# **The conceptual model of the folk tale and its machine implementation**

## **The structure of the fairy tale**

### **introduction**

Based on the axiomatic theses listed below, I aim at a reliable approach to the formalization of the fairy tale and consequently intend to enable the automatic analysis of the fairy tale.

The theses are as follows:

1. The content of a fairy tale consists of a sequence of individual actions.
2. Each action in a fairy tale is led by a leading character or several such characters.
3. Each appearance of such a character or several characters triggers a new plot.
4. The motif, as a conceptual term for the smallest, inseparable element of content in the text, corresponds to an action that is presented in the course of such an appearance or, in other words, such a scene (a motif must not be larger than the scene).
5. The marking of the objectively recognizable scene in the text corresponds to the marking of the plot sections and thus also to the marking of the motifs.
6. The categorization of the plot sections and the characters involved corresponds to the definition of content constants for the entire genre.
7. The schematization of the lawful and consistent reciprocity between these constants corresponds to the modeling of a universal structure of the genre.
8. After the formalized description of the content constants and their mutual relationships in the fairy tale, it is possible to automatically annotate and evaluate each individual representative of this genre and create a machine-generated picture of similarities and differences between the compared data.

The most important finding of this work is the identification of a previously missing objective criterion for capturing the motif as the smallest narrative content element in the fairy tale (points 1 – 4 above).

The first part of this work describes the system of action categories and the categories of action-bearing characters developed on the basis of these findings. This is continued with an attempt to schematize the universal structure of the fairy tale. The next section of the work is devoted to the provision of markup as a means of standardized coding of the content. The work then presents some modules of the artificial assistant for the semi-automatic annotation of the folk tale as well as tools for the visual representation of the mutually compared data. The digital infrastructure discussed in this section of the work is fully accessible through the GitHub publication.

### **State of research**

In general, this work presents a digital development of the method of comparative fairy tale research and thus of the scientific heritage of international fairy tale research. The focus is on the achievements of the historical-geographical method [[1]](#footnote-1)of the Finnish school and the structuralist method developed by VJ Propp [[2]](#footnote-2).

Whether it is explicitly mentioned or not, the definition of the structure of the fairy tale genre serves as the target instance for any comparative analysis of fairy tales. For the historical-geographical method, this target instance consists in dividing the entire repertoire of fairy tales into types, these types into episodes, and these episodes into motifs.

The terminology used to describe the structural elements of fairy tales is often inconsistent, unless there is a uniform method of structuring them. This structure goes back to Aarne 's attempt at systematization . According to his model, the fairy tale is made up of the following elements ( Aarne , 1913: 65):

1. Narrative = element of the highest category,
2. Main part = middle element,
3. Main move = the smallest element.

This structure can be seen even better when it is put into practice. When analyzing the fairy tale *ATU 670 - Man who knows animal languages,* Aarne segments the content as follows:

*"The content of the fairy tale can be divided into three main parts: the first deals with the learning of animal language, the second with the quarrel between the man and the woman caused by the knowledge, which leads the man to prepare for death, and the third with the end of the quarrel.*

*The main features of the first part of the fairy tale are the man who knows the language of animals or the main character of the fairy tale and the opportunity to learn the language together with its teacher”* ( Aarne 1914:24).

The main parts correspond to the narrative features and thus represent the elements of the middle category. The main features of the parts in turn correspond to the components of the middle elements and thus represent the elements of the lowest hierarchy level.

A significant refinement of this classification in the corresponding terminology goes back to Kaarle Krohn. According to his recommendation, the parts of the story are called "episodes" and the parts of the episode are called "moments" (Krohn, 1926: 29). Although he argued for the use of the term "moment" instead of "motif", the latter was later preferred. Finally, the terms "type", "episode" and "motif" have become established terms for the differentiated presentation of the parts of the story, which are hierarchically segmented according to the working method of comparative fairy tale research (cf. Anderson, 1934/40: 515).[[3]](#footnote-3)

An empirically observable sign that one could use to define the above-mentioned elements is missing in the description of the structure. An exception is the entire narrative, according to which one can define a fairy tale type. For example, if a story is told as an independent text, regardless of the fact that in one case it can be told as part of another story, it is still recognized as an independent type and given its own number.

The formalistic approach in fairy tale research places particular emphasis on the possibility of generalizing individual similar actions in the fairy tale and speaks about a certain number of so-called functions that make up the entire composition of the fairy tale. The physical features in the text, by the help of which we can recognize the presence of these elements, are not established. Indications that these functions correspond to predicates or are associated with concrete protagonists cannot refer to any objective sign. In turn, the generalization of the actions performed by a character may depend on which part or facet of the plot is prioritized by the researcher.

The confirmation of the alternation between the supposed structural elements in the text by an objective sign that regularly occurs in every single text distinguishes the present approach from all previous teachings on the structure of the fairy tale and thus represents other basic premises for the discussion of the same problem.

### **Characters in fairy tales**

Note: The characteristics of the characters are derived from the analysis of the plots in the fairy tale and not the other way round, as is presented in the structure of this paper. The structuring of this paper, however, in a wrong way, aims to save the discussion of the plots in the fairy tale from the need to comment on numerous new terms. This would be unavoidable if this paper were to follow exactly the actual research process.

The identification and categorization of the characters in the fairy tale first required the interpretation of the world order that functions in the fairy tale. Based on the findings of empirical research, the world order known to date in the fairy tale can be described as a hierarchical system of the following actors:

**Head:**

In most cases the head is a king or the head of the family; rarely a deity appears in this position.

**Successor:**

The head of state’s child who is capable of acting is considered the successor.

**Offspring:**

The offspring refers to the representatives of the third generation, initially the child of the successor.

**Equal of the successor who comes from outside the clan or family:**

The peer is a person with the same rights as the successor, but does not belong to the successor's clan or family.

**Priest / Mage / Totem:**

This figure can be distinguished by special features, abilities or skills, such as a beard, age or ugliness. A magician can also be the spirit of a deceased family member. Totems or personified natural phenomena such as wind, frost or sun often have magical powers. Representatives of this category are outside the power of other ordering subjects in the same world.

**Subject / Professional / Official:**

Typical subjects include representatives of dishonest professions such as millers, shepherds, watchmen or porters, but also other professionals such as goldsmiths (the blacksmith, like the healer, is more like a priest), innkeepers, coachmen, small traders; as well as officials, for example ministers, officers, advisers, etc. We should distinguish subjects/professionals/officials as representatives of the class from successors with the same professional attribute. A good example would be the frequent use of a soldier as both a hero and a subject in fairy tales.

**Property / Good / Thing:**

This includes material and/or immaterial inheritance that can function as an object of desire, a tool or a magical agent. Sometimes it can also be a helper, such as a ring or a lamp.

Each member of this order can be represented by **male** or **female** beings, but this does not automatically apply to property, goods and things unless they are personified.

This order applies to both worlds of the fairy tale. From the hero's perspective, these are referred to as his **own** and **other** worlds.

**The function** of the main character in the fairy tale is determined by their belonging to different worlds (their own vs. other people's) as well as by their status in these worlds. Overall, the following three categories of main characters can be identified:

A. Protagonists of your own world

B. Protagonists of the alien world

C. Protagonists of the neutral world (characters who are at home in both worlds).

The strangeness does not necessarily mean hostility; nevertheless, every conflict in the fairy tale reflects the conflict between the representatives of these two worlds.

Each protagonist can behave in a friendly or hostile manner towards his world. For protagonists without a fixed affiliation to one of the worlds, it is crucial whether they are on the side of the representatives of their own world or on the side of the foreign world. This results in two antipodes for each protagonist: the right ( ***r*** ) and the wrong ( ***f*** ).

In the following matrix we try to present and explain the described world order and its concrete representatives in the fairy tale:

The upper half of the x-axis corresponds to your own world (the hero’s world).

The lower part of the x-axis corresponds to the alien world (the world of the antagonist).

To the right of the y-axis is the field for correct action. This means:

* Representatives of their own world act in favor of their own world.
* Representatives of the alien world act in favor of their own world.
* Representatives of the neutral world act in favor of the hero's world.

To the left of the y-axis is the field for wrong actions. This means:

* Representatives of their own world act in favor of the foreign world.
* Representatives of the alien world act in favor of their own world.
* Representatives of the neutral world act in favor of the foreign world.

The y-axis shows the entities involved in both worlds.

The x-axis corresponds to the list of antipodes of the representatives of the members of the world order in the fairy tale. The names of these characters are each abbreviated with two capital letters. The lowercase letters ***r*** and ***f before the abbreviations indicate the*** *right* and *wrong* attributes of the corresponding protagonist.

The initials in capital letters are to be resolved as follows:

HH: Lord of the Hero

HD: Hero

RE: Object of recapture

HF: Helper

ST: Donor

VB: Connecting person

ZM: Magical remedy

BZ: Owner of the target object

ZO: Target object

HP: Master of the potential partner

PP: Potential Partner

AN: Antagonist

The marked cells in the fields and on the x-axis show the relationship between the members of the world order present in the fairy tale and the characters involved in the action, as well as their belonging to their own, foreign or neutral worlds.

|  |  |
| --- | --- |
|  | Representatives of your own world |
|  | Representatives of the neutral world |
|  | Representatives of the foreign world |

Below we look at these figures based on their *true* and *false* counterparts.

