GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Narrative Echoes across Time and Space: Comparative Analysis of Structural Similarities in Myth and Folktale Sequences

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
“Doctor rerum naturalium”
der Georg-August Universität Göttingen

im Promotionsprogramm “PhD Programme in Computer Science” (PCS)
der Georg-August University School of Science (GAUSS)

vorgelegt von Franziska Anne-Katrin Pannach
aus Löbau
Göttingen, 2023

Betreuungsausschuss

Prof. Dr. Caroline Sporleder
Institut für Informatik
Göttingen Centre for Digital Humanities
Georg-August Universität Göttingen

Prof. Dr. Ramin Yahyapour
Institut für Informatik
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
Georg-August Universität Göttingen

Prof. Dr. Christian Zgoll
Seminar für Klassische Philologie
Georg-August Universität Göttingen

Mitglieder der Prüfungskommission

Referentin
Prof. Dr. Caroline Sporleder
Institut für Informatik
Göttingen Centre for Digital Humanities
Georg-August Universität Göttingen

Koreferent
Prof. Dr. Ramin Yahyapour
Institut für Informatik
GWDG
Georg-August Universität Göttingen

Weitere Mitglieder der Prüfungskommission

Prof. Dr. Christian Zgoll
Seminar für Klassische Philologie
Georg-August Universität Göttingen

Prof. Dr. Wolfgang May
Institut für Informatik
Georg-August Universität Göttingen

Prof. Dr. Carsten Damm
Institut für Informatik
Georg-August Universität Göttingen

Prof. Dr. Bela Gipp
Institut für Informatik
Georg-August Universität Göttingen

Tag der mündlichen Prüfung: 28.11.2023

Für Ortrud Reh & Klaus-Dieter Amelang

Abstract

Originating in mythological research, the *Hylistic* theory [1] introduces *hylemes* as a basic plot unit, containing actions, states, or background information derived from a source. These statements are organized in *hyleme* sequences which are derived by domain experts and concern a narrative variant. By using those sequences, narrative variants can be compared across different source materials, such as text genres or sources in different languages, and subsequently structurally aligned with variants of the same narrative. Establishing appropriate methods for the automation of these *hyleme* sequene alignments is the central objective of this work. This thesis presents the first approach towards a *Digital Hylistic* theory. This work is related to the research discipline of Computational Narratology, and its related areas.

Two data sets are the basis of the conceptual and exploratory studies undertaken in this work: the German *hyleme* data sets consists of sequences from different cultural and temporal backgrounds, including Ancient Greece, or Mesopotamia. Those sequences are extracted by domain experts in the fields of Classics, Ancient Near Eastern Studies, and Religious Studies as part of the DFG-funded myth research group 2064 STRATA. They are derived from a multitude of sources, in different source languages, and from different genres and styles of narratives. The second data set is the first ever English *hyleme* data set, containing sequences describing a set of Zulu folktales. The data set is also the first approach towards a *hylistic* representation of folkloristic (in contrast to mythological) material.

This thesis follows a multi-method approach, grounded in the narrative theory of *Hylistics*, that carefully models *hylemes* and their properties as objects that can be processed and analyzed using methods from the fields of Natural Language Processing, Knowledge Engineering, and Formal Languages. This work approaches the comparability of *hylemes* from a semantic similarity point-of-view. The problem of aligning *hyleme sequences* is approached from different angles with a focus on the research question of the domain expert, e.g. comparison of variants of the same myth or exploratory alignment with the purpose to discover interesting patterns. This work does not aim to solve the *hyleme* alignment task, because alignment can be performed for various purposes.

All methods in this work are selected based on the appropriateness with respect to the *Hylistic* theory. States and events, conveyed by *durative* and *single-point* hylemes, are modelled and compared using fundamentally different methods. Previous work on story similarity suggests that many salient features contribute to the judgment of how similar two or more narrative variants are. Therefore, the similarity of background information and plot-driving actions is approached from different standpoints and with different methods. In combination, these methods can be used for alignment and the comparison of background information, which yields a holistic measure of the similarity between narrative variants.

Contents

List of Figures

List of Tables

Glossary

ANES Ancient Near Eastern Studies xiv, 13, 60, 63, 83, 93, 104, 138

ARM Association Rule Mining 46, 48

ATU Aarne-Thompson-Uther (tale type index) 51

BERT Bidirectional Encoder Representations from Transformers 52, 119

BLOSUM Blocks Substitution Matrix 22

BOW Bag of Words 57

CBOW Continuous Bag of Words 28

CL Computational Linguistics 7, 54, 83, 131

CNN Convolutional Neural Network 50

CS Classical Studies 7, 13, 83, 93, 104.
Computer Science 21, 60, 83.

DFA Deterministic Finite Automaton 110, 173

DFG Deutsche Forschungsgemeinschaft, German Research Foundation v, 60

DH Digital Humanities 7, 83

DNA Deoxyribonucleic acid 21, 22, 25, 127

FE Frame Entity 19, 20

FFNN Feed Forward Neural Network 28

GCN Graph Convolutional Neural Networks 50

GPT Generative Pre-Trained Transformer 113, 114

IC Information Content 30, 31, 56

KHM Kinder- und Hausmärchen (Grimm) 32, 52, 57, 59, 64

LCS Least Common Subsumer 31.
Longest Common Subsequence 115, 116, 120.

LCStr Longest Common Substring 115, 116, 122, 124

LDA Latent Dirichlet Allocation 32, 76, 77, 171

LSTM Long Short-term Memory 50

MSA Multiple Sequence Alignment 45, 58

NER Named Entity Recognition 26, 34, 68, 69, 89

NLP Natural Language Processing 7, 15, 17, 21, 50, 54, 59, 67, 89, 127, 131

NLTK Natural Language Tool Kit 139, 140

NPMI Normalized Pointwise Mutual Information 33

OCR Optical Character Recognition 142

OWL2 Web Ontology Language (2) 92, 104

PAS Predicate Argument Structures 19, 47, 48, 49, 100, 107, 151

PMI Pointwise Mutual Information 46

RNN Recurrent Neural Network 28

SKOS Simple Knowledge Organization System 93, 94

STRATA Stratification Analyses of Mythic Narrative Materials and Texts in Ancient Cultures, DFG myth research group 2064 STRATA v, 9, 60, 82, 93, 130, 153

SVP Separable Verb Prefix 65

TEI Text Encoding Initiative 151

TF-IDF Term Frequency Inverse Document Frequency 36, 37, 52, 86, 143, 144, 145, 148

TMI Thompson Motif Index 52, 53, 152

TSG Temporal Script Graph 45, 46

t-SNE t-distributed Stochastic Neighbor Embedding 28

TTL Terse RDF Triple Language (Turtle) 93

WHO World Health Organisation xiii, 2

WSD Word Sense Disambiguation 31, 56

XML Extensible Markup Language 52, 53, 151

Chapter 1

Introduction

RIKER
Greek, sir?

PICARD
The Homeric Hymns.
(glances up)
One of the root metaphors of our own culture.

RIKER
For the next time we encounter the Tamarians...

PICARD
(nodding)
More familiarity with our own mythology might help us relate to theirs.

Star Trek The Next Generation S5; E2: "Darmok"

1.1 Motivation

If we are in a medical emergency, we turn to the Rod of Asclepius which we know will guide us to a nearby medical professional. Why is the combination of rod and snake a symbol of the medical and pharmaceutical professions across the world? According to Wilcox and Whitham [12], the appearance of the rod as an emblem for medical professionals has been in use since at least the beginning of the 17th century. But the symbol of the snake curled around a staff itself has no immediate connection to the medical profession. It is only by recognizing the underlying mythical

Figure 1.1: Rod of Asclepius as depicted in the WHO logo Source: Wikimedia Commons, Public Domain

context that we understand why it was chosen to represent the medical field. The rod is an attribute of Asclepius, a Greek deity of medicine.[1] If we pay close attention, it is astounding how often we encounter mythical content in daily life. Even more astounding is how often these undertones go unnoticed or taken for granted.

Apart from mere mythological symbolism, like the usage of the rod of Asclepius, mythical content is also constantly re-adapted in popular fiction. From re-tellings of a popular story with relatively close resemblance to an ancient version, such as the epic sword-and-sandal movies of the 1950s and 1960s to modern adaptations like *Troy* (2004), to movies and TV series that re-use mythological content only loosely in their narrative (e.g. *Good Omens* (TV series, 2019-)).

Often popular fiction re-uses bits and pieces of mythical lore to draw on a certain amount of knowledge transfer from the recipient. For example, when Harry Potter encounters a Sphinx in *Harry Potter and the Goblet of Fire*, the reader immediately associates certain features with that encounter: a guardian, who has the head of a human and the body of a lion, and a riddle that is to be answered correctly in order to proceed.

