Decoding Southern African Folktales using *Hylistic* Analysis

Pannach, Franziska
Center for Language and Cognition (CLCG)
University of Groningen
f.a.pannach@rug.nl

Abstract

In this submission, the potential of *hylistic analysis* (Zgoll 2019, 2020) for the domain of Southern African folklore is demonstrated. For that purpose, narrative sequences of folktales are carefully re-constructed according to the new *hylistic* theory and analysed on a subset of 30 folktales from Henry Callaway's *Nursery Tales and Histories of the Zulus*, a unique parallel resource in English-vernacular isiZulu.

The dataset, consisting of 5176 distinct narrative statements, the so-called *hylemes* in 384 narrative sequences, is presented as a freely accessible resource which holds transformative potential for the re-interpretation and (automatic) processing of folktale narratives. It is the first step towards a thorough investigation of Callaway's collection, and an effort to the analysis of the tales' narrative structures. The dataset can be used as a resource for many future studies on folktale characters and plots as well as the study of historical isiZulu.

Keywords: Hylistic Analysis, Folktales, Dataset, Narrative Sequences, Digital Folkloristics, Vernacular Language, isiZulu, Topic modeling, Frame Semantics

1 Introduction

Digital Folkloristics is the study of folklore materials, including folktales, using digital tools and computational methods. It is a broad field that spans from the ethnographic study of born-digital materials (Tolbert & Johnson 2019) to folklore corpus construction and annotation, e.g. by Hagedorn & Darányi (2022), the digitization of motif indices, e.g. by Karsdorp et al. (2015), and the computational analysis of and tool-building for folkloristic materials in all their various forms.

In this work, I show how the novel *hylistic* approach (Zgoll 2019, 2020) can be applied to annotate narratives in Southern African folktales, on the example of Henry Callaway's collection *Nursery Tales and Histories of the Zulus*. The annotation is still a largely manual task; however the resulting data has immense potential for the computational study and further analysis of narrative materials (*Stoffe*) (Zgoll 2020).

After an explanation on how narrative events are annotated from the source text, this contribution presents two case studies on the application computational tools on data in *hylistic* format, compared to the original discursive text. First, the original text and the *hylistic* structure are processed using a frame semantic parser to see how the narrative events and their participants can be extracted from both representations. The output of the two approaches is compared. In the second case study, different topic modeling techniques are used to discover and investigate latent topics in the folktale sequences.

The paper is structured as follows: Section 2 frames this paper within related research. Section 3 explains how *hylistic* analyses is performed, Section 4 describes Callaway's tale collection and introduces the *hylistic* annotation of the data. The data set is then used in the two case studies mentioned above in Section 5. The paper ends with a discussion in Section 6 and an outlook on future work in Section 7.

2 Related Work

Folktales have been a target of study in Digital Humanities for many years, e.g. through semantic modeling by Declerck et al. (2012), Karsdorp et al. (2015), Declerck et al. (2016) or linguistic analyses (Karsdorp & Fonteyn 2019). The subdiscipline is often referred to as *Computational Folkloristics* (Abello et al. 2012), although folkloristics usually involves other research objectives as well, such as the study of folk song and performance, traditional dress, or idioms and vernacular expressions.

The annotation of structural features of folktales according to the formalistic approach of Propp (1968) has been performed manually, e.g. by Azuonye (1990), Okodo (2012), Lendvai et al. (2010), Pannach et al. (2021), in parts automated (Finlayson 2016, 2012) and attempted with assistance of large language models (LLMs) (Gervás & Méndez 2024). Finlayson showed that very complex narrative structures, such as Proppian functions, can be learned with carefully modeled pipelines, if high quality human-annotated data are available. (Finlayson 2017)

Frame-semantic parsing, in conjunction with other NLP technologies and sequence alignment algorithms, has been used to extract and align narrative structure from folktales and ritual description by Reiter et al. (2014) (similarly Pannach (2023*b*)).

However, automatic approaches towards African folktales, especially with regard to their narrative structure, are scarce. Annotated corpora often focus on different phenomena, e.g. on discourse (Aplonova et al. 2023) rather than the events that take place in a story.

3 Hylistic Analysis

The Hylistic approach originates in mythological research (Zgoll 2020) and has been largely applied in this discipline, for myths from Classical or Ancient Near Eastern sources so far (Zgoll & Zgoll 2020, Zgoll et al. 2023). However, as this article aims to demonstrate, it can be applied for the analysis of any form of narrative, fictional or non-fictional.

The output of a *hylistic* analysis is a so-called *hyleme sequence*, where a sequence represents one variant of a narrative material (or *Stoff*), as present in a resource. The medium of the resource can be textual, but also an object (e.g. a sculpture), an image (e.g. a comic strip or vase painting), or a oral version of a retelling. For the application in comparative folktale research, where oral and visual transmission play a vital role, the *hylistic* approach is therefore particularly well-suited.

Each hyleme sequence consists of elements called *hylemes*. A *hyleme* is a narrative statement which represents a vital element of the content, e.g. an event or a state. The order of *hylemes* can be either the logical chronological (fabula) or the order in which it is represented in the source (syuzhet).