Note: There are some reservations for the entire nomenclature of the main characters. The observance of these conventions is essential to avoid linguistic inconsistencies. Main characters (persons, beings, objects) are always in the singular and are referred to by the grammatical gender of the term in contrast to the gender and number of the instance. For example, the term rHD (real hero) is used for the female Cinderella ; the two older brothers are referred to as fHD (false hero), i.e. in the singular and masculine, etc.

**HH -** The hero's lord is the head of the family. If it is a royal family, the lord is the king. This must be distinguished from the king, who is the lord of the potential partner and is not the same as the king whom the hero serves.

When the hero's master behaves disproportionately towards the hero, he is defined as a false hero's master ( fHH ). Examples include: a father coveting his own daughter; he abandons his own son to destruction; or he abandons his children in the wilderness. In another scenario, a king entrusts the younger son with a life-threatening task due to the evil advice of his older sons. In all these cases, the father of the family takes on the role of a false hero's master ( fHH ), thus resembling the antagonist (see below).

**HD -** The hero is the standard for identifying all other characters in the fairy tale. As a rule, the hero is considered the successor, unless there is no master of the hero (HH) in the fairy tale. In the opposite case, if a protagonist in the fairy tale is equal to the head, it is necessary to look for the hero among the other protagonists. Furthermore, there can be both a searching hero and a banished hero. It is not uncommon for two heroes to act in fairy tales. These are: siblings (also half-brothers or half-sisters), an equal and the child of a hero's victim.

False heroes ( fHD ) are usually the hero's siblings, provided they were not born before the hero's activity began. The role of the fHD can also be taken on by the equal. This usually begins with the closure of the brotherhood and ends with betrayal of the hero.

**PP -** The potential partner represents the same class as the hero, but comes from the opposite world. Usually the partner has his own master (HP). The relationship between the potential partner (PP) and the hero (HD) always begins on the initiative of one of the two characters, most often by the male character, even when the male character is a potential partner and not the hero, as for example in ATU 425. Therefore, it is difficult to distinguish a female potential partner from a female target object (ZO, see below).The latter may marry the hero, but is not considered a potential partner for the hero, but rather a prey usually destined for the client, the hero's false master ( fHH ) or the real antagonist ( rAN ). The initiative to win such a figure comes from the rAN or the fHH . It is also difficult to distinguish a potential partner from the object of recapture (RE). One identifying characteristic is the object's marriageability, so it does not have to be a member of the hero's family. For example, in ATU 301, a freed woman is considered a potential partner even if she was previously abducted and the hero follows the king's order or inquiry to retrieve the abducted woman. This object is still considered a potential partner even if the hero does not want to marry her for various reasons, including those mentioned above.

In the case of a false potential partner ( fPP ), these are usually the same-sex siblings (sometimes half-brothers or half-sisters) of the correct potential partner or his peers.

**HP -** The potential partner's lord represents the same class as the hero's lord, but comes from the foreign world. His main function is to set the wedding tasks and general requirements, and if necessary, to pursue the suitors. The false potential partner's lord ( fHP ) acts kindly towards the hero.

**RE -** The object of recapture (RE) can be a person, an object, an animal or a magical tool. It is the object that was present in the hero's world before the hero's activities began and now needs to be retrieved. If the object is taken away, lost or disappeared after the hero's activities began, it retains the status it had before the crisis and does not need to be considered a RE. ( For example, the snake youth in ATU 425 remains a potential partner even if he has disappeared and is being sought by the wife (HD).) The existence of the need for recapture before the hero or helper's activities begin or he is born is important in order to define the sought-after object as an object of recapture (RE). The object of recapture is very active in most cases. The hero is tasked with retrieving the kidnapped mother. She helps the hero to spy on the whereabouts of the antagonist's external soul. The object of recapture belongs to the same world and can be represented by different classes: Successor, for example the siblings who were captured by the antagonist before the hero was born (ATU 328). Head, for example if it is about recapturing the kidnapped mother, who must also have been kidnapped before the hero began his activities. However, not before the hero was born, as is the case in type ATU 650A - The Strong Hanns. In some cases it is difficult to distinguish the RE from the potential partner. The girl from type ATU 301 can be both a bride and a stolen object that must be retrieved. In such cases, attention must be paid to the relationship between the affected protagonist and other protagonists. If a protagonist has his own master and is in the foreign world, this is the potential partner.

The RE can be false if it is directed against the hero. If the hero frees his brothers, who are being held captive by the AN, and they then throw him into the well, the brothers are the False Objects of Recapture ( fRE ) and not fHD (they cannot be these anyway if they disappeared before the hero began his activities).

**ZO & BZ –** The target object and the owner of the target object. The target object can be a person, an object, an animal or a magic tool. This object belongs to a different world than the hero and was not mentioned before. The target object can behave both actively and passively. For example, the real hero ( rHD ) is commissioned by the false master of the hero ( fHH ) to fetch a lion (ZO). The lion resists, but is defeated and taken away. The rHD is commissioned by the real master of the hero ( rHH ) to fetch a horse (ZO). The hero steals the horse, which is inactive but later helps the rHD solve other tasks by acting as a helper.

In most cases, these objects have their owners. The commissioned hero often comes into direct contact with the owner when conquering the object. It is common for the owner himself to be requested as the target object, which subsequently leads to the punishment of the antagonist (typical examples ATU 328, 531). The female figure of the target object can also be distinguished from the potential partner (PP) in that she is not subject to any rule. According to her status, the female figure of the target object is supposed to be a mage or an object or possession of the mage. However, this does not mean that all figures in the fairy tale who are endowed with magical properties belong to the target object category. An example of this is the fairy tale ATU 313, in which the youth ( rHD / rPP ) ends up with the magician ( rHP / rHH ) and falls in love with his daughter ( rPP / rHD ), with whom he later flees from the rHP / rHH . A target object can also be an abstract task, such as plowing a field, fetching wood, collecting customs, etc., which are not considered protagonists of the fairy tale.

A false target object is one that acts in the hero's favor. For example, the organ of an animal is sought and the animal gives the hero its young. The beauty follows the hero without resisting, etc. However, it is common for these objects to later help the hero. In the case of magic objects, it could be a false object if it is exchanged, for example the water of life.

**ZM –** A magic tool can only be defined as such if it does not function as a target object. It does not have to be specifically sought out and acquired, but can be given as a gift, ordered, found or stolen. The magic ring that the hero receives as a reward for his good deeds is not a target object, but a magic tool (ZM). In contrast, the water of life that the hero procures on behalf of his father is the target object (ZO) and not the magic tool, even though it does indeed have magical properties. If a real ring is replaced by a fake one, it is the fake magic tool ( fZM ). The same applies to the case of the exchanged water of life ( fZO , see above).

ZM can also be considered a special type of HF. This can be the case if the ZM is given as a gift by the ST. The gift from the ST is usually considered a ZM if, after the object is handed over, no interaction is developed between the ZM and the HD, but the ZM fulfills various orders (magic ring vs. ghost), transforms into various obstacles, etc.

**ST & HF –** The Donor and the Helper. The Donor can be a human, an animal, a plant or a mythical creature. The Donor provides the hero with a necessary object, a skill or information, but has no direct effect on the hero. For example, the Donor cannot revive the hero or grant him asylum or protection. It is important that the Donor only provides the hero with supportive and not incriminating information. For example, instead of saying that a monster has kidnapped a king's daughter, it would be more appropriate to show the hero the way to the kidnapper or give him a weapon or show him a trick on how to defeat the monster.

The helper (HF) is someone who accompanies the hero and acts together with him. Working together with the hero to solve the task or resolve the emergency is an important factor in distinguishing the helper from the donor. The function of the donor is to help the hero, but in such a way that he does not act as a collaborator. Sometimes the functions of donor and helper can be intertwined in one character, such as an animal brother-in-law from ATU 552. He first shows the hero the way to the antagonist and thus takes on the function of the donor. Then, however, when the hero is killed in a duel with the antagonist, the same brother-in-law brings him back to life and thus fulfills the function of the helper. Depending on his status, a helper can be a mage or a relative of the mage. Almost all protagonists in the fairy tale can donate and help, including the antagonist. The donor or helper himself, however, can do nothing other than donate and help. In the world of mages and totems, there are no classes or hierarchies: A young snake is rescued, and the snake's father gives the rescuer a wishing stone. Another snake is rescued, and it gives the rescuer the ability to understand animal language.

The false donor ( fST ) or false helper ( fHF ) is someone who helps the evil one or gives him something, for example the witch commissioned by a false hero. The beings who keep watch over the evil target (ATU 551) are also considered false helpers.

**VB -** The connecting person is represented by a human or mythical, but more anthropomorphic being. Their main function is to mediate between the hero and another character in the plot or to inform the hero about the existing crisis or necessary burdens. Typically, subjects are used in the role of the connecting person, such as an old hostess, an innkeeper or a shepherd. The hero's master, usually the hero's mother, is also often found in this role.

A connecting person becomes the false one when he acts in favor of the antagonist. Often, this role is played by servants, advisors, etc. who are employed by the evil king. The action of the false connecting person is not always the opposite of what the right one does. Denouncing and taking on the role of an informer or initiating malicious orders are considered typical.

**AN –** Antagonist (opponent, pest): The antagonist belongs to the evil world (foreign world) and can be represented by practically all classes. As indicated above, his function can also be taken over by a false protagonist, but from his own world. Particularly common cases are the false master of the hero ( fHH ) and the false hero ( fHD ).

A false antagonist, i.e. an evil person who acts in the hero's favor, is not encountered but is theoretically conceivable. Several antagonists can appear in a fairy tale. If the dragon is killed and his sister tries to take revenge, the dragon's sister is also an antagonist.