Even in our everyday language, we use mythical concepts to communicate certain ideas. Idiomatic expressions such as "Achilles' heel" or "Trojan Horse" only evoke the desired imagery if the recipient knows at least the basic concepts behind those expressions. Yet the wide-spread use of these idioms suggests that most people recognise those basic concepts.

So why is it that people continue to be drawn to ancient myths? This thesis cannot give a comprehensive answer to this question, if that is even possible. However, the following key aspects shall serve as pointers.

[1] https://www.wikidata.org/wiki/Q83223

Myths are fascinating. Mythological plots have a way of being told and retold that resonates with the human mind. In the sword-and-sandal movies mentioned above, we witness some of the monumental imagery and philosophy which mythological storytelling yields. Mythological plots deal with the big questions that we as humans ask ourselves: How far are we willing to go for love? How strong is our faith? What do we, and what did ancient people believe in?

The re-use of mythological plots, motifs and settings is continuous. Movies, TV shows, and books still draw on that familiar imagery because it fascinates and entertains us, just as it fascinated people in the near and distant past.

Myths explain the world around us. Myths have educational value on different layers. One value for the recipient today is certainly that they tell us about our (global) past. A popular and intuitive example is the genre of (great) flood or deluge myths [13]. Flood myths date back millenia, with the earliest examples from the 19th to 18th century BC.[2] Interesting for this particular genre is not only the broad spectrum of mythological plots that describe great floods in different geographical and temporal settings, but also the scientific effort that has been spent by researchers to search for "the one true great flood". To try and pinpoint the corresponding floods to a real-life event has proven to be almost impossible, if not moot. Scientific research has shown that there were great floods in ancient times, so it is not far fetched to believe that the people who originally crafted the stories that would later become myth had some first hand experience with them. But human nature wants more: It wants to identify exactly where a flood happened and how much of the popular story, the near extinction of human life, is safe to be held to be true. Furthermore, flood myths are often told by means of a *Pourquoi story* or *Aitiology* which allows us to study the world views of ancient peoples, their belief systems on the cause of things like natural disasters.

Additional to mythical-historical storytelling, we also learn about the educational value of myths for the ancient recipients. Myths supply the recipient with a model that explains the past, the present and the future at the same time [14]. To give another example, we can learn from the myths of the Mesopotamian she-demon Lamaštu[3] how vulnerable certain demographic groups, such as pregnant women, babies, or elderly people are, and that they need special protections from the various diseases that Lamaštu is associated with. One might argue that the idea of Lamaštu strangling babies is an *Aitiology* of what we today call SIDS[4] [16]. Myths also teach us quite descriptively which animals can be dangerous, which herbs and fruits are beneficial for our health, or which geographical areas to avoid.

Educational subjects in mythological plots can be plain or hidden. In Mesopotamian invocations we see how practical instructions on performing rituals are woven into mythological plot, a feature that Burkert [10] describes as 'particularly profound'. The ritual instructions for the priest or

[2] MS3026 `https://www.schoyencollection.com/literature-collection/sumerian-literature-collection/sumerian-flood-story-ms-3026`, Atra-Hasis MS2950 `https://www.schoyencollection.com/literature-collection/babylonian-literature-collection/atra-hasis-a-ms-2950`

[3] On the question whether or not Lamaštu should be called a goddess, please see [15, p.6]

[4] Sudden Infant Death Syndrome

priestess are embedded into a plot, creating two (or possibly more layers): the instructions and the story. Ritual and myth are interwoven. If the myth aspect of this relationship gets lost, we are left with tradition, without much significance for our belief system (imagine the action of setting up a Christmas tree). The story is needed in order to understand the reasoning behind the ritual instructions. In doing so, the ancient people of Mesopotamia did us a great favor. How else would we know how to interpret the action of a priest in ancient Uruk, when he places a gold necklace around a statue? Even today, ritual descriptions are not plain instructions for a series of actions. Nobody would assume that Christian people lining up on a Sunday to receive the Lord's Supper are doing so, because they are hungry. We know how to interpret the (ritual) behaviour from the story, and we create the mental connection because the ritual object, e.g. the necklace, has an affinity to the mythological story it references, like the shining light of the sun.

In these senses, myths are educational. But more importantly, myths give us the illusion that we understand how the world around us works. Until scientific discovery takes over that function of explaining how a thing came to be, it is the myth that we employ. And in turn, science can explain and discover something that was already present in the mythical context [17], but previously misunderstood or unexplained.

Myths create an emotional response. A myth communicates themes that always invoke feelings in some way. When we read about the Lamaštu crawling under the bed sheets at night, possessing and killing innocent children or puerperal or pregnant women by bringing disease and fever [15], it brings out a natural basic fear of harm done to us when we are in our weakest states. Just like horror novels or movies, myths play with our emotions and remind us of our own mortality. In the same way, the flood myths described above also evoke an emotional reaction in the recipient. They are special in a way, because "[t]he flood is the work of the gods, but at the center is the fate of man."[5] [13, p.15]

Moreover, we find mythological content so entertaining that certain features appear again and again in human storytelling. As an example we can consider the *trickster* motif. A trickster is a cunning figure in a narrative, who overcomes his or her opponents by means of imaginative use of his or her resources. According to Caduff, trickster heroes are a "relatively archaic product of mythological thinking." [13, p.280] Often, trickster tales tell a story of the trickster hero overcoming an imbalance of power [18, 19]. The human nature wants to resolve the power imbalance and roots for the trickster, even if the cunningness of his or her approach has fatal consequences for the trickster's opponent. Trickster motifs are one of the examples where we can observe a close resemblance between myth and folktale. Be it *UHlakanyana* in Henry Callaways collection of Zulu folktales [7], the *Poor Man of Nippur* [10] or modern variations, such as *Catch me if you can* (Movie, 2002), trickers are found across many, if not all, cultures and times. Myths invoke those emotional responses through mental pictures that are familiar to the (ancient) recipient: a snake is dangerous, the death of a loved one is tragic, the sun rising in the morning creates hope and life.

[5] Original citation translated by the author.

Myths show us aspects of human nature. Myths across cultures portray feelings that are inherent to the human nature: love and deceit, fear and pain, hope and longing. Mythological plots contain plot elements that are so embedded in the relationships humans share with each other that their appearance naturally appeals to us, like in the timeless love story between *Orpheus* and his wife *Eurydice*. There are certain ground truths of human nature that surface in myth (and any other form of story telling for that matter) time and time again. Mothers who love their children so much, they willingly sacrifice themselves for them. Honor and courage prevailing over dishonor and cowardice. Those aspects of human nature are valid beyond cultural and temporal boundaries and are ultimately what sets us apart from animals. Those universal values and truths are elements of a myth because they are inherent to human nature.

Myths have sociological function. Apart from entertainment and (historical) education, myth also has an important social purpose. Campbell argues that myth has inherent sociological functions. He claims that "it has always been on myths that the moral orders of societies have been founded" [20, p.20]. More so, in embedding morals into society, he suggests that myths have an inherent "life-supporting nature" [20, p.20].

The arguments presented in this sections are by no means exhaustive and could each easily fill an entire monograph. They are merely given as a pointer on why the study of ancient myth is relevant today. They should also serve as an incentive to study myth in a broader context, in and outside of ancient studies. The quote at the beginning of this chapter is taken from an episode of the popular 1990s TV series "Star Trek: The Next Generation", in which Captain Picard encounters a civilisation that communicates exclusively through mythical and legendary metaphors and comparisons, the *Tamarians*. In learning about the meaning of the metaphors, Picard slowly begins to understand the imagery that the Tamarians aim to invoke and the message they try to bring across. When we study ancient myths and legends today, we do the same.

It is easy to think that the post-modern world is free of the influences of ancient myth. However, once we start observing our surroundings, our speech and idioms, the media and advertisements we consume, we realise that myth is–in fact–everywhere.

1.1.1 *Myth* and mythological research

The question what exactly a myth constitutes is very complex. Upon studying the different aspects of the term, one might come to the conclusion that there are as many definitions of *myth* as there are accounts of biblical floods. For instance, in his treatise on ancient deluge or flood myths, Caduff [13] does not differentiate much between the forms myth and saga. Instead, he argues that the narrative of a saga largely concerns human beings, whereas the myth is centered around beings that are worshipped as deities. In deluge or flood narratives, however, the flood is brought by the deities, but the core narrative is the effect it has on the mortals.

Burkert points out that the concept of myth is a culture- and language-sensitive complex, that must also take into consideration the different termini that a language provides to distinguish certain genres, styles and narratological situations [10]. He comes to the conclusion that a wider definition of the term myth is achieved if we consider them “applied narratives” which verbalise the hyper-individual and collectively influential features of reality [10, p.65].