Hyleme functions encode whether a *hyleme* describes a *single-event* or a state (*durative*). Thus a *hyleme* can have one of four functions: *durative-initial*, *durative-resultative*, *durative-constant*, or *single-event* (standard hyleme). The functions are visualised in Figure 1. *Durative* hylemes are true over a certain "distance" in the sequence, either during the entire sequence (*durative-constant*), or only until a certain point (*durative-initial*), or from a certain point onwards (*durative-resultative*). Both *durative-initial* and *durative-resultative* types are therefore context-sensitive. The values of the function of the *hyleme* depends on the semantic information of another hyleme. For instance, if a hyleme "The king marries a young woman" (*single-event*) is present, the hyleme "The young woman is the queen" is *durative-resultative*. It only becomes true at a certain point in the narrative. In the same manner, the hyleme *The king is happy* in Table 1 is a result of the queen giving birth, and does not end within the scope of the sequence. It too, is thus *durative-resultative*.

A statement that is true at one point during the narrative is classified as *single-event*. This can for instance be an action like *opening a door*. The annotation of *hyleme* functions can be performed with overall very good inter-annotator agreement (Pannach 2023*a*). The following sequence (Table 1) gives examples derived from the annotated data for each of the functions.

Table 1: Examples for hyleme functions

Function	**Example**
durative-initial	"A woman is pregnant."
single-event	"The woman gives birth to a child."
durative-constant	"The woman is a queen."
durative-resultative	"The king is happy."

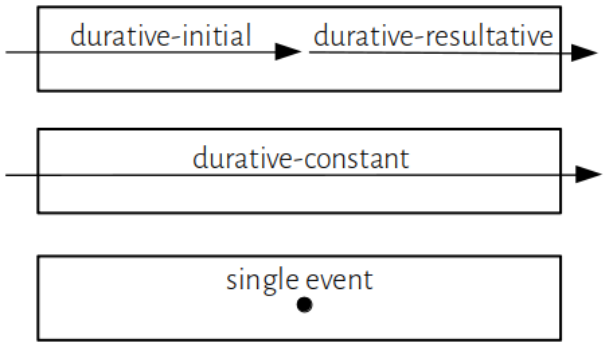


Figure 1: Visualization of hyleme functions

Hyleme Component	**Example**
Hyleme subject	the man
Subject det.	old
Hyleme predicate	pull
Predicate det.	with all his strength
Hyleme object	the root

A *hyleme* consists of three basic components, the hyleme subject, predicate, and object. A hyleme has at minimum a subject and a predicate. If a hyleme has multiple participants, this might result in multiple hyleme subjects, e.g. "The woman, the husband, and their children go to sleep." In this case the hyleme has three subjects. Therefore, the term *subject* does not strictly refer to the grammatical function of the subject. Hylemes are always in active voice, and present tense. Hyleme sequences are *not concurrent* text, and thus do not use discord markers and cross-hyleme co-references. Instead, proper names are always repeated.

Important information that is not directly stated but implied, that can be derived with confidence, is indicated by square brackets, e.g. "Unthlatu asks what Untombinde is doing [in his father's kraal.]"

Each hyleme component can be further specified by using *hyleme determiners*, e.g. adjectives or adverbs/adverbial clauses. Thus, the hyleme "The old man pulls the root with all his strength." can be split into the following hyleme components:

In an ideal setting, each sequence (per context-window) would be extracted by multiple extractors at the same time, and compared and discussed among domain experts. In the case of the presented dataset, the sequences are the interpretations of the author. Due to the nature of the tales being close to an orally narrated version (e.g. often accompanied by repetitions and addresses to the audience), most of the narratives in Callaway's collection are straightforward and thus easier to follow than others, e.g. mythological narratives based on ritual descriptions. In the latter case, there is often much more room for interpretation due to certain aspects not directly mentioned in the source or missing, and they require very detailed expertise, e.g. in Ancient Sumerian and Mesopotamian cultures, and expert-driven discussions.

In the context of this study, *hylistic analysis* was chosen over other narrative modeling and analyses techniques due to two reasons: Firstly, unlike other narrative theories, like Proppian Morphology, *Hylistics* shows a greater versatility across domains (i.e. it can be easily applied to new data as long as it is narrative in the broader sense). It generalizes well and does not come with a priori assumptions on narrative structure.

Secondly, the resulting sequences model core narrative structure without focusing solely on the text (i.e. also including implied narrative components). Other theories and annotation schemes for narrative events, especially those that are easier to automate, naturally have to focus on what is present in the text, potentially missing essential aspects.[1]. For example, we can imagine a folktale like *Snowwhite and the Seven Dwarfs* told as a monologue of the main character. As human annotators, we would recognize the character speaking, while extraction algorithms might not be able to identify the character.

4 Data and Approach

Hyleme sequences are extracted manually by a domain expert in a series of steps. First the domain expert identifies a relevant source, e.g. a text and the context window that contains the narrative variant. Secondly, hylemes are extracted in the standardized format, as described above. The extraction steps are visualised in Figure 3. After the identification of the context window, and the subsequent extraction of hylemes, they are ordered logically (*fabula*; *Stoff*-chronologically). The resulting *hyleme sequence* then represents one narrative variant of a *Stoff* (see Zgoll (2020)).