### **Plot sections in fairy tales and their organization**

The segmentation of the fairy tale is based on two empirically recognizable entities: the full story and the motifs created by the appearance of plot-bearing characters. These motifs function as the smallest sub-story within the full story.

In the context of fairy tale segmentation, the importance of the second entity, namely motifs, is crucial. If we take into account that folk tales do not contain detailed portraits, descriptions of nature or depictions of feelings, the universal applicability of the appearance of plot-bearing characters as a metric for fairy tale segmentation becomes undeniable.

The appearance of the main character in the fairy tale is equivalent to the scenes explicitly marked in the drama. If we were to record fairy tales as dramas, we could divide the text into segmented parts that correspond to the motifs. An example of this is the segmented summary of the fairy tale type *ATU 551 - The Water of Life* (see below).

The appearance itself can be presented in the text in various forms, such as:

1. Report about what a protagonist is currently experiencing or suffering from, for example: The king goes blind (King).
2. Report about what is happening between several protagonists, for example: The king instructs his son to guard his grave after his death (king, youngest brother, older brothers).
3. Report on the interaction between two parties and a third party without the latter being actively involved. For example: The youngest brother gives his sister to the suitor as his wife (suitor, youngest brother, the sister does not appear here at all).
4. Report on how two characters interact with each other, for example: The youth kills the monster (youth, monster) in a duel.
5. A special case is the narrative within the narrative, as for example in *ATU 449 -* *Sidi N'Mu'an* .

The list can be expanded.

The motifs before they are integrated into the full story can be organized in the middle structure. This structure is traditionally known as an episode.

The appearance of a leading character in a fairy tale can introduce, continue or conclude an episode. Each episode represents a unit of the following action phases:

Crisis is a situation that needs a solution. Each motif refers to a crisis and represents a phase in its course.

– Krise auslösen) – und – Krise nachvollziehen – und – (Krise bekämpfen – oder – Krise missachten) – und – (Krise lösen – oder

Each motive must be determined according to its **value** , **meaning** and **consequence .**

Depending on the **value,** a motive can be positive, negative or ambivalent.

The **consequence** of the motive determines its belonging to a certain phase of the episode and thus the mutual tension between two motives that are causally linked: *obviousness* vs. *action* (with two results: *solution* vs. *triggering of the crisis* ).

As for the **purpose** of the motive, a motive can have either a *bonus* or *a burden outcome* . In some cases, an outcome can have both purposes (more on that below). A motive with a bonus outcome serves to advance the acting character (it provides them with equipment, items, information, promises, or helpers that can help them in times of need). A motive with a burden outcome, on the other hand, requires the consumption of the collected bonus resources in order to bring the crisis closer to being resolved.

The standard for determining the **value** of a specific motive is the hero's factor alone, not the character acting (this means the action without the hero's participation). The hero's participation in a motive makes it possible to clearly determine the value of the motive based on its effects; motives without the hero's participation are problematic. In such cases, the evaluation of the motive depends solely on the results of the action. If the action has negative results for the hero, then the motive is evaluated as negative. If the motive has no direct result for the hero, then the value of the motive is assessed by its absolute value. For example, if a false hero fails the test of the potential donor, this is a bonus motive with a negative value (the reason for this evaluation is the perspective of the true hero: for him, the same outcome would be the same as a bonus motive with a negative consequence).

Before the **crisis becomes apparent** , it must first have been triggered. The triggering of the crisis is the result of an **action** . This assumes that an action that resolves an existing crisis can simultaneously trigger a new crisis or make it apparent.

The following example: the hero, who is faced with a difficult task, turns to the donor , whom he once helped, and asks him for something in return. This is a manifestation of the crisis, for the solution of which the donor must be engaged in addition to the hero. The donor gives the hero instructions with which he can come closer to solving the task set before him. For example, a trick for capturing a mount. The hero follows the donor's instructions and takes hold of the object, which he later hands over to his client.

If we take a closer look at this chain of actions, we will notice that equipping the hero with instructions is not only seen as a solution to the crisis (fixing the lack of a capture method), but also as an indication of an impending crisis. The capture of the target object (here a horse) is considered to be such a crisis.

A crisis does not exist as an independent motive, but is either the consequence of an action or the obviousness of such an action result.

**awareness** is the situation in which the crisis is made known or recognized. Like anything else, crisis awareness can be negative, positive or ambivalent. Examples:

* The dwarf tests the politeness of the character in question in order to help him. This is a positive intention aimed at helping someone.
* A girl realizes her flaw and becomes jealous of her sister's beauty. This represents a negative intention aimed at corrupting someone.
* The sign at a crossroads contains a warning that must be ignored in order to achieve success. The sign therefore plays the role of an ambivalent manifestation, that which is right is called wrong and vice versa.

The reason for the crisis and the crisis made manifest can sometimes seem completely wrong. For example: The king feels threatened by the hero (the reason for the crisis) and wants to get rid of him. Therefore, he gives the hero an impossible task, which he must complete to die. However, the manifestation of the crisis is not what the crisis is called, such as "You must be eaten by the lion," but rather "You must get a lion's skin." Because the manifestation aims to ruin the hero, it is negative despite its content.

Acknowledging the crisis is followed by an appropriate **action** . A positive action either creates a bonus for further actions or allows the consequences of a negative action to be remedied (resolving the existing crisis). A negative action, on the other hand, leaves the existing crisis in place or triggers a new crisis. Actions with dilemmatic values can do both, such as feeding the bird with one's own flesh or killing one's own children to save the petrified friend.

The action can be postponed. In such cases, the manifestation of one crisis is followed by the manifestation of the next crisis. It can also happen that actions performed within the framework of a crisis are followed by a new action from the framework of the same crisis. Usually, such an action represents an **added value** of the previous action, especially in so-called bonus episodes. Although the added value of the action is established as an autonomous motive, it is indistinguishable from the previous motive in terms of its value and meaning (hence its name - added value). In other words, the result of what and how was acted is completely dependent on the actual action or even included in it. For example: the older brother behaves cheekily towards the dwarf and misses the bonus from him. The behavior of the false hero can be perceived here as both the action and its result. In another variant, the false hero is not only denied potential help, but can be paralyzed or otherwise prevented (here paralysis is considered the added value of the negative action).

The results of the segmentation of the fairy tale into plot sections can be summarized as follows: The smallest empirically recognizable entity for the structural segmentation of the fairy tale corresponds to the appearance of the character(s) carrying the plot, i.e. the scene. The change between the scenes in the story corresponds to the change between the segments of the fairy tale. The smallest category of these segments are motifs, motifs form episodes, which in turn are responsible for the formation of the full story. A motif is either an obviousness or the same as the action(s) of the crisis addressed in an episode and begins and ends in the course of an appearance or a scene.

A scene consists of at least one motif and can contain the motifs of more than one episode. An episode is made up of the motifs that reflect the initial and resolution phases of the crisis it addresses; these are the motifs of the categories 'obviousness' and 'action'.

Motifs from one episode can be separated by motifs from other episodes, resulting in an episode that encompasses practically the entire story. A schematic representation of such a constellation could look like this:

[Story [Episode 1 [Motive 1 . 1 + motif 1 . 2 ]] + [Episode 2 [Motive 2.1 + Motif 2.2 + Motif 2 . 3 ]] + [Episode 1 [Motive 1 . 3 ]]]

Where each plus sign is optional, and each square bracket separated by the plus sign indicates the boundary between two different scenes.

## **Mark-up**

### **tuple *m***

The markup will help us to annotate fairy tales in such a way that all their content properties can be compared using these annotations alone. The development of such a markup is possible due to the maximum compression of the content-structural information. The requirements for the higher expressive potential, user-friendliness and compactness of the markup must not be lost sight of.

***m*** for short) we have identified above , we try to represent it as a 4-tuple of the following attributes:

Attribute ***a*** records the belonging of the motif to the type of story ( ATU Index ), ***b*** records the belonging of the motif to a consequence (action or manifestation) as well as to its value (negative, positive, ambivalent), ***c*** summarizes the content of the motif in one wording, and ***d*** records the nomenclature of the characters involved in the motif.

Here we give the description of the variations of each of these attributes in detail:

#### **Attribute a**

Every motif is part of an episode. Episodes form the full story. So every episode and hence every motif are components of a story type. These are, as is known, recorded in an international type catalog and indexed accordingly. We designate this property of the motif with the lowercase letter ***a*** (for ATU) and the subsequent type number, for example,

*a300* for the type *ATU 300 – Dragon Slayer* .

This time we restrict the set ***a*** to the indices of the fairy tales,

#### **Attribute b**

Marking the belonging of the motif to the specific sequence is possible by appropriate conventional signs:

|  |  |
| --- | --- |
| Character name | The consequence (type of motif) |
| *Ef* | Obviousness of the crisis |
| *Ha* | Action in the crisis |

The variations according to the different value of each of these sequences can be distinguished as follows:

|  |  |
| --- | --- |
| Sign | The value of the consequences (type of motive) |
| F (capital *Ef* ) | Evident crisis with positive value |
| f (small *Ef* ) | Evidence of the crisis with negative value |
| Ff (both *Ef* ) | Evident crisis with ambivalent value |
| H (capital *Ha* ) | Action in crisis with positive value |
| h (small *ha* ) | Action in crisis with negative value |
| Hh (both *Ha* ) | Action in crisis with ambivalent value |

In summary, the set of variations of element *b* looks like this:

We do not mark motifs with double functions such as ***Ha*** *and* ***Ef*** and prefer them according to their primary functions ***Ha*** *or* to be called ***Ef .***

#### **Attribute c**

The content of the action represents a text and is therefore the most informative part of the markup element. This text corresponds to the content value of the action that took place within the framework of an episode phase, e.g. " obtaining a remedy ".