Myths can be classified across many different dimensions, e.g. location (e.g. Norse myths), time (e.g. Ur-III myths), by source or author (Homerian Myths), or topic (e.g. deluge myths). This thesis cannot answer the question what a myth is. Instead, we follow the definition of C. Zgoll [21, pp.14], that myths are neither text nor a literary genre, but a collection of different representations of a narrative, a *Stoff*.[6] This means that a *myth* is independent from its medium (text, picture or object) or genre (e.g. poems or travelogues). Instead, it is a *Stoff*, i.e. a collection of all possible variants of a certain narrative (see Section 2.1).

Narrative and cognition are fundamentally connected, because without a mental process, a recipient will be unable to comprehend a story and place it within a frame of reference. When we read stories, we are able to construct a mental representation of the narrative and its circumstances. The rules and circumstances within a narrative together with the plot form a coherent picture for us as readers, because we can comprehend a fictional (or non-fictional) world beyond what our own experience and world-knowledge provides us with. If humans did not have this ability, no fictional story would be comprehensible for us. We can relate to a story of a person trapped in a landslide, without having first-hand experience of it. Herman called these mental models the “storyworlds” [22].

Therefore, the significance of narrative in general, and myths and folklore in particular, is that we get to mentally experience something that happened (real or fictional), without having to experience it for ourselves first. When it comes to myth however, what we mentally experience today is (probably) fundamentally different to what the ancient recipients experienced. Our story-experiencer ancestors have one advantage over us, that is that they shared a collective mental model when they heard or read about a myth. Thus, mythological sources could afford to often leave out many aspects which were commonsense for the ancient recipient, that we cannot easily infer from our world-knowledge. We today have to carefully reconstruct what that mental picture was, in order to properly understand a myth. This is, among other things, a core task of the *mythological scholar*. We can see the *mythological scholar*, which I will call *domain expert* in this thesis, as the *archaeologist of the myth*.

Like the *archaeologist* on an excavation, the *mythological scholar* has a set of tools that they can employ for the extraction, categorization, and comparison of *myth Stoffe*. Some of those tools sets having been used traditionally for a long time, much like chisel and brush. Other tool sets are new and can help to better understand the core concepts of *myths* and their inter-relations with each other. One of those new *theories of myth*, based on narrative fundamentals, is the *hylistic*

[6] I distinctly exclude the reading of the term *myth* as “urban legend” or “something that is not true”.

theory [1, 21]. With the *hylistic* theory, a practical framework for the narrative analysis of myths across cultures is laid out, which will–among other things–ease the task to re-create the mental model of ancient peoples.

Analysing and comparing *Stoffe* is a challenging task, because variants of the same *Stoff* can be represented differently in different sources, they can be fully formulated or only alluded, sources vary in genre, language, and medium. The *hylistic* theory [1, 8, 21] (see Section 2.1) overcomes the different representations by introducing *hylemes* as "minimal action bearing units" [1, 21]. *Hylemes* are statements that describe an action, state, or information present in a *Stoff*-variant, e.g. "Orpheus goes into the netherworld". From these individual statements the *hyleme sequences*, i.e. the sequence of events and states in logical order, is constructed. One *hyleme sequence* represents one narrative variant. In result, these *hyleme sequences* can be used to compare variants of the same *Stoff*, as well as different *Stoffe*.

Digital, or computer-assisted study of (ancient) *myths*, especially with regard to their plot structure, is relatively scarce. Computational (Ancient) Mythological Studies[7] are often concerned with creating digital resources, e.g. collections of sources, or networks describing the interconnection of characters across sources. However, with *hylistic* data, it is possible to study *myth Stoffe* with computational methods, e.g. stemming from natural language processing or knowledge engineering, without having to process the source material in its original language, e.g. ancient Sumerian, which can be challenging.

For the *hylistic* approach, no dedicated digital toolset has been established yet. For that purpose, this work aims to investigate which digital methods can be applied to *hylemes* and *hyleme sequences* to facilitate the comparison of *Stoffe* and *Stoff variants*.

1.2 Research Objectives

The thesis is interdisciplinary in nature and is related to the fields of digital humanities (DH), computational linguistics (CL) and natural language processing (NLP), (digital) narratology, mythology and folkloristic studies, and to a certain extent classical studies (CS), and assyriology.

The study of myth through digital methods is challenging, because myths and folklore, widely do not follow commonsense reasoning. Concepts that are physically impossible in real-life, are accepted and presented as fair and realistic. Lévi-Strauss described this as: "Any characteristic can be attributed to any subject; every conceivable relation can be met." [14, p.429] Furthermore, he comes to the conclusion that the semantics of a myth variant can only be studied if the myth is studied as a whole, or at least as completely as possible. In order to model mythological concepts, we need to include linguistic considerations, but we cannot limit ourselves to them [14].

[7]Computational or Digital Mythology also occasionally refers to the study of *social* myths and their spread in the digital space, which is not what is studied in this thesis.

Hylistic analysis of a *Stoff* follows ten general steps [8], which are presented in Section 2.1. In this work each of those steps will be approached computationally in some way (outlined in Table 2.2 in Section 2.1).

Examples in this thesis are often drawn from existing *hylistic* research data (see Section 4), or from own examples. Occasionally, these free examples are drawn from the *Harry Potter* universe, because a) the broad audience can be safely assumed somewhat familiar with the content, and b) the story contains various mythological and narrative features that make excellent examples for many phenomena and their *hylistic* representation.

In this thesis, the comparison of *Stoff* variants and *Stoffe* (to a certain degree) will be approached by establishing different measures for *hyleme* and *hyleme sequence* comparison. *Hyleme sequences* are compared by means of sequence alignment, following the assumptions that similar *Stoff* variants share similar actions, agents, and settings.

The resulting alignment of sequences yields essentially three types of information: 1.) the *hylemes* that are aligned with each other contain the semantic overlap between *Stoff* variants, 2.) the *hylemes* that are not aligned contain semantic divergence of *Stoff* variants, i.e. the actions or states that are different in the sequences that are compared, 3.) the gaps in the alignment indicate that an action or state is present in one sequence, but not in the other. These kinds of information are an important basis for comparative myth analysis. Figure 1.2 shows an example for two sequences of *hyper-hylemes* (i.e. general plot units, see Section 2.1 and Section 6.3), with two gaps. The interpretation of those gaps rests on the domain expert. For instance, the gap can be the result of a style choice by the author of the source (e.g. to start a narration *in medias res*), or that the author assumes the familiarity of the recipient with the *Stoff* and therefore does not state the information (e.g. the descend into the netherworld). Another interpretation could be that the missing *(hyper-)hyleme* might be an optional *Stoff* unit that alludes to another *Stoff*, like *Orpheus being killed by the Thrakian women.*

Now, if we want to compare the *hylemes* directly, e.g. the *hylemes* that represent the *hyper-hyleme* "The inhabitants of the netherworld set conditions for the ascent of Orpheus and Eurydice", we see that these general plot units (*hyper-hylemes*) are conveyed by *hylemes* with a certain level of lexical variation, e.g.

- Sequence 86: Orpheus convinces Dis's wife by singing to the lyre.
- Sequence 107: Orpheus beguiles Pluto and Kore (= Persephone) with music.

We see that the *hylemes* use different predicates, and that the inhabitants of the netherworld are represented differently (Dis is an alternative name for Pluto). We also see that "singing to the lyre" is a specification of "with music".

In order match *hylemes* for the purpose of the alignment, these similarities need to be modelled. Therefore, *hyleme* pairs that are similar are identified by a set of different methods that is proposed

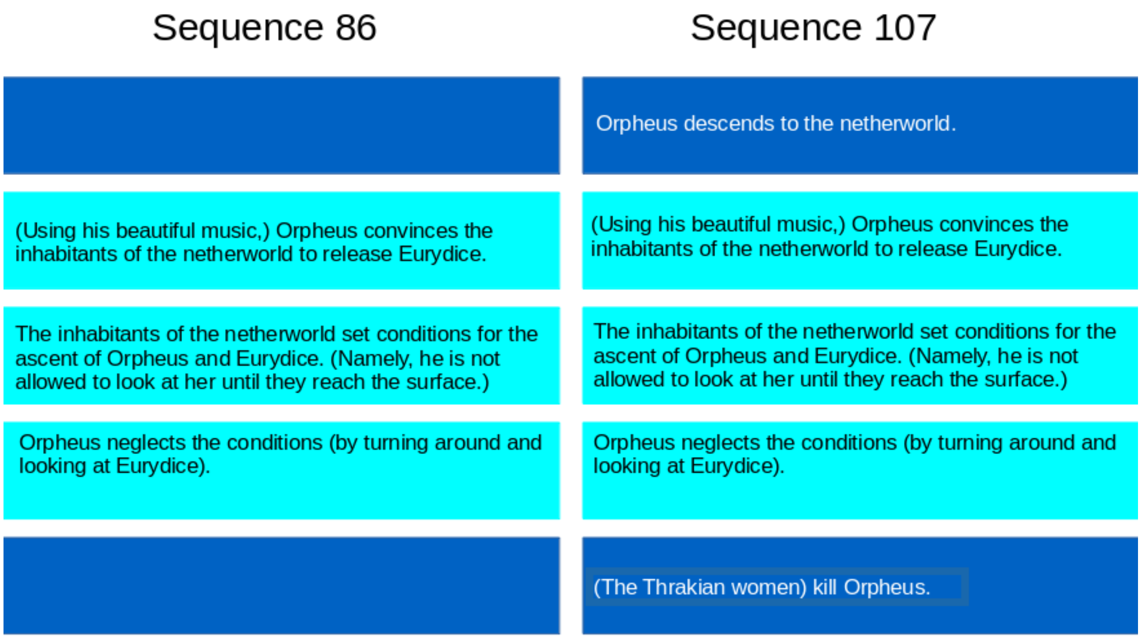


Figure 1.2: Example alignment of two sequences (see Section 6.3)

in this thesis. Depending on the purpose of the comparison from the domain expert's perspective these different methods can be combined, e.g. stricter methods where the domain expert wants to study nuances, and more lenient methods, where the domain expert is interested in finding general analogies (e.g. matching actions only).