In order to facilitate automated comparability between the sequences, certain narrative events were annotated in a standardized way (without using a specific controlled vocabulary), which I strongly recommend. For instance, *return* events were always annotated as *CHARACTER returns to PLACE*, even if the text used different verbalisations such as *return, come back, go back to, arrive back at* and so on. Direct speech was largely avoided, unless the exact phrasing was of significance (e.g. in chants or forms of magic spells), instead they were annotated as communication events *CHARACTER says that .../CHARACTER answers that...*

In total, 383 hyleme sequences have been constructed for 30 tales in Callaway's collection. As expected, the hyleme labels are unequally distributed. Most hylemes (4176) are regular hylemes corresponding to single-events, which mostly encode important actions, such as hunting, leaving/returning, or marrying. Of the *durative* hylemes, 535 are *durative-constant*, 322 are *-resultative*, and 143 are *-initial*, see Figure 4. The sequences contain 68 distinct named entities (50 PER, 4 LOC, 14 MISC). The sequences range in length from one hyleme to 88 hylemes (depending on the length of the text in the context-window, here: between the subheadings).

It has to be noted that some of the original tales, and especially Callaway's original comments and footnotes contain terms and language that are not acceptable today, e.g. referring to people. Those terms have not been used in the hylemes. The data is publicly available (in csv format) under a Creative Commons license (CC-BY-SA 4.0)[2]

5 Application of the Data

Data in hylistic format can be applied in various studies in Computational Narratology. Its main advantage is that narrative events and states in *hylistic* format are easier to model and harmonize than in concurrent narrative texts. Especially in the literary domain, events are often not communicated directly. Events or event participants may be implied, phrased as direct or indirect speech/thought, or communicated through figurative language, such as similes "[...] as pleasant as a grain of sand in the eye [...]" (Eliot 1994). Literary texts are therefore complex to automatically process with regard to narrative structure and participants, even if the text is in modern, standard English, and analyses are even more difficult in other languages, especially low-resource languages.

5.1 Case Study 1: Frame Semantic Parsing

Due to the standardized structure of the hylemes, frame semantic parsing yields more informative results compared to the application on the original text. For demonstration purposes, I apply frame semantic parsing to the following original quote (Callaway 1868):

"Uthlakanyana opened the door. (1) "Are you still angry, my uncle? (2) Do you no longer cry out so as to be heard; for I thought you were screaming? (3) My uncle, speak. (4) Why are you silent? (5) Just play your calabash, that I may listen and hear." (6) At length he entered ; when he came, the cannibal was dead. (7) He took him out of the house, and took possession of it. (8) He slept, and was happy now. (9)" (Callaway 1868, p.34)

The paragraph can represented using three hylemes:

1. Uthlakanyana opens the door.
2. Uthlakanyana mocks the cannibal.

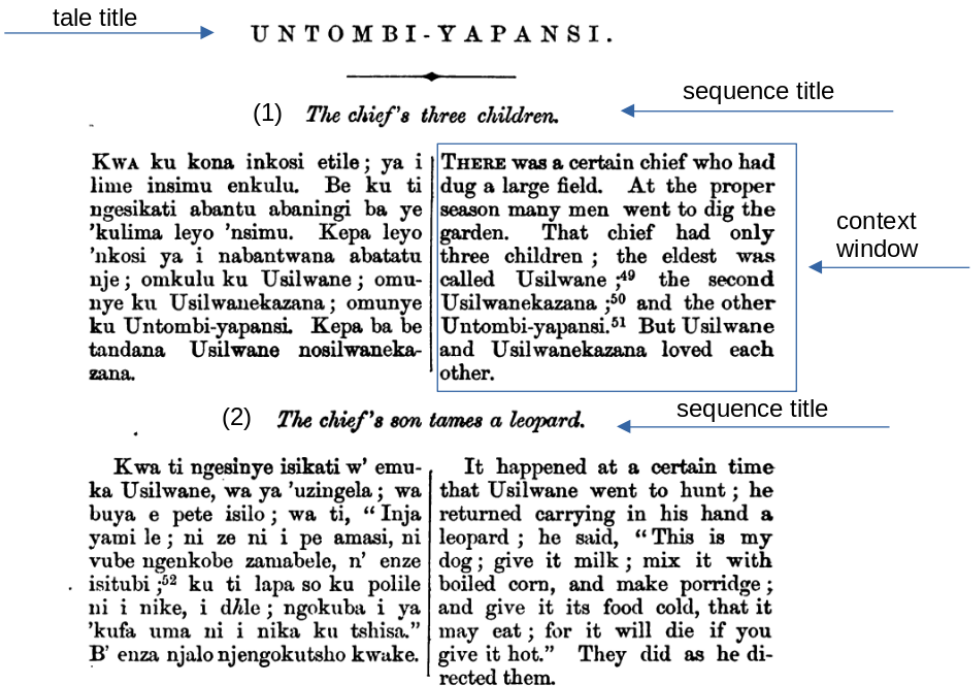


Figure 2: Excerpt from Henry Callaway's Nursery Tales and Histories of the Zulus (Callaway 1868)