In contrast to other attributes of the tuple, whose variations have already been predetermined, a final completion of the wording list, i.e. variations for the attribute ***c*** *, is* rather unthinkable. There can always be variations that require new wording. For this reason alone, designing a model for the prescribed production of the wording values is one of the most important tasks in building the markup. Without solving this task, standard use of the markup would be doomed to failure.

Thanks to the use of the tuple construction of the markup element, it is possible to use one and the same wording more than once in certain cases. We call this category of wordings the cloned variant.

Cloning or unchanged application of a wording is usually possible by switching between the obviousness as an order and the result of the subsequent action (wrong or right). The variation of the wording is achieved by changing ***b attributes. The application by the d*** attribute values is also possible to further contextualize the difference between the contents covered by one and the same wording.

The following example:

*F:Remedies \_obtain:rHD\_rHH*

***F )*** given by the hero's master to the hero with good intentions to procure him an urgently needed remedy.

By replacing ***F*** with ***f*** and ***rHH*** with ***rAN*** , the same wording could have summarized a significantly different content. Namely, that would be the order given by the antagonist with bad intentions, to let the hero perish while fulfilling the impossible task, in this case, in obtaining the cure. The corresponding markup element looks like this:

*f: obtain remedy:rHD\_rAN*

By further changing the values of attribute ***b,*** it is also possible to indicate the added value of the action that has a positive, negative or ambivalent outcome. The following markup element can be resolved as follows: The hero ends his action to find the cure with positive consequences and hands over the required target object to his agent :

*H:Remedies \_obtain:rHD\_rHH\_rZO*

If an incident were to occur in which the correct water of life was exchanged for the simple water by the wrong hero (here by jealous brothers), this case could also be covered by the same wording, with the values of ***b*** and ***d*** attributes changed:

*Hh:Remedies \_obtain:rHD\_rHH\_fZO*

Another possible use of the clone method is to summarize the first meeting between the hero and his donor or future companion. The latter can be either a helper or a false hero. Such cases usually involve the test of the acting or traveling character and his reaction. The older brothers cheekily answer the question of the little man they met on the way and fail in completing the task. The youngest brother behaves politely towards the little man and thus ensures his help, which can take different forms. The encounter with the dwarf and its process can be recorded, for example, using the following wording *Polite\_Hidden* Prefixes ***Ef*** and ***Ha*** could have indicated the phase of the meeting: *F:Polite\_Behavior* as [veiled] obviousness of the prerequisite for potential help and *H:Polite\_Behavior* as the hero's positive reaction to this prerequisite or *h:Polite\_Behavior* as the negative, rude reaction of the false hero on the same test.

Not all consecutively connected phases of one and the same action can be captured by one wording, i.e. by the clone. In order to be able to maintain the connection between these actions or phases of action even in these cases, a **derivation method** is used. A characterizing keyword takes on the function of an axis, whereby several wordings that follow one another in terms of content remain connected despite their differences, e.g.

h:securities\_impose > H:securities\_impose

Most often, this method is used to summarize such action phases as the initiation and resolution of the crisis, or the manifestation of a predetermined crisis and its occurrence, e.g.

*h: cancel* ***contact*** > *F: restore* ***contact*** > *H: restore* ***contact***

*h* : *causing\_disease* ***>*** *F :* ***resolving\_disease*** > *H :* ***resolving\_disease***

The next method for compactly building the markup is **serialization** of some wording. It is crucial that the action belongs to the same action class. The following action classes are the most common:

* Taboo – Wordings with the keyword Taboo form a list of the obviousness of individual prohibitions, e.g. *" f: Taboo Do not open the door"* . They all have a single continuation, which is: *" h: Taboo breaking follows"*
* Instructions for Action – The character who is hired to help gives the hero instructions on how to act in a situation. Most often, this involves how the hero can fix, grab, or get something.[[4]](#footnote-6)
* Behavior that requires explanation – This keyword applies to the clone wordings that describe questionable behavior of characters. These are usually the characters that the traveling hero encounters and inspires him to search for the corresponding explanation, e.g. *F:Behavior\_of\_fraudsters\_that\_requires\_explanation &* H *:Behavior\_of\_* *fraudsters* .
* Pretense - Like the taboo, pretense in fairy tales is only guided in the sense that it must be followed without fail. So on the one hand we have the list of various pretense as obvious facts and a single wording that applies as a consequence to all these cases of pretense. This is as follows: *H: Pretense \_follow up.*
* Misdeed - The keyword forms the wording for the development of various cases of betrayal. It can connect four categories of wording : (1) misdeed in the planning phase, (2) committing the act, (3) instructions on how to remedy this act and (4) remedying the act. The content of the misdeed is only given in the second category, all others then point to the course of the action in its various phases. The following example:
* resolve – is a common keyword for solving the stressful situation (the crisis). The following example:
* received – is considered a common keyword for the added value of the interaction between the hero and the donor leading to the bonus-giving outcome. The following example:
* seize – summarizes the action in which the target object or its owner is first captured. The following example:
* acquire – is used as the same keyword, which records the communication and fulfillment of an order (but not the initial capture of the target object or its owner). The following example:
* pursue – is considered a generalized form of such behavior as: accepting or rejecting an offer, as well as a correct or incorrect reaction to danger, seduction, provocation or deception. The following example:

The wordings with the keywords *remedy* , *obtain* , *obtain* , *seize* , *pursue* may already be cloned or derived variants.

Another relatively important part of the markup consists of so-called monowords. They usually indicate results or intermediate results of the actions that are captured by the cloned or derived words, or represent the so-called added value of actions that do not come to an end in the course of the story . For example:

*H: Be born*

*H:Tell your own life story*

*H:Marriage ,*

*H:Punishment ,*

*h:Death of a family member*

*h: monstrosity etc.*

Methods such as cloning and deriving the wordings promote the uniform and user-friendly application of the markup developed from them and ensure a high quality of the data annotated with them. The uniform application of the markups is also supported by the wording catalog, which enables the user to make the values recorded in the wording list mutually comparable and easy to find.[[5]](#footnote-7)

#### **Attribute d**

As is already known, there can be a total of 24 characters in the fairy tale genre. Usually one to three such characters can act in one appearance. Acting means that the character appears in the story and not that such a character is mentioned, meant or ordered by a character acting in the episode).

Taking into account the number of heights of the combination between the figures, we had to refrain from recording the order of the figures according to a syntactical function, such as subject / direct object / indirect object. Instead, we reduced the interactions with differently combined and / or identical figures to one combination each.

( ab|ba ) = (ab)

( abc|acb|cab|aba|bac|bca ) = ( abc )

The characters that carry the plot are combined in a fixed order (see the table). This means that the two-part combination is carried out as follows:

* First, the rHD is combined with all the others,
* Secondly, the rHH is also combined with all others, except the already used rHD ,
* Third, the rRE is combined with all others except rHD and rHH , etc.

Thus, the set of variations of ***d*** elements consists of the subsets of figures acting individually, in pairs or in threes:[[6]](#footnote-8)

{

*{ rHD , rHH , rRE , rHF , rST rVB , rZM , rBZ , rZO , rHP , rPP , rAN , fAN , fPP , fHP , fZO , fBZ ... }*

*{ rHD\_rHH , rHD\_rRE , rHD\_rHF , rHD\_ST , rHD\_rVB , rHD\_rZM , rHD\_rZO , rHD\_HP , rHD\_rPP ... }*

*{ rHD\_rHH\_rRE , rHD\_rHH\_rHF , rHD\_rHH\_rST , rHD\_rHH\_rVB , rHD\_rHH\_rZM , rHD\_rHH\_rBZ ... }*

}[[7]](#footnote-9)