Furthermore, this work makes use of the *hylistc* feature of *hyleme type* (see Section 2.1) to compare *Stoff*-variant-inherent background knowledge separately from the plot, by using knowledge engineering methods (*shallow* or *minimal ontologies*).

1.3 Contributions

Before I elaborate on the contributions of this thesis, I would like to clarify what this project is not: It would have been easy to take the data that was provided to me by the STRATA research group as it was initially presented to me, and from a computational linguistic standpoint assume it to be similar enough to some pre-existing research objective, such as event prediction or classification, then subsequently refine and improve existing approaches and algorithms. I could have added a few % F1-Score and leave it at that. However, this would have been short-sighted and a neglect of the theoretical foundation laid out by the *hylistic* approach. Instead, the main aim of this thesis is to approach *Hylistics* from a more holistic point-of-view. This work is the first large study of approaches towards *Digital Hylistics*. Every digital method that is used in this work has been carefully selected according to its appropriateness towards the theory, even if it came at a cost of being perceived as not complex enough or not computationally advanced enough.

That being said, the thesis presents the following contributions:

1. In this thesis, I open the field of *Digital Hylistics*, which will lead to many different research objectives in the future, as discussed in Section 8.1. Each of the methods and approaches I used in Chapters 4 to 7 can be expanded, refined, improved and applied to new *hylistic* data sets.

2. Additionally, I present a new data set, which is the first large-scale English *hyleme* data set and the first data set in the field of Folkloristics constructed using *hylistic* analysis, advancing the fields of *hylistic* research and Digital Folkloristics at the same time.

3. This work presents the first large-scale annotation study on *hyleme* types.

4. Additionally, I show how *hyper-hylemes* can be employed to construct a formal grammar that can describe a *Stoff*. For that purpose, a case study on *Orpheus and Eurydice* is presented.

5. Based on the *hyper-hyleme* sequences, this thesis presents the first case study of ideal *hyleme* sequence alignment, using 18 variants of the *Orpheus and Eurydice-Stoff*.

6. For the similarity of hylemes, the first *hyleme* semantic similarity annotation was performed.

7. In order to achieve automatic identification of alignment candidates (i.e. *hylemes* that are similar in two or more sequences), a set of methods is proposed and evaluated with respect to its applicability for different research questions and considerations.

1.4 Structure of the Thesis

This work is structured as follows: Chapter 2 gives an introduction to the underlying narrative, linguistic, and technical foundations of this thesis. In Chapter 3.1, I frame the *hylistic* theory [21] in the context of related narrative theories. I discuss other related work, e.g. on events and narrative structure modelling, and digital (humanities) approaches to myths and folktales in Chapter 3. Chapter 4 introduces the data sets that are used in this work, and demonstrates the results of an exploratory study of *hylistic* data in comparison with other data sets. Knowledge Engineering approaches for modelling background information in *Stoff* variants are presented in Chapter 5. Chapter 6 demonstrates an ideal modelling and subsequent alignment of *Stoff* variants pertaining to the *Orpheus and Eurydice-Stoff*. Computational approaches and experiments are presented in Chapter 7. This thesis ends with a discussion and conclusion in Chapter 8.

Chapter 2

Theoretical Foundations

This chapter gives an introduction into the theoretical foundations of this thesis. In Section 2.1 the narratological foundation, i.e. the *hylistic* theory is introduced. Related theories that are not directly applied in this thesis but are important to frame the *hylistic* theory, are discussed in Section 3.1 as part of the literature review. In Section 2.2 (computer-)linguistic foundations are introduced. The technical foundations, including natural language processing tasks and methods, are discussed in Section 2.3.

2.1 Narratological Foundations

Mythological contents are communicated in various ways. The obvious form, which directly comes to mind, is textual representation. But even texts are an exceptionally diverse medium, which can exist in different languages, genres, and forms. The study of myth becomes more complex when we take into considerations non-textual forms of media, such as Greek vase paintings, Mesopotamian cylinder seals, mural paintings, mosaics and other modes of image-based story telling.[1]

In this work, we apply the *hylistic* approach [1, 21], a unifying[2] theory to myth that overcomes the difficulties of working with intra- and inter-medial myth representations. *hylistic* analysis is performed by domain experts and aims to derive a representation of the core narrative (or story) from the discourse. For a discussion of different narrative theories, please refer to the Related Work Chapter 3.1. In a *hylistic* analysis, the media of the source and the textual representation, e.g. how colorful the wording is, is only of secondary interest. The reconstruction of the plot and related background information is performed on the plot relevant aspects of the source. With that, it becomes easier to compare versions of narrative variants that have fundamentally different styles.

The basic unit of analysis used in this thesis is the *hyleme*. A *hyleme* is a semi-formal statement

[1]This can be extended towards any other form of narrative media, for instance modern literature, comics, mangas, or video games.

[2]In this context *unifying* refers to language independent, genre-independent, and medium-independent.

pertaining to a variant of a narrative[3]. *Hylemes* are defined as "minimal action-bearing units" [1, 21].[4] A *hyleme sequence* is a *sequence* of hylemes reconstructed from a source, including implied statements, which reconstructs the plot and context of a narrative variant (*Erzählstoff*-version). The entirety of *(Erzähl-)Stoff*-versions pertaining to the same narrative material is the *(Erzähl-)Stoff* [8]. Since a specific *Stoff* can always be extended, by discovery of previously unseen versions of the past, and by extension and remix in the future, a *Stoff* is never complete. "The shape of a Stoff is open, both with a view to the past as well as into the future. [...] the potential in a particular Stoff is inexhaustible. The actual maximum of the complete spectrum of possible Stoff variants is infinite" [21, p.22].

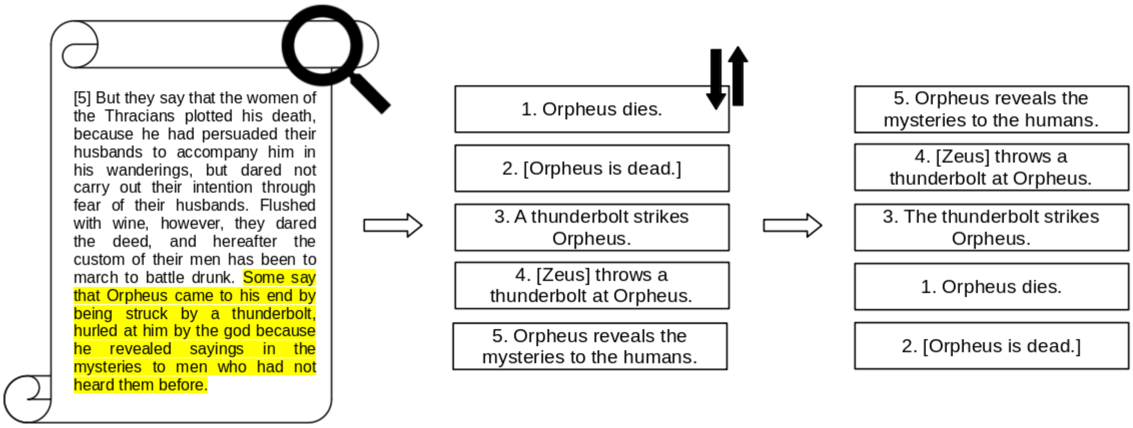


Figure 2.1: Basic steps of hyleme extraction, Example taken from Pausanias' description of Greece

In Figure 2.1 the basic steps of *hyleme* extraction are illustrated. Firstly, the domain expert identifies a potential narrative variant from a source. These sources can be presented in various forms of media, such as cuneiform tablets, paintings, or illustrated handwritten manuscripts. If sources are presented in textual form, they can belong to different genres, such as different kinds of prose, travelogues, mythological collections, poems, or ritual instructions.