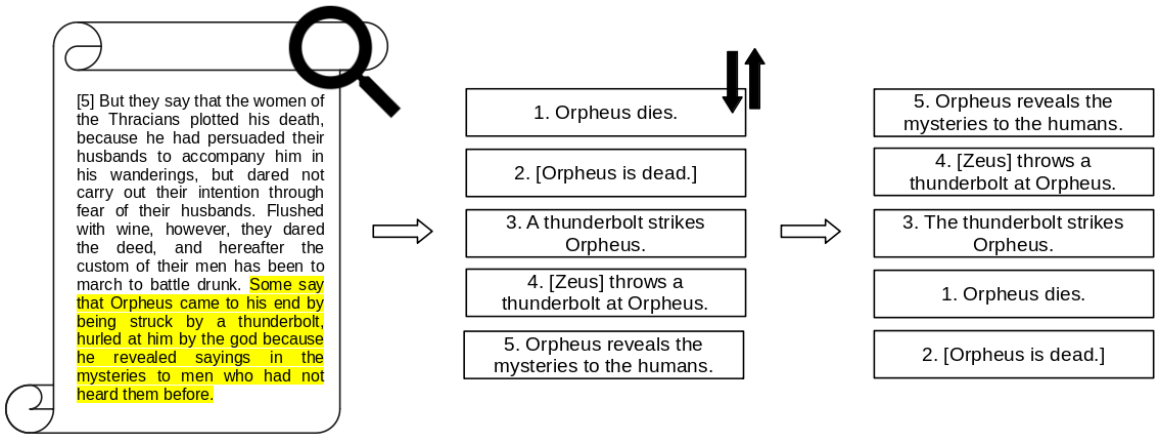


Figure 3: Hyleme extraction steps on the example of Orpheus' death in Pausanias' Description of Greece, Source: Poster accompanying (Pannach 2023a)

3. The cannibal is dead.

In this case, the direct speech is subsumed by using *to mock* as the hyleme predicate, because the content of the mocking does not carry narrative relevant information, as the trickster hero knows that the cannibal is dead. In contrast, "I too am ignorant how those animals got into my bag" (ibid.) translates to the hyleme "Uthlakanyana says he does not know how the animals got into his bag" because the statement is a lie, which is relevant information for character analysis.

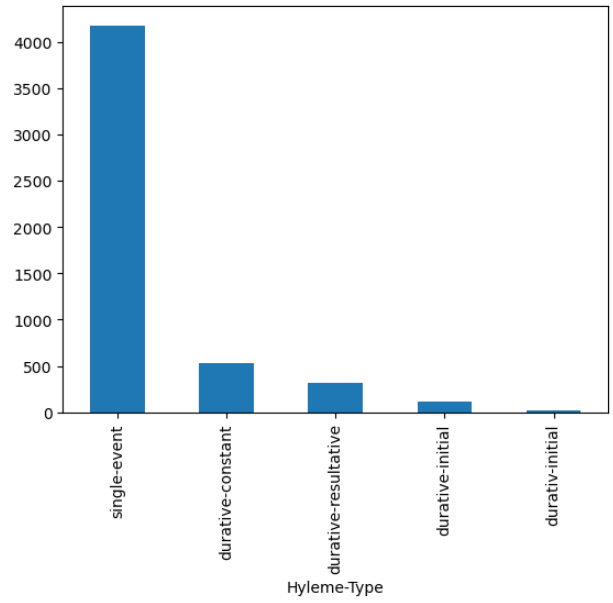


Figure 4: Distribution of hyleme functions in the data set

The application of a frame semantic parser (Chanin 2023) yields the frames and frame elements (FEs) as shown in Table 2. While the content of the frames is mostly correctly parsed, we can see that the results are only correct on a superficial semantic but not on a narrative level. For example, from the phrase "that I may listen and hear", the parser invokes both an active (listen) and a passive (hear) perception frame, when the intended meaning is of course that the dead cannibal can no longer play on the calabash drum, the object which the trickster desired all along. Additionally, the parser cannot distinguish between actual narrative events and frames invoked by direct speech, e.g. that calling the cannibal uncle does not imply a kinship relation between the two characters. Additional steps, such as verifying the statements in the direct speech against the events that happen in the narrative require a higher level of semantic inference, e.g. world knowledge that a dead cannibal cannot speak, and thus the content of the utterance is not true.

Frame-semantic parsing on hylistic data is perceivably straightforward, as shown in Table 3. The frame assignment *people_by_vocation* might be open to discussion, and the parser missed to assign a frame to the *mocking* event, which is a characteristic action of a trickster. However, since the hylemes already reduce the input to the parser to the narratologically relevant events and states (*opens(Uthlakanyana, door), dead(cannibal)*), the results are better suited for further analysis, especially when multiple narratives are compared.

Frame-semantic annotation and parsing can give us insights into narrative events and their participants, both on the extracted sequences and on the full-text. However, even if we manually annotated the frames on the full text of the story and thus created a perfect dataset, we would still miss important events - namely those that are implied or not directly stated in the text, e.g. the mocking as an action of the main character. Thus, the additional step of *hylistic analyses* which makes explicit what is implicitly present in the text, is still valuable.

5.2 Case Study 2: Topic modeling on Hyleme Sequences

In the second case study, I want to demonstrate how topic modeling techniques can be used to investigate the hyleme sequences for latent topics. For this purpose, latent topics are extracted by employing BERTopic (Grootendorst 2022) on the individual hylemes.

The technique is applied to the data in hylistic format with masked entities and vernacular terms, i.e. "[PER] takes [PER] by the arm." This is due to the result of a previous study (Pannach 2023*b*), which largely translated the tales themselves as topics when vernacular terms and proper names are unmasked in the training data. While this could have some application to discover tales from a large dataset of hylemes, it does not provide much insight into the individual latent topics across the sequences.