### **Sample example**

In the following, we will try to illustrate the connection between the empirical object and the model based on it using an example. To do this, we will segment a classic fairy tale text according to its scenes. [[8]](#footnote-10)Each segment is summarized and provided with the corresponding markup element.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. | Appearance: |  | The king |
|  | Contents: |  | A sick king is dying. |
|  | *Note:* | 1 | *a551 h:Disease :rHH* |
| 2. | Appearance: |  | The three sons and the old man |
|  | Contents: |  | A wise old man points out to the desperate king's sons that the water of life is hard to obtain. |
|  | *Note:* | 2 | *a551:* *F:Remedies \_obtain:rHD\_rVB\_fHD* |
| 3. | Appearance: |  | The father and the eldest son |
|  | Contents: |  | The eldest son asks his father for permission and sets out to find the water of life. |
|  | *Note:* | 3 | *a 551:F : Advertising\_for\_the\_contract:rHH\_fHD* |
| 4. | Appearance: |  | The eldest son and the dwarf |
|  | Contents: |  | The older son shows arrogance towards the dwarf. The dwarf becomes angry with the older son. |
|  | *Note:* | 4  5 | *a 551:F : Polite\_behavior:rST\_fHD*  *a551:h: Polite\_behavior :rST \_fHD* |
| 5. | Appearance: |  | The eldest son and the dwarf |
|  | Contents: |  | The eldest son is locked in a crevice in the rock. |
|  | *Note:* | 6 | *a551:h: Arrest or block:rST \_fHD* |
| 6. | Appearance: |  | The father and the middle son |
|  | Contents: |  | The middle son asks his father for permission and goes in search of the water of life. |
|  | *Note:* | 7 | *a 551:F :Recruiting\_the\_contract:rHH\_fHD* |
| 7. | Appearance: |  | The middle son and the dwarf |
|  | Contents: |  | The middle son also shows arrogance towards the dwarf. The dwarf becomes angry with the middle son. |
|  | *Note:* | 8th  9 | *a 551:F : Polite\_behavior:rST\_fHD*  *a551:h: Polite\_behavior :rST \_fHD* |
| 8th. | Appearance: |  | The middle son and the dwarf |
|  | Contents: |  | The middle son is also locked in a crevice in the rock. |
|  | *Note:* | 10 | *a551:h: Arrest or block:rST \_fHD* |
| 9. | Appearance: |  | The father and the youngest son |
|  | Contents: |  | The youngest son asks his father for permission and sets out to search for the water of life. |
|  | *Note:* | 11 | *a 551:F :Recruiting\_the\_contract:rHH\_fHD* |
| 10. | Appearance: |  | The youngest son and the dwarf |
|  | Contents: |  | The youngest son behaves politely towards the dwarf. The dwarf gives the youngest son the necessary instructions. |
|  | *Note:* | 12  13  14  15 | *a 551:F : Polite\_behavior:rHD\_rST*  *a551: H:Polite \_Behavior: rHD\_rST*  *a551:H:* *Instructions\_Appease guard:rHD \_rST*  *a551: h:Contact \_cancel:rHD\_rST* |
| 11. | Appearance: |  | The youngest son and the guard |
|  | Contents: |  | The youngest son appeases the guard who lets him pass. |
|  | *Note:* | 16  17 | *a551: H:Guard \_appease:rHD\_fHF*  *a551: H:Guard \_escape:rHD\_fHF* |
| 12. | Appearance: |  | The youngest son and the enchanted creatures |
|  | Contents: |  | The youngest son takes the magic objects. |
|  | *Note:* | 18 | *a551: H:Magic \_obtained:rHD\_rZM* |
| 13. | Appearance: |  | The youngest son and the owner of the water of life |
|  | Contents: |  | A love affair develops between the youngest son and the owner of the water of life. |
|  | *Note:* | 19  20 | *a 551:F :* *Instruction\_to\_grasp\_object:rHD\_fBZ*  *a551: h:Contact \_cancel: rHD\_fBZ* |
| 14. | Appearance: |  | The youngest son and the water of life |
|  | Contents: |  | The youngest son obtains the water of life. |
|  | *Note:* | 21 | *a551:H:* *Object\_ grab:rHD \_rZO* |
| 15. | Appearance: |  | The youngest son and the dwarf |
|  | Contents: |  | The youngest son asks the dwarf to release his brothers, but ignores his warning. |
|  | *Note:* | 22  23  24 | *a 551:F :* *Restore\_contact:rHD\_rST*  *a 551:F :* *Avoid betrayal:rHD\_rST*  *a551:h:* *Avoid\_betrayal\_disregard :rHD \_rST* |
| 16. | Appearance: |  | The youngest son and his brothers |
|  | Contents: |  | The youngest son meets his freed brothers and tells them of his success. |
|  | *Note:* | 25 | *a 551:Hh :arrest\_or\_ban\_remove:rHD\_fHD* |
| 17. | Appearance: |  | The kings from distant lands and the youngest son |
|  | Contents: |  | On the way back home, the youngest brother supports the rulers of three different countries with the help of his magical objects. |
|  | *Note:* | 26  27  28 | *a 551:F :Rescue\_of\_those\_in\_need:rHD\_rHF\_rZM*  *a551: H:Rescue\_of\_those\_in\_need :rHD\_rHF\_rZM*  *a551: h:Contact \_cancel:rHD\_rHF* |
| 18. | Appearance: |  | The kings from distant lands and the youngest son |
|  | Contents: |  | On the way back home, the youngest brother supports the rulers of three different countries with the help of his magical objects. |
|  | *Note:* | 29  30  31 | *a 551:F :Rescue\_of\_those\_in\_need:rHD\_rHF\_rZM*  *a551: H:Rescue\_of\_those\_in\_need :rHD\_rHF\_rZM*  *a551: h:Contact \_cancel:rHD\_rHF* |
| 19. | Appearance: |  | The older sons watching their youngest brother sleep |
|  | Contents: |  | The older brothers become jealous and secretly exchange the water of life for ordinary water. |
|  | *Note:* | 32  33 | *A 551:f :* *Planning\_misdeed:rZM\_fHD*  *a551:h:* *Misdeed\_ Appropriate:rZM \_fHD* |
| 20. | Appearance: |  | The youngest son and her father |
|  | Contents: |  | The obtained remedy makes the king even sicker. |
|  | *Note:* | 34  35 | *a551:* *h: procure remedies:* *: rHD\_rHH\_fZO*  *a551: h:disease \_remedy:rHD\_rHH\_fZO* |
| 21. | Appearance: |  | The elder sons the father and the remedy |
|  | Contents: |  | The older sons heal their father and slander their youngest brother. |
|  | *Note:* | 36  37  38 | *a551:* *h: procure remedies:* *:* *rHH\_rZO\_fHD*  *a 551:Hh :resolve\_disease:* *rHH\_rZO\_fHD*  *a551: h:Defamation \_of\_disloyalty:* *rHH\_rZO\_fHD* |
| 22. | Appearance: |  | The older and youngest brother |
|  | Contents: |  | The older brothers silence the youngest brother. |
|  | *Note:* | 39 | *a551:h: Silence :rHD \_fHD* |
| 23. | Appearance: |  | The King and the Executioner |
|  | Contents: |  | The king orders the death of his youngest son. |
|  | *Note:* | 40 | *a 551:f :Death order:rVB\_fHH* |
| 24. | Appearance: |  | The elder son and the executioner |
|  | Contents: |  | The executioner refuses to kill the prince and he flees. |
|  | *Note:* | 41  42 | *A551:H:* *Triggering\_feelings\_of\_compassion :rHD \_rVB*  *a551:H:* *Death order\_ escape:rHD \_rVB* |
| 25. | Appearance: |  | The father, the three foreign kings |
|  | Contents: |  | The three kings whom the prince once helped visit the kingdom. The king learns the truth and forgives his youngest son. |
|  | *Note:* | 43  44 | *a551: H: Restore contact:rHH\_rHF*  *a551:H:* *misdeed\_reveal : rHH \_ rHF* |
| 26. | Appearance: |  | The King and the Executioner |
|  | Contents: |  | The king learns from the executioner that his youngest son has been released and recognizes the slander. |
|  | *Note:* | 45 | *a551: H: Defamation \_fix:rHH\_rVB* |
| 27. | Appearance: |  | The owner of the magic potion |
|  | Contents: |  | At the same time, the owner of the water of life prepares to search for her lover. |
|  | *Note:* | 46 | *a 551:F :Restore\_contact:fBZ* |
| 28. | Appearance: |  | The eldest son and servant of the owner of the water of life |
|  | Contents: |  | The eldest son also fails the identity test. |
|  | *Note:* | 47 | *a551: h:Identity \_test:fHD* |
| 29. | Appearance: |  | The middle son and the servant of the owner of the water of life |
|  | Contents: |  | The middle son also fails the identity test. |
|  | *Note:* | 48 | *a551: h:Identity \_test:fHD* |
| 30. | Appearance: |  | The youngest son and the owner of the water of life |
|  | Contents: |  | The youngest son passes the test and is recognized. |
|  | *Note:* | 49  50 | *a551: H:Identity \_test:rHD\_fBZ*  *a551: H: Restore contact:rHD\_fBZ* |
| 31. | Appearance: |  | Father and eldest son |
|  | Contents: |  | The eldest son confesses the truth to the king. The criminals flee and never return. |
|  | *Note:* | 51  52 | *a551:H: Fix confidentiality\_ obligation :rHD \_rHH\_fHD*  *a551: H:Punishment :rHD\_rHH\_fHD* |

*.*

### **motif, episode, fairy tale**

The organization of the content elements in the fairy tale can be observed empirically on two structural levels: the motif level and the level of the entire story. The first can be recognized by the beginning and end of the scenes in the story, the other by the beginning and end of the entire text.

The full story as the highest instance of the analysis corresponds to a macrostructure that consists of the microelements organized in the mesoscopic structural elements. If this definition is adopted for the full story, the motif is to be defined as the smallest and thus microstructural element of the system. What occurs between the full story and the motif belongs to the mesoscopic structural element. The episodes are to be counted as such.

The mesoscopic nature of the episode automatically implies the ability to represent the character of micro- and macroelements. Hence the possibility arises that an episode can be both a building block and a structure that itself consists of such building blocks.

Using our sample example, we will try to present the connection between the structural elements ***motif > episode > fairy tale*** . To do this, we will present the annotation of the text as a list of motifs (see the annotation values in Table 2) and record the connections between the motifs in episodes.

An episode tells the story of a major or minor crisis. The episode follows a structure that deals with the triggering/evident action, the resolution of the existing crisis or the triggering of a new one.

As already known, the resolution of the crisis is represented by the motive with the attribute values ***b=H/ Hh*** . The triggering or manifestation of the crisis, on the other hand, is captured by a motive that precedes the last one and has a comparable wording (attribute value ***c*** [[9]](#footnote-11)). We qualify this connection between the motives as first-degree causality.