Secondly, the domain expert extracts the *hylemes*. *Hylemes* are not direct quotes from a text, but abstractions or specifications of the content, referring to the plot or the background information (e.g. characterizations of locations and characters, relationships between characters, or habituals). A *hyleme* follows a semi-standardized form: *Hylemes* contain one finite main verb, are presented in active voice and present tense. There are four kinds of *hyleme components*: *subjects*, *predicates*, *objects* and *determinations* (of the subject, predicate, or object). For instance, the *hyleme Huge masses of water cover Tyros* can be split into the core components: *masses of water* (Subject), *huge* (Subject

[3]Since this work deals with different types of narrative, mainly folklore and myth, I use the term *narrative variant*, when I make a general statement, and *myth variant* if I make a statement that only applies to the mythological domain or a specific myth.

[4]Technically, only *single-event hylemes* (see below) are action-bearing units, *durative* hylemes contain background information that is important in the context of a (myth) narrative. Hence, *durative hylemes* can be defined as *action-context* units.

determination), *cover* (predicate), *Tyros* (Object).

Consequently, *hylemes* have at minimum one *subject* and a *predicate*. *Hylemes* may contain objects.[5] Entities that are not directly named in the source but perform an action or are otherwise plot-relevant, (like an unnamed narrator, e.g. "I saw the cat."-*NN sees the cat.*) are represented as *NN*.[6] Furthermore, *hylemes* always include explicit mentions of characters and places (where present in the source or inferable), that means that character and place names (*named entities*) are repeated, co-references are avoided unless they occur in the same *hyleme*. For example, *Innana fastens a brooch to* ***her*** *side* is acceptable but *Innana goes to the netherworld -* ***There she*** *looks for her husband* is not. *Hylemes* are independent statements, they are not connected to each other using discourse markers, such as *before* or *after*. However, *hylemes* are not *lexically* standardized.

The domain expert decides on the lexical representation of the *hyleme*, based on the source text, the conventions of the discipline (e.g. Ancient Near Eastern Studies ANES, Classical Studies (CS), or Religious Studies), and to a certain degree the preference of the expert. Hence, multiple sequences pertaining to the same narrative may be subject to lexical variation, paraphrases, summarization of multiple *hylemes* and slightly different inference of implied *hylemes*.

Hyleme extraction includes the inference of *implicit* or *implied hylemes* (represented by square brackets). In many cases, *hyleme extraction* is reiterated many times and community-driven by a group of domain experts. In that sense, *hyleme sequences* are extremely high quality data represented as a reconstruction and abstraction of a *Stoff*-version from a source. The *hyleme* extraction process is not always straightforward when the mythological domain is concerned, because sources do not always reiterate a complete narrative. *Hylemes* can be contained or entailed only by allusion ("Hades shows mercy to Alcestis, Protesilaos, Eurydice and Orpheus."), or subsumed under a broad statement ("When Orpheus went into the netherworld,..."). Additionally, some individual *hylemes* may also be represented by a large portion of the text (expansion of the hyleme [8]).

From the extracted *hylemes*, the domain expert orders the sequence chronologically (as in *logical* order, not necessarily the order in the source). This ordered sequence of events and states is the *hyleme sequence* pertaining to the narrative variant presented in the source.

Figure 2.2 shows how a *hyleme sequence* can be mapped to the corresponding text, here Plato *Symposium 179d*.[7] We see that not all *hylemes* and *hyleme* components are directly mentioned in the text. Some elements have to be inferred from the source or through domain and world-knowledge of the domain expert. For example, *Eurydice* is not directly mentioned in the text. Therefore, we have to infer her character from what we already know about the myth. Those *hyleme* elements or *hylemes* are therefore implicit.

[5] These categories do not exactly mean the same as the grammatical categories by the same name, e.g. the conjunction in *Dumuzi and Innana go to the netherworld* implies that the *hyleme* has two subjects.

[6] Multiple *NN* might be distinguished by indices, e.g. *NN1*, *NN2*

[7] Text: `http://www.perseus.tufts.edu/hopper/text?doc=Plat.+Sym.+179d&fromdoc=Perseus%3Atext%3A1999.02.0028` (CC-BY-SA), Annotation Tool: `https://catma.de/`

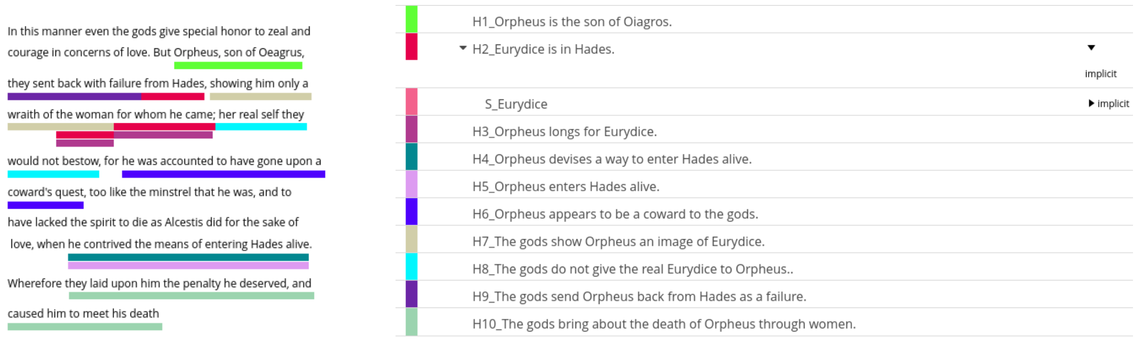


Figure 2.2: Plat. Sym. 179d, Orpheus and Eurydike, *hyleme* sequence (right) mapped to corresponding text (left), Tool: CATMA

The color coding in Figure 2.2 illustrates that the order of the *hylemes* in a sequence can differ greatly from the order of information in a text. The former denotes the core narrative or *Stoff*-version, while the second represents the order of the discourse. *Hyleme* sequences present the order of the narrative how it logically unfolds, not typically how it is told.

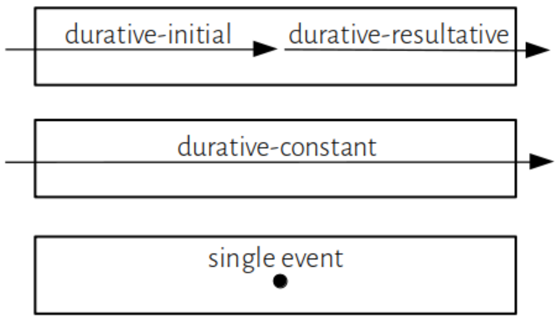


Figure 2.3: Types of hylemes

The domain expert assigns event categories/temporal semantic categories to the *hyleme sequence*, as illustrated in Figure 2.3. There are two categories of *hylemes* which differ in their truth value across the sequence. The *single-event* hylemes are true at exactly one point during the course of the sequence. *Durative* hylemes on the other hand are true over a certain temporal/chronological span of the sequence. *Durative* hylemes are divided into three sub-types: *durativie-initial* hylemes, which are true from before or at the beginning of the *hyleme* sequence, *durative-resultative* hylemes, which are true from a certain point in the sequence onwards, and *durative-constant* hylemes, which are true during the course of the entire sequence and possibly before and after the sequence has ended. Examples for the different *hyleme* categories are shown in Table 2.1.

Durative-resultative hylemes often define the boundaries of a *Stoff*-version [8]. They are the result of a *single-event* hyleme, or multiple preparatory hylemes, or the entire sequence. For instance, the *hyleme Harry Potter attends Hogwarts school of Witchcraft and Wizardry* in Table 2.1 is a state that concerns the protagonist, which is enabled by preceding hylemes, such as *Harry receives his Hogwarts letter, Harry learns that he is a wizard* and *Hagrid takes Harry away from the Dursleys*. Similarly, *durative-intial* hylemes define a state at the beginning of a narrative variant. Their truth value changes over the course of the narrative. Thus, both *durative-initial* and *durative-resultative* hylemes are context-sensitive. They are connected to *single-event* hylemes by causal-links. *Durative-initial* hylemes can for instance serve as the *motivation* of a character to perform certain actions (*single-event* hylemes), and *durative-resultative* hylemes are the consequent states invoked by an action, e.g. *Harry kills Voldemort, Voldemort is dead. (durative-resultative)*

Table 2.1: Examples for the different *hyleme* types

Hyleme	Hyleme Type
"Harry Potter is a wizard."	durative-constant
"Harry Potter lives in 4 Privet Drive, Little Winging, Surrey."	durative-initial
"Harry Potter boards the Hogwarts express."	single-event
"Harry Potter attends Hogwarts school of Witchcraft and Wizardry."	durative-resultative

Another important additional concept in *Hylistics* is the *hyper-hyleme*. A *hyper-hyleme* is an abstraction and summarization of a number of plot components, "represent[ing] longer episodes or even an entire *Stoff*" [21, p.40]. When contained directly in a source text, they are often used to allude to a part of a *Stoff*, or summarize a group of events. They can be present in a source in the "style of a chapter heading" [21, p.40], or by referencing a known *Stoff* by mentioning specific, identifying details, e.g. *when the minstrel went into the netherworld* as a reference to the *Orpheus-Stoff*.