Here, I compare the results of different topic modeling strategies on hylistic data and on the English tale texts using BERTopic. Namely, the standard configuration of BERTopic, a model with reduced outliers, a KeyBertInspired model [3], and lastly a model with a preset number of topics $k = 85$. The

Table 2: Results of Frame Semantic Parsing on original text, (S = Sentence Index)

S	Frame	FE	Value	FE	Value
1	Closure	Agent	Uthlakanyana	Container_portal	the door
	Connecting_architecture	Part	door	-	-
2	Emotion_directed	Experiencer	you	Time	still
3	Temporal_collocation	Trajector_event	cry out so as to be heard	-	-
	Communication_noise	Speaker	you	Message	so as to be heard
	Perception_experience	Phenomenon	you	-	-
	Awareness	Cognizer	I	Content	you were screaming
	Make_noise	Sound_source	you	-	-
	Kinship	Ego	My	Alter	uncle
4	-	-	-	-	-
5	Perception_active	Perceiver_agentive	I	-	-
5	Perception_experience	Phenomenon	that	Perceiver_passive	I
6	Arriving	Theme	he	-	-
7	Dead_or_alive	Time	when he came	Protagonist	the cannibal
8	Removing	Agent	He	Theme	him
		Source	out of the house	-	-
8	Buildings	Building	house	-	-
8	Possession	Owner	He	Possession	of it.
9	Sleep	Sleeper	He	-	-
9	Emotion_directed	Experiencer	He	Time	now

Table 3: Results of Frame Semantic Parsing on hyleme sequence text, (H = Hyleme Index)

H	Frame	FE	Value	FE	Value
1	Closure	Agent	Uthlakanyana	Container_portal	the door
2	People_by_vocation	Person	cannibal	-	-
3	Dead_or_alive	Protagonist	The cannibal	-	-

number of topics discovered by each configuration is shown in Table 4. One of the advantages of BERTopic compared to other approaches, such as LDA (Blei et al. 2003) is that the former does not require pre-processing of the data, and thus does not to the same degree suffer from error propagation due to performance issues in the pre-processing stage. However, since the original text is derived from the scanned PDF files of the book [4], the text had to be extracted using an OCR tool [5]. Afterwards, a language detection library [6] was used to extract the English text, and the tale texts were split into sentences using the NLTK (Bird et al. 2009) sentence tokenizer. All these additional steps require either a lot of manual data cleaning, or introduce errors in the resulting topics.

Table 4: Number of topics discovered by model

Model	$\#\text{Topics}_{h}$	$\#\text{Topics}_{orig}$
Standard configuration	156	72
Reduced Outliers	155	72
KeyBERTInspired	142	64
preset	85	60

In the standard configuration, we can see distinct topics derived from the hylistic data set that refer to individual semantic fields, as shown in Figure 5. For instance, topic 0 includes terms that describe royalty, such as kings and quens; whereas topic 3 describes terms that refer to becoming a member of

the royal family (e.g. *woman-bride-queen*). Topic 2 includes terms about animals, such as *hyenas and cattle*. Notably, topic 1 and topic 6 are very similar, containing terms that refer to *hungry cannibals*.

In Figure 8, the intertopic distance maps of the standard model, the KeyBERTInspired model and the preset-85 model are illustrated. It shows that there is some overlap between the topics, e.g. topic 58 and topic 130, highlighted in red in Figure 7b, describe weather phenomena. They include words like *thunderstorm, raining, rain, thunders, starts* (topic 58) and *clouds, sky, fog, cloud, sun* (topic 130). This way, the intertopic maps can be investigated for topic clusters, such as those referring to siblings (e.g. *sisters, brothers, twins*, topics 18 and 25, highlighted in orange), farming (e.g. *cattle, cattlepen, pasture, grazing*, topics 68,83,88,111,119, highlighted in blue). The topics are generally easy to interpret and contain words related to motifs that appear across the hyleme sequences, which allows interesting insights into the themes of the folktales.

Additionally, we can add inspect new hylemes for the most similar topics contained: e.g. the fictional hyleme "[PER] cries at the [LOC]" is most similar to the topic (KeyBERTInspired) containing terms like *cries, crying, screams, enraged, upset* and *she*. However, BERTopic does not yield good results for sequenes of hylemes in the presented data set. It yields only three topics, one containing largely stopwords, one referring to cannibals, and one containing vernacular terms and terms related to family structure. The project repository makes the graphs, intertopic distance maps, as well as the associated jupyter-notebooks available for further inspection.

The topics trained on the original text are more prone to noise. For instance, in Figure 6 shows topic 2 with vernacular terms and prefixes, that should have been filtered out by the language detection. Topic 4 includes artifacts from masking the named entities with [PER] tags. Nevertheless, BERTopic found similar topics in both data sets, e.g. in Figure 6 the cannibal topic (topic 1), or the royalty topic (topic 7). Inspecting the intertopic distance maps in Figure 7 closer reveals more "noise topics", e.g. including abbreviations from Callaway's footnotes, or more vernacular terms. The topics are also semantically closer than for the model trained on hylistic data, creating more clusters, such as the family cluster, such as a cluster about cooking and the preparation of sour beer, (topics 15, 19, 58, highlighted in orange in Figure 8b).

6 Conclusion

This article presents the application of the *hylistic theory* for the domain of folklore. It introduces briefly what *hylemes* and *hyleme sequences* are, and how they are extracted from narrrative texts. Based on the annotation, it further introduces a freely available dataset of over 5000 *hylemes* based on Henry Callaway's collection of tales of the amaZulu. On this basis, two case studies demonstrate what kind of further analyses can be performed on *hylistic* data using computational tools, namely frame-semantic parsing and topic modeling. Both techniques are compared to their application on the English translations of the original texts.