Other motives that relate to the same crisis should be in close proximity to the motives mentioned above and have the same ***d*** attribute value. We call this connection the second degree causality. Lastly, there are the elements attached to the motives of the last category with the same ***c*** attribute value. This is called the third degree causality.

The determination of causality between the motifs of an episode or episode cluster begins from bottom to top. The last motif with the attribute ***b=H/ Hh*** should correspond to a motif in the upper part of the list that has a comparable ***c*** attribute value.

The next step is to identify the motifs that are in the immediate vicinity of such motifs and have the same ***d*** values.

The fact that this research reveals complex clusters of episodes or that certain motifs appear simultaneously in several episodes is completely normal and indicates the fact that the consecutive conclusion of a crisis and thus also of an episode can mean the beginning of a new crisis or episode. The following example:

The hero is faced with the choice of either behaving just as cheekily towards the dwarf as his older brothers *( h: Polite \_Behavior:rST\_fHD )* or, in contrast to them, behaving politely *.* The hero behaves politely, so chooses a positive solution to the crisis *( H: Polite \_Behavior:rHD\_rST )* . As an added value of the solution, he receives instructions on how to appease the gatekeepers so that he can gain free access to the target object. The motive is as follows: *H: Instructions \_Appease\_gatekeepers:rHD\_rST*

Although the motif's affiliation with the episode involving the founder is undisputed, it also shows an equally clear connection with the episode that follows it. This develops in a different place and with the participation of completely different characters. The hero, equipped with the founder's (here the dwarf's) instructions, wins the sympathy of the suffering guard and provides free access to the target object (here the remedy):

*H:Pfortner \_appease:rHD\_fHF*

*H:Pfortner \_escape:rHD\_fHF*

Here are some examples of the episodes identified according to this rule based on the list of motifs in the sample example annotated above:

1. Example:

**First degree causality**

**Second degree causality**

**Third degree causality**

*H:Action instruction \_grasp object:rHD\_fBZ*

*h:Contact \_cancel:rHD\_fBZ*

*H:Object \_grab:rHD\_rZO*

*Q: Restore contact: fBZ*

*h:Entity \_test:fHD\_fVB*

*h:Entity \_test:fHD\_fVB*

*H:Entity \_test:rHD\_fBZ*

*H: Restore contact:rHD\_fBZ*

1. Example:

**First degree causality**

**Second degree causality**

**Third degree causality**

*h:disease :rHH*

*F:Remedies \_obtain:rHD\_rVB\_fHD*

*F:Rescue \_of\_those\_in\_need:rHD\_rHF\_rZM*

*H:Rescue \_of\_those\_in\_need:rHD\_rHF\_rZM*

*F:Rescue \_of\_those\_in\_need:rHD\_rHF\_rZM*

*H:Rescue \_of\_those\_in\_need:rHD\_rHF\_rZM*

*h:Remedies \_obtain:rHD\_rHH\_fZO*

*h:disease \_remedy:rHD\_rHH\_fZO*

*Hh:Remedies \_obtain:rHH\_rZO\_fHD*

*Hh:disease \_fix:rHH\_rZO\_fHD*

*h:Defamation \_of\_infidelity:rHH\_rZO\_fHD*

*F:Rescue\_of\_those\_in\_need :rHH\_rHF*

*H:Resolve slander:rHH\_rHF*

The task of how this rule can be used for the purpose of fully automatic analysis of the fairy tale remains to be done.

## **Machine analysis of the fairy tale**

### **introduction**

Given their simple forms and structures, genres of oral tradition are popular objects for demonstrating the effectiveness of some analytical and generative models of artificial intelligence. Although many of these attempts refer to the scientific heritage of the Finnish School, in particular to the ATU and MIT classification systems, they generally limit themselves to the automation of these instruments without fundamentally reforming their basic method (working method of the Finnish School) in accordance with the requirements of the new approach (methods of digital humanities) ( Declerck et al., n.d.; D'Huy 2019: passim). [[10]](#footnote-12)It seems to be even more popular to formalize the text blocks captured by the structuralist analysis of the fairy tale, so-called Propp functions, and to test their machine recognizability.[[11]](#footnote-13)

In contrast to these attempts, the present work is not about the application of digital methods in fairy tale research, but about the digital expansion of the method of fairy tale research. The most important task of the new approach is the formalized development of the research object (the structure of the fairy tale). It is actually self-evident that a formalized research object must be machine-understandable.

In order to illustrate the articulated difference between the research based on this strategy and the results of the already published papers in the field of electronic fairy tale research, we will try to discuss one of the most recent publications on this topic as an example. It is the article by J. Eklund and others entitled: *"Teaching Tale Types to a Computer: A First Experiment with the Annotated Folktales Collection"* ( Eklund et al. 2023: passim).

New instrument vs. old method. Method developed according to new knowledge.

As the article in question shows, the authors aim to get the computer to predict whether the texts belong to the known fairy tale types. The basis for the classification is the International Fairy Tale Catalogue (ATU) with annotated types. The machine recognition of the type is based on the content properties of the text, which in turn are identified by the content units recorded in the motif catalogue by Thompson (MIT). In summary, the computer is able to ensure the automation of the manual work done in the creation of fairy tale catalogues.

For example, with regard to the concrete text, it consists of a chain of motifs such as B312.2, B11.2.3.1, B11.10, T68.1, D1975, D1978.2, B11.11, H105.1, K1933, K1932, H151.2 and H105.1 and should therefore with a higher probability [[12]](#footnote-14)be assigned to the type *ATU 300 - The Dragon Slayer .*

The efficiency of such an undertaking should not be underestimated. Nevertheless, it is questionable whether it really exceeds the requirements and performance of the applied classification. [[13]](#footnote-15)Automated recognition of the text without critical examination of the legacy of comparative fairy tale research from the time before the digital humanities can be nothing other than the application of digital methods in fairy tale research (see above). As already indicated, we distance ourselves from this approach and try to overcome the weaknesses of the previous method of comparative fairy tale research by expanding this method based on computer-aided solution models.

The reason for the creation of this project lies precisely in the need for a model based on philological analysis that can be used for the machine analysis of fairy tale texts. Formalized description of the structure of the fairy tale opens the way to machine processing of the research object. This means automation both of the segmentation of the fairy tale into structural elements of different sizes and of creating a summary picture based on the comparison of the research objects selected depending on the question, such as: comparative analysis of type X based on repertoires Y and Z. Combination of motifs within episode X. Status of the plot characters in motif Y and much more.

The solution of the machine analysis of the fairy tale as the next step after the formalization of the research object is associated with the development of the appropriate digital infrastructure. Below we will consider some of the most important elements of this infrastructure.

### **Text corpus**

The text corpus is an XML file based on the TEI standard and has the root element < teiCorpus >. [[14]](#footnote-16)The texts within the corpus can be in different languages, but must be standardized according to the Open Data requirements and encoded in Unicode. Targeted access to the texts from different languages or groups can be done using the metadata recorded in the <TEI> element. In this case, it is recommended to use the xml:id attribute on the <Text> element. This attribute consists of four parts that contain information about the license conditions [[15]](#footnote-17), about the belonging of the text to a certain repertoire, the language of the text and the serial number of the text among other materials with the same belonging to the repertoire.

The following examples

* *deu\_deu\_1* = 1. Text from the German repertoire in German
* *deu\_eng\_1* = The same 1st text from the German repertoire in English
* *deu\_eng\_2* = 2. Text from the German repertoire in English etc.

For the names of repertoires and languages, abbreviations according to the ISO 639-3 standard are used.

In the TEI corpus, the fairy tales are recorded individually in separate TEI elements. For each fairy tale, there is a coded text version or a text version made available for coding. The basic texts are located in various repositories and are referenced to each other and to the coded text. The texts that do not meet the OpenData requirements are presented in the text corpus in a derived form and provided with bibliographic information on the primary text source.

The main task of the corpus is to ensure standardized segmentation of the data and targeted access to this data. The data model developed makes it possible to differentiate between scenes in the text and to identify motifs embedded in them. The ***< seg > element*** and the tuples of attributes ***a, b, c*** and ***d embedded in it are*** responsible for this. The optimal number of motifs within a scene is five. This means that two attributes ***b*** and ***c*** can be recorded for each motif, as well as two additional common attributes ***a*** and ***d*** for all five tuples.

An example of annotating the 10th paragraph in our text would be as follows (N stands for zero):

*< seg*

*a 1:ana = “a551“*

*b 1:ana = “F“ c1:ana = “ Polite\_Behavior “*

*b 2:ana = “H“ c2:ana = “ Polite\_Behavior “*

*b 3:ana =“H“ c3:ana=“ Instructions\_to\_appease\_guard “*

*b 4:ana =“N“ c4:ana=“ N“*

*b 5:ana =“N“ c5:ana=“ N“*

*d:ana =“ rHD\_rHF “ >* The meeting scene between the hero and the founder

*</seg>​​*

### **Artificial assistant for semi-automatic annotation of texts**

A central component of the machine analysis of fairy tales is an artificial assistant with a prediction function. [[16]](#footnote-18)This assistant is responsible for recognizing the text sections in a large amount of data that have the desired content properties.

The development of such an assistant requires the fulfillment of the following three tasks:

1. Selecting a suitable approach for machine text analysis.
2. Adaptation of this approach to the research question.
3. Reformulating the knowledge gained about the research object into the language of the selected approach.