On the *hyleme sequences*, the domain experts can then perform inter- and intra-myth comparison. There are ten core steps that scholars perform in order to *hylistically* analyse and compare a *Stoff*-variant, subsequently a *Stoff* and finally different *Stoffe* [8]. This includes the *hyleme* extraction, categorization and ordering, but also comparison along various dimensions, including the *TTPPE* conditions: *topics*, *time*, *place*, *protagonists*, and *events*. Each of the ten steps have some form of representation in this thesis. In Table 2.2 the steps are listed with the corresponding chapters of the thesis.

As of date[8], *hylistic* analyses have been applied in the mythological domain [1, 23, 24] and folkloristics (in this thesis). Many other potential domains and areas of application are yet unexplored. I outline potential projects, and their significance for Natural Language Processing (NLP) and Computational Linguistic (*CL*) research areas in Section 8.1.

[8]February 13, 2024

Table 2.2: Ten steps for *hylistic* narratological analysis [8] and representation in the thesis

No	Analysis Step [8]	Aspect	Chapter
1	Identification, categorization, standardization	Identification, standardization Categorization	4.1 4.5
2	Analysis of textual representation	Comparison of hylemes and hyper-hylemes	6.2.3
3	Textual ommisions vs. complete narrative material	Folktale hyleme data set data set construction	4.1
4	Order of the text vs. order of the narrative material	Folktale hyleme data set data set construction	4.1
5	*Erzählstoff*-versions, Changes in *TTPPE* conditions	Generative Grammar, Named Entity Recognition, Hyleme matching approaches	6.1.1 4.3 7
6	Combination of multiple narratives in a text	Folktale hyleme data set data set construction	4.1
7	Analysis of compression ratio	Modelling of hyper-hylemes	6.1.1
8	Stratigraphical analysis of narrative material(s)	Minimal-Ontologies	5
9	Comparative analysis of different versions of narrative material	Ideal alignment, Hyleme-matching	6.3 7.2
10	Inter-hylistic and Inter-textual analysis	Minimal-Ontologies	5

2.2 Linguistic Foundations

In this section, the most relevant linguistic foundations for the work presented in this thesis are introduced. This includes tasks (e.g. Verb Lemmatization), concepts (e.g. Verb aspect), and measures (e.g. Cohen's κ for inter-annotator agreement).

2.2.1 Verb Lemmatization

Verb lemmatization is the (largely language-dependent) task of finding the *lemma* of a verb from its inflected forms. Hence, a *lemma* is a "group of word forms that are related by being inflectional forms of the same base word" [25, p.245]. Consequently, *lemmatization* is either a manual task (text annotation), or a NLP task. For instance, the English word 'fed' is the past particle of 'to feed'. Therefore, the *lemma* is *feed*. The performance of *lemmatization* methods and algorithms is dependent on the complexity of the language, mainly how prevalent inflections are. There are different approaches towards *lemmatization*, dictionary approaches perform well on common verbs, but usually fail to account for rare words and neologisms. Rule-based approaches employ inflection tables and/or morphosyntactic or grammatical rules. They generalize well and can potentially be applied to unseen words. However, they do not perform well on irregular verbs (such as "was" as the simple past of "to be"). Machine learning approaches, especially those employing neural networks, need a lot of training data to predict *lemmata* correctly. Under-resourced or low-resource languages suffer from data poverty and might have to use data leverage approaches, such as using data from related languages or retraining models originally developed for related tasks or languages (*transfer learning*). Alternatively (or additionally), intricate feature engineering, and/or a combination of dictionary-based and rule-based approaches with machine learning methods can yield promising results.

2.2.2 Verb Aspect and temporal-semantic values of utterances

The term *aspect* describes a universal (i.e. language-independent) grammatical feature of a verb. *Aspect* is a grammatical category that defines the contrast between perfective and imperfective meaning. *Aspect* is either inflectionally or syntactically signaled in a sentence [26]. In that sense, *aspect* "can be described either as a completed whole, or as something 'ongoing, in progress' or simply 'existent' for a given point in or period of time." [26, p.13] *Aspect* is related, but not strictly the same as *Aktionsart*, which is "the manner in which some event is integrated in the imagined stream of time." [26, p.14] This includes the question whether an event is *telic* (aimed at achieving a goal) or *atelic*. *Aktionsart* is a lexical category, describing the manner in which some event takes place [26].

Moens and Steedman define the *aspectual class* or *type* as a semantic category that "a speaker predicates of the particular happening that their utterance describes, relative to other happenings in the domain of the discourse" [2, p.16]. In that, *verb aspect* is the feature which determines the temporal-semantic value of a sentence or utterance in the context of a discourse. More specifically, it defines if a statement concerns a point in time, or an interval which is complete or ongoing.

Moens and Steedman [2] derive four fundamental classes of *contigency*: *culmination*, a punctual action or event that results in a change of state, *point* expressions with no change of state in the context of the narration, *process* expressions which have no immediate result in the discourse, and *culminated processes* that have a defined result (or culmination). Grammatical tense alone often does not give sufficient information to determine the temporal-semantic value. Consider the following examples which all use simple past:

- Tom Riddle woke up the Basilisk. (culmination, change of state)
- Harry fought Tom Riddle. (culminated process)
- Harry walked over to Ginny. (point, no change of state)
- Harry waited for a while. (process)

However, tense can influence the (re-)interpretation of an utterance, e.g. progressive auxiliaries invoke processes, even if the base form of the verb indicates a *point* (*hit/was hitting*) and perfect auxiliary may transform an utterance into a culmination (*run/had run*). Other modifiers, such as temporal adverbials (e.g. *often*) may also influence the temporal-semantic value.

From the *contingency classes*, Moens and Steedman derived their definition of *events* as "happenings with defined beginnings and ends." [2, p.17] In contrast to events, *states* have no defined beginning and end (within the context of a discourse or narration). They can be classified into *habitual states* (ongoing processes), *consequent states* (results of culminations), *progressive states* (processes that are in progress at the time of narration), and *lexical states*. Figure 2.4 illustrates the *contingency classes* with a few example verbs that commonly assume a type.

	EVENTS		STATES
	atomic	extended	
+conseq	**CULMINATION** recognize, spot, win the race	**CULMINATED PROCESS** build a house, eat a sandwich	understand, love, know, resemble
-conseq	**POINT** hiccup, tap, wink	**PROCESS** run, swim, walk, play the piano	

Figure 2.4: Contigency Classes and event/state distinctions, Source [2, p.17]

Related annotation studies concerning aspectual classes/event models, are discussed in Section 3.3.1.

2.2.3 Predicate-Argument Structures

Predicate-argument structures (PAS) are semantic representations of a verb and its arguments derived from phrases. For instance, *Harry boards the train* induces the predicate-argument *boards(Harry, train)*, in which the predicate *boards* defines the relationship or action performed between the agent *Harry* and the patient *train*. Inflected forms of the predicate (*boarded, boards, has boarded*) are represented by the same predicate-argument structure. The selectional preferences of the predicate influence which assumptions can be made about the arguments, for instance that the patients is some form of a transport vessel. The argument slots in a predicate-argument structures are usually filled with the (proto-)agent and (proto-)patient in a phrase. Intransitive verbs (e.g. *sleep*) have a single argument structure (*sleeps(Harry)*).

However, predicate-argument structures can suffer from ambiguity in the source phrase, e.g. *Ron and Harry are old friends*, in which *old friends* is a predicative that can be represented as *old(Ron)* $\wedge$ *old(Harry)* $\wedge$ *friends(Ron,Harry)* or *long_time(friends(Ron, Harry))*.

Predicate-argument structures are a way to represent semantic primitives, which can serve as a basis of many knowledge engineering and natural language processing tasks. Some event modelling approaches using predicate-argument structures are discussed in Section 3.3. Predicate-argument structures are related to the subject-predicate-(object) structure of *hylemes*. For the data and approaches in this thesis, *hylemes* and their components are the fundamental semantic primitives and contain the vocabulary for the knowledge representation. However, world knowledge on selectional preferences might not always be fully applicable for the *mythological* and *folkloristic* domain, because some predicates may take arguments that are not commonsense in real world circumstances, e.g. *build(animal, palace)*.

2.2.4 Frame Semantics

A *frame* pertains to a mental concept that is invoked by a lexical item, e.g. a noun or a verb. The *frame* of a concept generalizes instances in which that concept is used, through a set of lexical items that refer to the same frame, e.g. *sleep*, *nap*, *snooze*, *catnap*, and abstracts semantic roles that co-occur when an utterance refers to that frame. For instance, consider the FrameNet [27][9] entry for *Sleep* as presented in Figure 2.5. It defines the core frame entities (*FEs*), in this case the *Sleeper*, i.e. the entity that sleeps. The core *FEs* co-occur when the frame is invoked.[10] Non-core *FEs* do not automatically co-occur when a frame is invoked, for instance the mental picture of sleep does not automatically include a time-frame or a location.