7 Future Work

As demonstrated, data in *hylistic* format is easier to process with natural language processing tools than concurrent, discursive text. With the presented data set, Callaway's tales can be studied from various perspectives, e.g. analysing characters and their associated behaviour using sentiment analyses techniques, the tales' event structures (e.g. by identifying recurring patterns and spheres of actions) and many more. This paper uses two basic case studies for demonstration purposes, which can be extended into more detailed analyses both by adding more data, e.g. from other parts of Southern African, or from a technical standpoint, e.g. by comparing different implementations of the frame-semantic parser. While the (high quality) annotation of narrative structure is still largely a manual task, especially in a domain that differs in form, con-

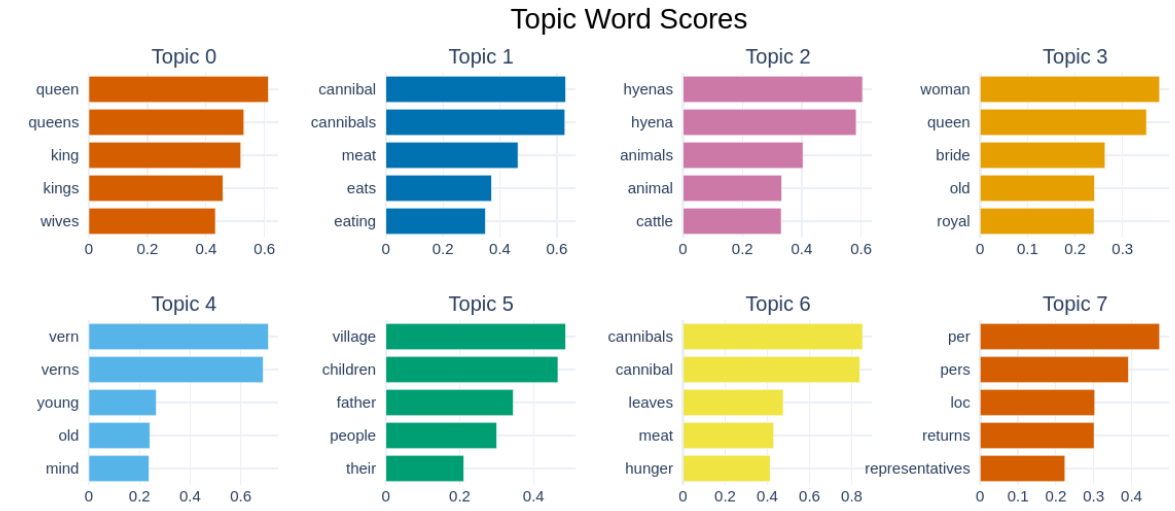


Figure 5: Top 8 topics derived from individual hylemes using BERTopic (KeyBERTInspired)