The last task is particularly important when it comes to matching the conceptual and empirically observable elements. A scene within the story is considered to be an empirically observable boundary for the beginning or end of a motif. For the machine, this means a character string in which the smallest recognizable element is a word and the largest is a paragraph. On a content level, this corresponds to the text section with changed characters, their attributes, locations and actions. We can teach the machine to recognize this change, but we must first make clear what the connection is between the raw, training, test and target data.

The raw text, as we usually see it, is perceptible through the following explicitly recognizable signs: word, sentence, paragraph and text. However, the boundaries of the text parts organized in this way do not necessarily correspond to the scene change in the text that we are looking for.

Preparing the raw text to form training data requires converting the so-called actual characters into target characters. In doing so, it is often necessary to correct the number of paragraphs in the text. The paragraphs can either be combined or split into several parts. In some cases, a single sentence can correspond to such a paragraph.

Measurement data is the same as raw data, it contains bibliographic metadata and represents well-formed XML data . Target data is annotated measurement data. This can be partially or completely annotated . The already annotated parts of the measurement data are not included in the automatic analysis of the texts. However, they can be used as training data.

For the automatic detection of motifs and episodes in the text, we use the logistic regression algorithm ( LogisticRegression ). A model driven by this algorithm tries to find features applicable to the detection of the sought episodes (positively labeled records) as well as their opposites (negatively labeled records), and arranges them on both sides of an imaginary dividing line.

The decisive factors for embedding the elements in the model are the characteristics of the positively and negatively rated data sets, scaled according to the TF\*IDF values. Before the TF\*IDF values are calculated, the data is cleaned, whereby so-called stop words and other words that are irrelevant for the comparison are removed from the data sets.

Due to the typically limited number of positively labeled training data, a multiple cross-validation method is applied. This method divides the entire training data into equal sections that are used as validation and training data. The trained model should now be able to separate the characteristic features (word forms) for positively and negatively labeled data and represent them as a sorted list of the corresponding character groups.

The trained model is extended by the prediction function. [[17]](#footnote-19)Although the function fails in some cases, it can still be considered a reliable tool. The exact reason why the prediction function fails in some cases is not clear, but it is likely that this is partly due to the method of matching the similarity between the model and the measurement data. The sigmoid function, which can be successfully used to calculate the distance between the model and the measurement data, might be hampered in this process because the dataset often represents both classes (both positive and negative). In such cases, the assumed similarity or difference between the compared contents might be erroneous because both negative and positive features can be present simultaneously in the text part being checked.

Taking the above factor into account, it seems to be much more effective to examine the data only based on the positive features with the highest coefficients. This means that the data sets are no longer compared using the model, but using the prototypical meta-motif extracted from the model. This works like a search query to the document collection.

Ein Bild, das Screenshot, Reihe, Rechteck, Diagramm enthält.

Automatisch generierte Beschreibung

Fig. 6

The features are extracted by calculating the threshold value for the top classification features. This value is to be found at the point in the feature chain where the weights of the positive and negative features or coefficients are compared.

Using the graphical method, this point can be located as follows: Figure 6 shows the oscillation between the maximum and minimum coefficients of the negatively and positively assessed classification features. The maximum number of features (z) applicable to both categories is a variable and can be adjusted according to the situation; this time it is 200. Blue columns represent the positive features, while red columns visualize the negative features (due to the large number, the features overlap (see black word cloud in the lower part of the graphic). To visualize the desired point, the existing graphic is divided at the meeting point of the red and blue columns and the left half is moved to the right until the smallest red column is placed under the largest blue column. If you pay attention to the size of the blue and red columns, you can easily determine where they coincide (see Figure 6.1).

Without a graphical representation, the same threshold can be calculated using the following formula:

The equation is:

, *where ,*

the threshold is equal to *.*

Since the top classification features represent nothing more than a simple list of positively rated elements or words, they can be formed using a simple step function (corresponds to the green graphic in Fig. 6.1):

Ein Bild, das Screenshot, Reihe, Rechteck, parallel enthält.

Automatisch generierte Beschreibung

Fig. 6.1

The list of characteristics defined in this way is stored as a prototypical meta-motif in a data set and added to the measurement data.

The recognition of the desired motif in the measurement data is carried out by classification based on cosine similarity. [[18]](#footnote-20)Each dataset is compared with a synthetic dataset, which we defined above as the prototypical metamotif.

### **Testing the Assistant**

To test the assistant's prediction function, we will conduct the following experiment. First, we will create a text corpus from data that can be assigned to the fairy tale genre with a high degree of certainty. For this, we will use the German-language fairy tale collections published in the TextGrid repository.

We select the fairy tale texts by Jacob and Wilhelm Grimm (215), Heinrich Pröhle (150), Karl Bartsch (52 + 47) and Ludwig Bechstein (81) and combine them in an XML-TEI corpus. In addition, 12 further texts from the collection "European Fairy Tales" are included. For legal reasons, these texts are presented in a derived version. In total, our corpus contains 557 texts.

Now we select an example of a fairy tale type ATU 300 – The Dragon Slayer from the collection of Karl Bartsch and annotate the text using the markup for the content annotation of the fairy tale.

The selected text represents a complex fairy tale consisting of four type parts. After the initial episodes from types ATU 567 (from eating the extraordinary bird heart to escaping) and ATU 303 (raising two similar-looking brothers to the brothers' separation and exchange of life indicators), the episode of type ATU 300 - The Dragon Slayer begins. This includes the content from the inquiry about the human sacrifice to the marriage of the rescued princess. The text continues as type ATU 303. It reports how the youngest brother is petrified by the witch, but is then rescued with the help of his brother.

For the experiment, we will only extract the text part of type ATU 300 in this fairy tale, i.e. label it using the markup. This means that the model is only trained using positively labeled data sets. As part of this process, the tokens are extracted from all labeled data sets and a vocabulary is created.

The vocabulary is cleaned of so-called stop words and other words that are not essential for the comparison. The latter are considered rare words that only occur sporadically in the data sets and therefore play no role in comparing these data sets with others. The parameter min\_df = n is responsible for cleaning the vocabulary of isolated words. Given the small number of data sets and the short list of tokens found in these data sets, the value of the min\_df parameter is set to the minimum value ( min\_df = 2) in our experiment. This means that every word recorded in the vocabulary must appear in at least two different data sets.

Next, the word forms present in the vocabulary are scaled according to their TF\*IDF values. The TF (Term Frequency ) of a token is calculated as:

Frequency of the token in a data set / total number of tokens in the data set

The IDF (Inverse Document Frequency ) of a token is calculated as:

Log e (total number of records + 1 / number of records with the searched token + 1) + 1

rescaled using the L2 normalization procedure . The coefficient thus calculated determines the value of each individual token in the vocabulary. The ranking of the tokens in the vocabulary makes it possible to effectively correct the number of tokens in the vocabulary.

Segmenting the training data into natural plot segments is particularly important for the approximation between the created vocabulary and the model of a specific story. It ensures that words that occur repeatedly in two consecutive plot segments (obviousness of the crisis - confrontation with the crisis) are represented with separate data sets. This gives them higher TF\*IDF weights or even includes them in the vocabulary.

In the next step, a measurement data corpus is aggregated from the remaining (unlabeled) data records in the text corpus. The previously created model data record is also inserted into this measurement data corpus. Then all other data records from the measurement data corpus are compared with the model data record to predict which of them are closest to it.

The procedure includes the following steps:

1. Creation and cleaning of the vocabulary: The vocabulary is cleaned of stop words and other words that are not essential for the comparison.

2. Calculation of TF\*IDF values: The TF\*IDF values are calculated for each word in the vocabulary.

3. Creation of sparse matrices: For each data set, a sparse matrix is created that represents the TF\*IDF values of the contained tokens.

4. Comparison of the matrices: Each of these matrices is compared with the matrix of the model dataset.

The calculation of the similarity between the data sets is based on the cosine similarity formula, which is defined as follows:

Here , ***a i corresponds*** to a feature (token) from the model data set scaled in a TF\*IDF value, a ***b i*** corresponds to a feature (token) from the arbitrary measurement data set scaled in a TF\*IDF value. The closer the value of the calculated cosine similarity is to one, the more plausible the predicted content similarity between ***a*** and ***b*** data sets is. This method makes it possible to quantify the similarity of the data sets and to identify those that are closest to the model data set.

The data sets with the highest coefficients, i.e. the most probable predictions, are at the top of the list. Using Hans-Jörg Uther 's reference work , the "German Fairy Tale Catalogue", it is easy to check whether the algorithm has found all texts or parts of them contained in the text corpus that belong to the fairy tale category being sought (this refers to type ATU 300 - The Dragon Slayer).

The search in the catalogue shows that all the relevant texts were indeed reliably recorded by the algorithm. They are as follows:

|  |  |
| --- | --- |
| Text in the catalogue | Text ID in XML text corpus |
| Bartsch 1879f. I, 474ff. | @n=“dummy20“ |
| Bechstein/ Uther 1997 I, No.49 | @n=“dummy122“ |
| Grimm/KHM (1857) No. 60 | @n=“dummy484“ |
| Pröhle 1853, No. 4 | @n=“dummy234“ |
| Pröhle 1853, No. 5[[19]](#footnote-21) | @n=“dummy256“ |

The method shown in this experiment is particularly useful when creating the training data. With the training data obtained in this way, it is possible not only to make the prediction more precise, but also to focus it on smaller content elements. To demonstrate this property, we conduct another experiment.