Frames follow a hierarchical structure, that means that the *Sleep* frame has a parent frame, *Sleep-wake-cycle*, which includes other frames, such as *Waking Up*. Through this framework, specific knowledge can be incorporated into frames, by further specification of a concept through a sub-

[9] `https://framenet.icsi.berkeley.edu/`

[10] If an utterance were to just refer to the general subject, e.g. *Sleep is important* it would not neccesarily invoke the frame.

Sleep

Definition:

The Sleeper stays in an altered state of consciousness with greatly reduced external awareness.
We SLEPT.

FEs:

Core:

Sleeper [Slp]	The entity that is sleeping.

Non-Core:

Degree [Deg] **Semantic Type:** Degree	The degree to which the Sleeper is asleep.
Duration [Dur] **Semantic Type:** Duration	This FE identifies the Duration of time for which a state holds or a process is ongoing.
Manner [Manr] **Semantic Type:** Manner	Any description of the sleeping event which is not covered by more specific FEs, including epistemic modification (probably, presumably, mysteriously), secondary effects (quietly, noisily), and general descriptions comparing events (the same way). Fitzhume SLEPT soundly
Place [Place] **Semantic Type:** Locative_relation	Where the event takes place.
Time [Time] **Semantic Type:** Time	This FE identifies the Time when the sleep occurs.

Figure 2.5: FrameNet entry for *Sleep*

frame. A concept inherits the slots from its ancestor frames, and can incorporate new slots (*FEs*) [28]. Thus, frames can be used for common sense reasoning and as a resource of world-knowledge.

Frames usually concern one static concept, like a specific situation or activity. They are related to *scripts* in that a frame can have a dynamic procedure (*script*) attached to it, which models prototypical everyday situations, e.g. the *Restaurant* frame can have a *Eating-at-restaurant script* attached to it. *Scripts* are thus an extended common sense knowledge representation. They have been employed in event modelling and prediction in different ways. *Scripts* as event representation frameworks are discussed in Section 3.2.

2.3 Technical Foundations

In this section, the technical foundations of this thesis are introduced. Many methods originate from the field of natural language processing and computational linguistics, but some stem from other disciplines of the computer sciences (CS), such as bioinformatics, knowledge engineering or theoretical informatics (formal languages). First, the relevant (string) alignment and string distance measures are introduced. Then the NLP tasks of named entity recognition and (verb) lemmatization are explained, before this section moves on to explain methods of distributional semantics which can be applied in semantic similarity estimation tasks. We introduce topic models, that are applied later in this thesis in Section 4. Lastly, knowledge graphs and semantic nets are introduced.

2.3.1 Sequence Alignment and String Distance Methods

Alignment Methods

Early sequence alignments were used in biological science to find a plausible alignment between two sequences, which is the first step towards establishing a similarity judgement of two DNA sequences or protein structures. DNA and protein sequences undergo evolutionary change, which results in additions, deletions or substitutions of parts of the sequence. For instance, DNA sequences are constructed from pairs of nucleobases (ACGT[11]). They can be represented as long strings where each literal represents one base (pair). Alignment of multiple DNA sequences is used to find indications on whether or not two (or more) sequences are related [3].

Long string or sequence alignment methods use a scoring function to achieve the optimal (local or global) alignment between two sequences. These scoring functions commonly assign scores to three phenomena in a potential alignment: the match, the mismatch and the gap. Gaps often occur when one of the sequences includes an additional item (e.g. a base) stemming from an insertion/deletion, that is placed between two partial strings that both sequences have in common. Gaps mainly occur when the sequences that are to be compared are of different lengths. For instance, the string *AAAATTTT* would be aligned to the string *AAAACTTTT* using a gap in the fifth position: *AAAA-TTTT*. Each possible alignment is assigned a score, based on the operations performed to transform one sequence into the other. In our case, the *cost* of the alignment *AAAA-TTTT* would be 1 (if we assume that a score of 1 is assigned to a gap). Another possible alignment is *AAAA**T**TTT-* (one mismatch, one gap). This alignment would be assigned a score of 2 (if we assume a gap cost of 1 and a mismatch cost of 1). Hence, the alignment *AAAA**T**TTT-* is more costly and therefore less favourable than the alignment *AAAA-TTTT*. However, in evolutionary biology certain mutations are more likely than others, so scoring functions may take these factors into account by introducing probabilities [3, p.2].

[11] Adenine, Thymine, Guanine and Cytosine

When processing long sequences of DNA, protein structures or other biological sequences, the scores are often arranged in a substitution matrix. A common matrix design is BLOSUM (Blocks Substitution Matrix) [29]. If deviations between sequences are punished using a negative score, the alignment task is a maximization problem. If positive scores are assigned, it becomes a minimization problem. The overall goal is to find alignments that preserve as many matching elements between the sequences as possible. Therefore, in the above example the alignment *AAAA-TTTT* is preferable over *AAAA**T**TTT-*.

From the scores assigned to possible alignments between two (or more) sequences, alignments can be produced. For that purpose, we can differentiate betewen *global* and *local* alignments. A *global* alignment approach is used to find the optimal alignment between two (or more) complete sequences. A popular global alignment algorithm is the Needleman-Wunsch algorithm [30, 31]. The Needleman-Wunsch algorithm is a *dynamic programming* approach that tries to find the optimal alignment between two complete sequences (*pair-wise alignment*).

The Needleman-Wunsch algorithm constructs an alignment matrix F from two sequences $s_1 = x_1, .., x_m$ and $s_2 = y_1, ..., y_n$. After initializing the matrix $F_{0,0} = 0$, the matrix is filled recursively using previous alignment solutions for partial sequences $x_1, .., x_i$ and $y_1, ..., y_j$, starting from the top left [3], or bottom right [30] cell. For each index i, j the best alignment up to that index is calculated based on the previous values $F_{i-1,j}, F_{i,j-1}$, and $F_{i-1,j-1}$.

$$F_{i,j} = max \begin{cases} F_{i-1,j-1} + s(x_i, y_j), \\ F_{i-1,j} - d, & [3] \\ F_{i,j-1} - d. \end{cases} \tag{2.1}$$

The gap cost d can be assigned either to sequence x or y, and $s(x_i, y_j)$ is the cost of a match or a mismatch.

The score in the final cell can then be interpreted as the score of the optimal alignment. The alignment itself can then be derived by traversing the matrix back from the final value to the first value (*traceback*). From each index (i, j) the traceback returns to the cells from which the value was derived, i.e. either $(i-1, j), (i, j-1)$, or $(i-1, j-1)$. Figure 2.6 shows an example matrix F and the traceback for the optimal alignment for the sequences *HEAGAWGHEE* and *PAWHEAE*. The optimal alignments for the sequences are *HEAGAWGHE-E* and *–P-AW-HEAE*.

The computational cost to fill one cell of the matrix is four operations (three sums and a maximum), the matrix has $(m+1) \times (n+1)$ cells (where m is the length of sequence s_1 and n is the length of sequence s_2). Therefore, the computational complexity of the Needleman-Wunsch algorithm is $\mathcal{O}(nm)$.

In contrast to a *global* alignment, which operates on full sequences, *local* alignment algorithms only align partial sequences (*subsequences*), mainly those that are most similar. For sequences that

		H	E	A	G	A	W	G	H	E	E
	0	**−8**	**−16**	−24	−32	−40	−48	−56	−64	−72	−80
P	−8	−2	−9	**−17**	**−25**	−33	−42	−49	−57	−65	−73
A	−16	−10	−3	−4	−12	**−20**	−28	−36	−44	−52	−60
W	−24	−18	−11	−6	−7	−15	**−5**	**−13**	−21	−29	−37
H	−32	−14	−18	−13	−8	−9	−13	−7	**−3**	−11	−19
E	−40	−22	−8	−16	−16	−9	−12	−15	−7	**3**	−5
A	−48	−30	−16	−3	−11	−11	−12	−12	−15	**−5**	2
E	−56	−38	−24	−11	−6	−12	−14	−15	−12	−9	**1**

Figure 2.6: Example Needleman-Wunsch score matrix and alignment, Source: [3, p.21]

are by default very dissimilar, local alignments help to identify common patterns. In this work, the Smith-Waterman [32] algorithm is used for local alignment. The Smith-Waterman method is similar to the Needleman-Wunsch approach, with the added constraint that the cells of the matrix F can also assume 0 values.

$$F_{i,j} = max \begin{cases} 0, \\ F_{i-1,j-1} + s(x_i, y_j), \\ F_{i-1,j} - d, \\ F_{i,j-1} - d. \end{cases} \quad [3] \tag{2.2}$$

The context window of the alignments are defined through the occurrence of the 0 values. A 0 value is assigned, if all other options would yield in a negative score. As soon as a cell assumes a value of 0, a new local alignment starts. The corresponding sequence for the best alignment is then not derived from the bottom right corner, but from the highest value in F, to the closest cell containing a 0 value. This way, the Smith-Waterman approach yields the best local alignment between two sequences. Figure 2.7 illustrates the alignment on the same example as used in Figure 2.6.