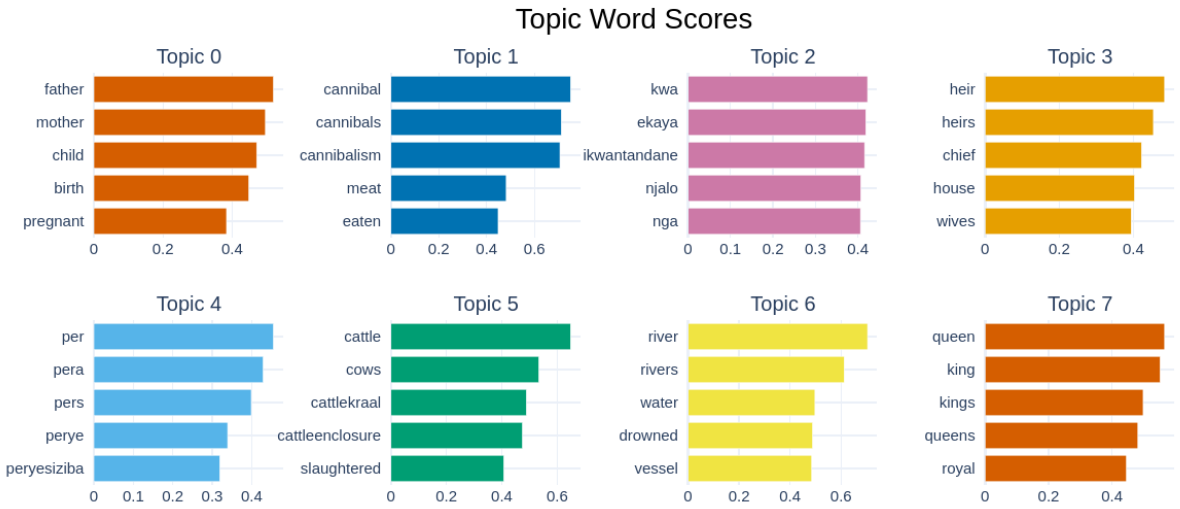
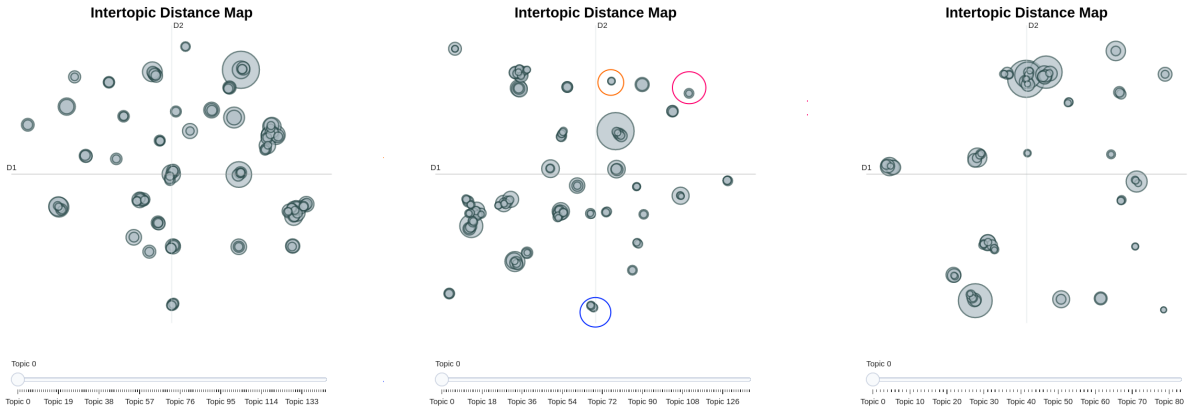
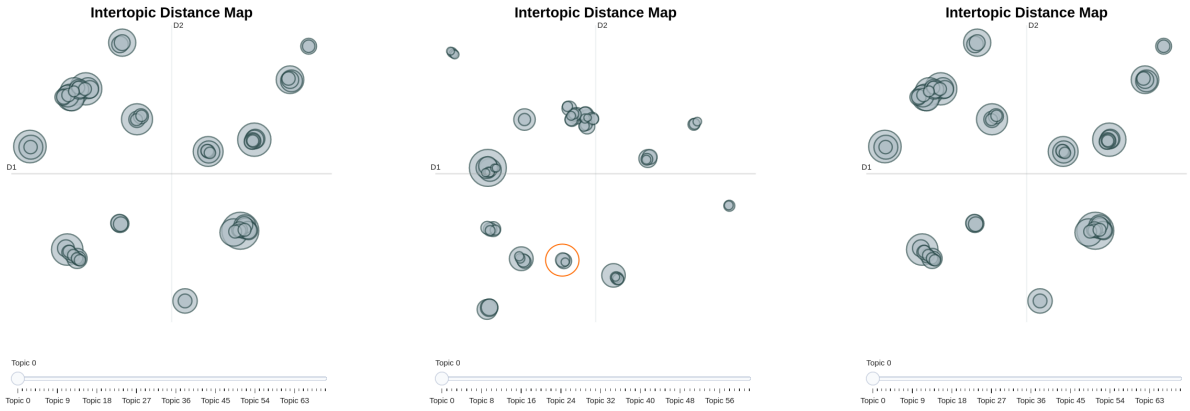


Figure 6: Top 8 topics derived from original text using BERTopic (KeyBERTInspired)



(a) Standard model *(b) KeyBERTInspired model* *(c) preset model with* k=85

Figure 7: Intertopic distance maps for standard configuration, KeyBERTInspired and preset-85 topic models



(a) Standard model *(b) KeyBERTInspired model* *(c) preset model with* k=85

Figure 8: Intertopic distance maps for standard configuration, KeyBERTInspired and preset-85 topic models

tent and presentation from other narrative texts, such as news paper articles, future research might explore if *hylistic annotation* can be assisted by semi-automatic extraction of candidate events and states, e.g. using Large Language Models.

Since the data is available under a Creative Commons license, the sequences can be re-used as input for further studies by interested domain experts and laymen scholars. Furthermore, domain experts and experts in vernacular language are invited to explore the remaining tales of the collection and add their own annotations.

In the future, it is planned to map the annotations to the respective text passages in the vernacular language and in the English translation and create an annotated edition available in XML-TEI format.

I hope that this contribution motivates more annotation studies on traditional African (and other) folktales using the *hylistic* approach. This would allow a larger scale (inter-cultural) comparison of the narrative structure, patterns, and representation of characters of folktales.

Notes

[1] I compare *hylistic analysis* to a different event and state annotation model by Gius & Vauth (2022) in Pannach (2023*a*).

[2] `github.com/FPannach/Callaway_Hylistics/`

[3] `https://maartengr.github.io/BERTopic/api/representation/keybert.html`

[4] `http://emandulo.apc.uct.ac.za/metadata/FHYA\%20Depot/1188/1207/index.html`

[5] tesseract Version 4.1.1

[6] langdetect for Python `https://pypi.org/project/langdetect/`

Acknowledgments

This work was funded by the German Research Foundation (DFG) as part of the myth research group FOR 2064 STRATA, Stratification Analyses of Mythic Plots and Texts in Ancient Cultures.

The author would like to thank the reviewers for their insightful comments and helpful suggestions, as well as their time and effort.

References

Abello, J., Broadwell, P. & Tangherlini, T. R. (2012), 'Computational folkloristics', *Communications of the ACM* **55**(7), 60–70.

Aplonova, E., Jordanoska, I., Arkhangelskiy, T. & Nikitina, T. (2023), Speechreporting corpus: annotated corpora of west african traditional narratives, *in* 'Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)', pp. 26–31.

Azuonye, C. (1990), 'Morphology of the Igbo folktale: Ethnographic, historiographic and aesthetic implications', *Folklore* **101**(1), 36–46.