We train the model on the previously labeled datasets, including the above-mentioned texts of type ATU 551 "The Water of Life". [[20]](#footnote-22)The ID characters of these texts in the text corpus begin with the suffix " zyx \_". After aggregating a provisional text corpus from these and the datasets labeled in the first experiment (datasets with the ID suffix " dummy "), we can retrieve and view all the markup elements present in this corpus.

Next, we select one or more markup elements from this list and create a corresponding query. Suppose we are looking for the motive of the father's obvious illness. The corresponding query consists of the attributes "a551" and "Illness:". The colon after the word "illness" ensures that any further wording such as " fix\_illness " is ignored in the search.

The result of the subsequent machine analysis confirms the high precision of the prediction function used. The first hit corresponds to the part of the text with the motif sought. In addition, the difference between the coefficients of the first and the following data sets is so great that they must undoubtedly represent two contrastingly different predictions. In fact, as the above-mentioned catalogue confirms, no other texts with the same content are to be expected in the text corpus.

### **Tools for evaluating the data**

Content annotation makes it possible to evaluate a large number of texts in a uniform manner and with a precise focus on the specific research questions, and to present the results in a comprehensible and clear manner. Here we look at an application example with the aim of evaluating the normal form of a type in Western European cultural areas in a visualized form.

Before the evaluation, it must be predetermined (a) which type or type cluster we are looking for and (b) in which part of the data the search should be carried out.

Suppose we are interested in the standard form of the fairy tale type *a551 – Water of Life* in the European fairy tale repertoire [[21]](#footnote-23)(German, Austrian, French, Italian, Hungarian, Greek, Russian).

After entering the appropriate search parameters, the assistant accesses the annotated fairy tale texts with matching identifiers and presents the evaluation results in the form of graphs and a simple list of the motifs involved in the composition.[[22]](#footnote-24)

The graph shows a combined image of the motifs in the various texts that are linked to one another with the desired ***a*** attribute value. At the peripheries of the chain, there may also be some motifs that have other ***a*** attribute values. These in turn represent the peripheries of the motif chain that make up the type combined with the desired type. [[23]](#footnote-25)After entering the values of such ***a*** attributes, it is expected that the graph of the relationships of motifs within a type will be converted to the graph of the relationships of motifs within a type cluster.

The evaluation of the motif inventory in a type or type cluster is automatically followed by the aggregation of the library of motifs involved. This can, unless requested, represent a simple series of motifs without any indication of any connection. After entering a specific motif, the list is reorganized as follows: the entered motif is given the number zero and is placed in its place in the series of all motifs, the motifs ahead and behind it on the time axis are given corresponding numbering with minus and plus signs and take the place above or below the motif marked zero. If the chain of motifs of a type is separated by one or more motifs with a different ***a*** attribute, they are recorded without any indication of their position in the motif chain. This gives us a list of motifs that matches the summarized content of the evaluated type.

Given the confusion caused by the variations in motifs caused by the change of characters, it is recommended to condense the summary picture of the fairy tale's composition into a few elements. In this sense, it seems effective to remove the last attribute from the annotated markup element.

Example: In type a551 (=attribute a), the order "F" (=attribute b) to obtain the water of life (=attribute c) can be given to different protagonists (attribute d), be it the sick father ( rHH ) or the old man giving advice ( rST ). This automatically leads to two markup elements being created when summarizing essentially the same action :

a 551:F :obtaining\_remedies:rHH

a 551:F :obtain\_remedies:rST

By removing the attribute ***d,*** only one markup element remains, which, despite the loss of information, is able to convey a common picture of the chain of actions. The corresponding code can be found in the GitHub repository.[[24]](#footnote-26)

Ultimately, the scientific value and plausibility of such an evaluation depends entirely on the degree of representativeness of the research data collected. In this publication, we limit ourselves to demonstrating how the code works. The results obtained using the code on the Caucasian narrative material will be published separately.

## **outlook**

This work examines general rules for the organization and combination of content elements in all levels of the fairy tale: motif, episode, full story and gives the formalized description of the recognized orders. Based on the conceptual model of the genre thus captured, I design a machine learning model that allows us to automate the most costly and time-consuming tasks of the comparative fairy tale.

A particular added value of the research results achieved lies in the provision of an artificial assistant for the recognition of content-structural elements within an unmanageable amount of text data. The merging of the performance of two independent algorithms into a single one is particularly noteworthy. Only through this approach is it possible to successfully use the prediction function, even if the positive and negative classification features separated according to the model are kept together in a text section under investigation.

An important part of the infrastructure provided in the course of this work is the module for the visual representation of the evaluated normal forms of individual types. The application of the visualization method using digraphs seems to be particularly effective. The resulting picture provides a clear and complete idea of the popularity of individual types and the frequency of their mutual combinations.

The results of this study are currently being applied in practice to contribute to the development of North Caucasian folklore. The project, funded by the DFG, aims to create an electronic data corpus and, by annotating the content-structural properties of the texts collected in this corpus, to make them available for international comparative fairy tale research. The following are considered to be particular challenges:

1. linguistic diversity of the material [[25]](#footnote-27),

2. Limitations on the application of the Open Data Principles.[[26]](#footnote-28)

After only one and a half years of applying the theoretical and practical research results contained in this book, their sustainability is clearly evident. Particularly noteworthy is the ability to adapt to new development requirements. This is evident in the switch from CSV data collection to an XML data corpus and in the transition from basic segmentation procedures from episodes to the motif level.

With continued use of this working method, it is easy to predict that a special synergy effect will be achieved. As already indicated, comparative fairy tale research develops successively, following the model of a self-regulating system. This means, among other things, that each newly developed repertoire encourages the comparative analysis of another repertoire; both in turn contribute to the development of the next one, and so on. Taking this aspect into account, the additions to the individual infrastructural elements, such as markups or synthetic training data, are considered as an incidental contribution to each new use of the working method.

1. Anderson, Walter (1934/40): Geographical-historical method. *Dictionary of German fairy tales,* edited by Lutz Mackensen. Vol. 2, Berlin 1934/40, pp. 508-522; Röhrich, Lutz (1987): Geographical-historical method. In: *Encyclopedia of fairy tales.* Vol. 5, Berlin/New York, col . 1012-1030. [↑](#footnote-ref-1)
2. Propp , Vladimir J. (1975): *Morphology of the Magic Fairy Tale* , ed. Karl Eimermacher, Frankfurt am Main, Suhrkamp. [↑](#footnote-ref-2)
3. According to Anderson, the narrative is divided into episodes, which could also be called acts or chapters. "Each episode consists of several individual motifs or moments, which in turn contain a series of features It should be emphasized that the term train, of whose rows the motif consists, includes nothing other than variations of individual motifs. For comparison see Anderson 1923: 242-260. [↑](#footnote-ref-3)
4. Instructions for action must be clearly distinguished from a test position. This only occurs when the relationship between the hero and the donor or helper is clarified and the hero can enjoy guaranteed support from the donor. [↑](#footnote-ref-6)
5. For the full list of attributes see GitHub: edadunashvili / VerMa / kf / vmf\_c1-5.xsd [↑](#footnote-ref-7)
6. full list of elements see GitHub [↑](#footnote-ref-8)
7. For the full list of attributes see GitHub: edadunashvili / VerMa / kf / vmf\_d.xsd [↑](#footnote-ref-9)
8. Water of Life, KHM 97; cf. Blécourt , Willem de ( 2014): Water of Life. Encyclopedia of Fairy Tales. Col. 509-514. [↑](#footnote-ref-10)
9. This refers to cloned and derived wordings that make up the *c* attribute values (see the section *Attribute c* ) [↑](#footnote-ref-11)
10. to one of the youngest [↑](#footnote-ref-12)
11. The most recent attempts include the publications by Finlayson (2012, 2016) and Lendvai et al. (2010a, 2010b). [↑](#footnote-ref-13)
12. For the resolution of the evidence listed here, see MIT [↑](#footnote-ref-14)
13. This refers to the above-mentioned classification models ATU and MIT [↑](#footnote-ref-15)
14. Github : edadunashvili / VerMa /Textkorpus.xml [↑](#footnote-ref-16)
15. For this purpose, abbreviations cc are used for the released data and cr for the texts protected by copyright. [↑](#footnote-ref-17)
16. GitHub: edadunashvili / VerMa /erthaos\_11.ipynb [↑](#footnote-ref-18)
17. GitHub: edadunashvili / VerMa /erthaos\_11.ipynb ( cells #14-1). [↑](#footnote-ref-19)
18. Needham, Mark 2016 is considered exemplary . [↑](#footnote-ref-20)
19. The text represents a combination of types 303 and 300. In the Uther catalogue only recorded in the type ATU 303. [↑](#footnote-ref-21)
20. European Fairy Tales. CD Edition [↑](#footnote-ref-22)
21. Repertoire corresponds to the texts collected in the electronic (CD) edition of “European Fairy Tales” under the type number ATU 551. [↑](#footnote-ref-23)
22. GitHub: edadunashvili / VerMa / Evaluation\_abcd.ipynb [↑](#footnote-ref-24)
23. If a member of the searched motifs represents the beginning or the end of a text, this is marked by extra characters “ Anf- ” and “End-”. [↑](#footnote-ref-25)
24. GitHub: edadunashvili / VerMa / Evaluation\_abc.ipynb [↑](#footnote-ref-26)
25. About 25 indigenous languages are spoken on the present territory of the Federal Republic of Dagestan. [↑](#footnote-ref-27)
26. The only way out of this situation is to convert the original texts into derived research data. We often use texts that have been cleaned of stop words. References to the source data are always routinely verifiable when such derivations are released. [↑](#footnote-ref-28)