements, such as subject–verb–object triplets and sequences of triplets. This more semantic level of description strives to be compatible with the functions developed by Propp.

5. Acknowledgements

This work has been carried out within the Tunes and Tales project, which is funded by the Royal Netherlands Academy of Arts and Sciences through the Computational Humanities program.

6. References

Sándor Darányi and László Forró. 2011. Detecting multiple motif co-occurrences in the Aarne-Thompson-Uther tale type catalog: A preliminary survey. Presented at the second AMICUS workshop.

Daniel Dennett. 1995. *Darwin's dangerous idea: evolution and the meanings of life*. Penguin Books. Penguin Group, England, London.

Alan Dundes. 1997. The motif-index and the tale type index: a critique. *Journal of Folklore Research*, 34(3):195–202.

Heda Jason. 2007. About 'motifs', 'motives', 'motuses', '-etic/s', '-emic/s' and 'allo/s-', and how they fit together. *Fabula*, (48):85–99.

Vladimir Propp. 1968. *Morphology of the folktale*. Publication of the Indiana University Research Center in Anthropology, Folklore, and Linguistics. University of Texas Press, Austin.

Stith Thompson. 1946. *The Folktale*. Dryden Press, New York.

Stith Thompson. 1955–1958. *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jestbooks, and local legends*. Indiana University Press.

Hans-Jörg Uther. 2004. *The Types of International Folktales: a Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson*, volume 1–3 of *FF Communications*. Academia Scientarium Fennica, Helsinki.

Erica Van Boven and Gillis Dorleijn. 2003. *Literair Mechaniek, Inleiding tot de analyse van verhalen en gedichten*. Uitgeverij Coutinho.

Cornelis Joost Van Rijsbergen. 1979. *Information Retrieval*. Butterworths.