Bird, S., Klein, E. & Loper, E. (2009), *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* **3**(Jan), 993–1022.

Callaway, H. (1868), *Nursery tales, traditions, and histories of the Zulus, in their own words, with a translation into English and Notes: Volume I*, Springvale, Natal: Blair.

Chanin, D. (2023), 'Open-source frame semantic parsing', *arXiv preprint arXiv:2303.12788* .

Declerck, T., Klement, T. & Kostova, A. (2016), Towards a WordNet based classification of actors in folktales, *in* C. Fellbaum, P. Vossen, V. B. Mititelu & C. Forascu, eds, 'Proceedings of the 8th Global WordNet Conference (GWC)', Global Wordnet Association, Bucharest, Romania, pp. 83–87.
URL: *https://aclanthology.org/2016.gwc-1.13*

Declerck, T., Koleva, N. & Krieger, H.-U. (2012), Ontology-based incremental annotation of characters in folktales, *in* 'Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities', pp. 30–34.

Eliot, G. (1994), *Middlemarch*, Blackwood. public domain.
URL: *https://www.gutenberg.org/ebooks/145*

Finlayson, M. A. (2016), 'Inferring Propp's functions from semantically annotated text', *Journal of American Folklore* **129**(511), 55–77.

Finlayson, M. A. (2017), 'Propplearner: Deeply annotating a corpus of russian folktales to enable the machine learning of a russian formalist theory', *Digital Scholarship in the Humanities* **32**(2), 284–300.

Finlayson, M. M. A. (2012), Learning narrative structure from annotated folktales, PhD thesis, Massachusetts Institute of Technology.

Gervás, P. & Méndez, G. (2024), Tagging narrative with propp's character functions using large language models (negative results), *in* 'Proceedings of the Seventh International Workshop on Narrative Extraction from Texts (Text2Story 2024)'.

Gius, E. & Vauth, M. (2022), 'Towards an event based plot model. A computational narratology approach', *Journal of Computational Literary Studies* **1**(1).
URL: *https://doi.org/10.48694/jcls.110*

Grootendorst, M. (2022), 'Bertopic: Neural topic modeling with a class-based tf-idf procedure', *arXiv preprint arXiv:2203.05794* .

Hagedorn, J. & Darányi, S. (2022), 'Bearing a bag-of-tales: An open corpus of annotated folktales for reproducible research', *Journal of Open Humanities Data* **8**(16).

Karsdorp, F. & Fonteyn, L. (2019), 'Cultural entrenchment of folktales is encoded in language', *Palgrave Communications* **5**(1).

Karsdorp, F., van der Meulen, M., Meder, T. & van den Bosch, A. (2015), 'MOMFER: A search engine of thompson's motif-index of folk literature', *Folklore* **126**(1), 37–52.
URL: *https://doi.org/10.1080/0015587X.2015.1006954*

Lendvai, P., Declerck, T., Darányi, S. & Malec, S.

(2010), Propp revisited: Integration of linguistic markup into structured content descriptors of tales, *in* 'Digital Humanities 2010, London, United Kingdom'.
URL: *https://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-753.html*

Okodo, I. (2012), 'Obaraedo: Conformity to Proppian morphology', *AFRREV IJAH: An International Journal of Arts and Humanities* **1**(2), 100–111.

Pannach, F. (2023*a*), "Orpheus came to his end by being struck by a thunderbolt": Annotating events in mythological sequences, *in* 'Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)', Association for Computational Linguistics, Toronto, Canada, pp. 10–18.
URL: *https://aclanthology.org/2023.law-1.2*

Pannach, F. A.-K. (2023*b*), Narrative Echoes across Time and Space, Phd thesis, University of Goettingen, Germany. Available at `http://dx.doi.org/10.53846/goediss-10349`.

Pannach, F., Sporleder, C., May, W., Krishnan, A. & Sewchurran, A. (2021), 'Of Lions and Yakshis', *Semantic Web* **12**(2), 219–239.

Propp, V. (1968), *Morphology of the Folktale*, Vol. 10, University of Texas Press, Austin, Texas, USA.

Reiter, N., Frank, A. & Hellwig, O. (2014), 'An nlp-based cross-document approach to narrative structure discovery', *Literary and Linguistic Computing* **29**(4), 583–605.
URL: *https://doi.org/10.1093/llc/fqu055*

Tolbert, J. A. & Johnson, E. D. (2019), 'Digital folkloristics', *Western Folklore* **78**(4), 327–356.

Zgoll, A., Cuperly, B. & Cöster-Gilbert, A. (2023), In search of Dumuzi: An introduction to Hylistic narratology, *in* 'The Shape of Stories', Leiden: Brill, pp. 285–350.
URL: *https://doi.org/10.1163/9789004539761*$_0$13

Zgoll, A. & Zgoll, C., eds (2020), *Mythische Sphärenwechsel*, De Gruyter, Berlin, Boston.
URL: *https://doi.org/10.1515/9783110652543*

Zgoll, C. (2019), *Tractatus mythologicus*, De Gruyter, Berlin, Boston.
URL: *https://doi.org/10.1515/9783110541588*

Zgoll, C. (2020), Myths as Polymorphous and Polystratic Erzählstoffe, *in* A. Zgoll & C. Zgoll, eds, 'Mythische Sphärenwechsel', De Gruyter, Berlin, Boston, pp. 9–82.
URL: *https://doi.org/10.1515/9783110652543-002*