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

Figure 2.7: Example Smith-Waterman score matrix and alignment, Source: [3, p.23]

String Distance Methods

In contrast to the *alignment* methods described in Section 2.3.1, two sequences of characters (i.e. strings) s_1 and s_2 of arbitrary lengths can be *compared* for the assessment of string similarity. One way of deriving numerical values for string similarity is through string *distance* measures. Two of those methods are used in this work: the *Levensthein* [33] distance (edit distance) and the *Jaccard* [34] distance.

The Levensthein oder Edit Distance is a method that calculates the minimal number of single character operations–insertions, deletions and substitutions–needed to transform one string into the other. For instance, the strings *inherent* and *interest* can be transformed into each other using two operations, the substitutions *h-t* and *n-s*.

For longer, more complex strings a dynamic programming approach based on a similar method as described for the Needleman-Wunsch algorithm might be used. For that purpose, a matrix $D(s_1, s_2)$ for the strings $s_, = x_1, ..., x_m$ and $s_2 = y_1, ..., y_n$ is constructed, using the values of the previous cells and a cost function that defines the cost ω for the basic operations (deletion ϕ of x_i, substitution of x_i and y_j, and deletion ϕ of y_j).

$$D_{i,j} = min \begin{cases} D(x_{i-1}, y_j) + \omega(x_i, \phi), \\ D(x_{i-1}, y_{i-1}) + \omega(x_i, y_j), \quad [33] \\ D(x_i, y_{i-1}) + \omega(y_j, \phi). \end{cases} \tag{2.3}$$

To find the edit trace (which is a form of alignment), the matrix D traversed in the same manner as discussed in the previous section. String distance methods might add further constraints to possible solutions, e.g. that there can be a maximum of two consecutive deletions. Furthermore, each elementary operation might be assigned a weight, through which one operation might be punished more severely than the other, e.g. favouring gaps over insertions/deletions [33].

It has to be noted that string distance methods do not take semantic information into account, but strings that represent words might occasionally be semantically related if the string distance is low, e.g. because they share prefixes or have the same word stem ("escalate/escalation"). However, the opposite is not necessarily true ("boat/boot").

The *Jaccard coefficient* [34] is a measure originating in set theory. It is worth mentioning that the Jaccard coefficient is not per se a string distance measure, it can be applied to any set of items. In contrast to the alignment methods, and the Levensthein distance, the Jaccard coefficient does not consider the order of items. Instead, it is the intersection of two sets S_1 and S_2 over the union of both sets.

$$Jaccard(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{2.4}$$

If both sets contain the same items the Jaccard coefficient is 1. That means that for sequences containing a finite set of elements in different amounts and positions, like DNA sequences, the Jaccard coefficient is not suitable. For semantic similarity of words the coefficient is not particularly suited either, because the order of elements is not taken into account, e.g. the strings *ARTEMIS* and *SMARTIES* have a Jaccard coefficient of 1. For sentences, the Jaccard coefficient might be more suitable, if the grammatical function of words can be neglected (*bag of words*), because repetitions of the same salient words are less frequent.[12] Hence, the sentences *The cat sat on the mat* and *On the mat sat the cat* also have a Jaccard coefficient of 1. In this case, the distance measure is a more realistic approximation of similarity.

[12] Less salient words, e.g. determiners, might occur more frequently.

2.3.2 Named Entity Recognition

Named Entity Recognition (NER) is the task of automatically detecting and labelling named entities in textual data. The label refers to the type of entity. It is assigned according to a labelling scheme relevant to the data in question. Consider the following example:

> "Cooks Fähigkeiten brachten ihn gut voran. Mitte 1754 wechselte er unter finanziellen Einbußen zur Royal Navy, bei der er als Able Seaman auf der HMS Eagle anheuerte."[13]

Translation:

> "Cook's skills allowed him to advance well. In mid-1754, he transferred to the Royal Navy at a financial loss, where he signed on as Able Seaman on the HMS Eagle."[14]

This sentence can be annotated with different Named Entity labels, as follows:

> "$[_{PER}\ Cooks]$ Fähigkeiten brachten ihn gut voran. $[_{TIME}\ Mitte\ 1754]$ wechselte er unter finanziellen Einbußen zur $[_{ORG}\ Royal\ Navy]$, bei der er als $[_{TITLE}\ Able\ Seaman]$ auf der $[_{MISC}\ HMS\ Eagle]$ anheuerte."[15]

Here, the labels *PER*, *ORG*, and *MISC* are used for names of people, organisations, or other entities. Together with the *LOC* label for names of geographic locations, they form the core of many label sets. Additionally, the *TIME* label is used for time designations. CoreNLP's [35] regexner tool includes additional, more fine-grained labels, such as *TITLE* for (job) titles, or *NATIONALITY*. Named Entity Recognition (NER) approaches can either rely on manually annotated data, which is time consuming and costly, or use a semi-supervised or unsupervised approach. The WikiNER [36] model, which is used for the NER task in the spaCy[16] pipeline, uses a semi-supervised approach, in which they classify Wikipedia articles using Wikipedia's internal link structure, and features such as capitalisation. The performance on the WikiNER classifier, as part of the spacy pipeline and the three different spacy models for German[17] and English[18] are analyzed in Section 4.3.

[13] This example sentence was taken from the small Wikipedia data set that is used for comparison in Section 4 `https://de.wikipedia.org/wiki/James_Cook`
[14] Translated by the author.
[15] The tagset used for this example is part of the Stanford CoreNLP NER tagset. [35]
[16] `https://spacy.io/`
[17] *de_core_news_sm*, *de_core_news_md*, and *de_core_news_lg*
[18] *en_core_web_sm*, *en_core_web_md*, and *en_core_web_lg*

2.3.3 Word Embeddings

Word Embeddings are a word-in-context representation based on distributional semantics, derived from studies of synonymy and semantic similarity as a function of contextual correlates ([37, 38], see Section 3.6).

Word Embeddings represent a term by the context of terms that commonly co-occur with it. For example, the term *Quidditch* might not to be found in traditional dictionaries. If we wanted to derive the meaning of *Quidditch*, we could investigate in which contexts the term occurs. To illustrate this, the following four examples (derived from the English *Harry Potter* Wikipedia entry[19]) are given with a context window of four tokens.

1. ... tents put up for *Quidditch* tournaments are similar to ...
2. ... version of the sport *Quidditch* was created in 2005 ...
3. ... Conn used Snape's and *Quidditch* coach Madam Hooch's teaching ...
4. ... purported Hogwarts textbook) and *Quidditch* Through the Ages (a ...

We can see that the term *Quidditch* occurs alongside other terms like, *tournaments*, *sport*, and *coach*. From this "neighbourhood" we can already draw some conclusions what the term means. From terms in the context windows, a term-context vector (of dimension $(1, n)$ where n is the size of the vocabulary) can be created for the term *Quidditch*. In our case, the vector with be of dimensions (*1,32*).

While the *term-context* matrix combines all terms in the vocabulary, the *document-term* matrix contains all terms in the vocabulary across a collection of documents, containing counts of how often the terms occur in each document. The *term-context* or *document-term* matrix is the *embedding* of the terms or documents, where each vector corresponds to the contextual representation of a term or a document. The embeddings are thus a mapping of the n terms of the vocabulary to R^n. This means that two terms that occur in the same context are represented by close vectors (see below Section 2.3.3).

In the simplest form, the *term-context* matrix contains counts of how often two terms co-occur within a given context-window. The dimensions of the matrix are the unique terms of the vocabulary in the rows and columns. The items in the vocabulary depend on the research question, e.g. all terms in a document, only salient terms in the document (e.g. removing stopwords), or only lemmata. In our example, we would probably want to summarizes all occurrences of terms like *game* and *games*, or *catch*, including *catching* with the term *Quidditch*. Therefore, we would preprocess the text using a lemmatizer and only count co-occurrences of the lemmata *game* and *catch*.

The resulting vector is sparse, e.g. a lot of terms are part of the vocabulary[20], but never co-occur

[19] `https://en.wikipedia.org/wiki/Harry_Potter`

[20] According to hacker news user Robin_Message the seven *Harry Potter* books have a vocabulary of 21,441 unique words.

with our target term *Quidditch*. For that reason, dimensionality reduction algorithms, e.g. t-SNE (t-distributed stochastic neighbor embedding [39]), can be employed to reduce the computational complexity of processing those term vectors.

Word embeddings have the advantage that they are well suited for visualisation and can be employed for explorative studies through their arithmetic properties (*King - Man + Woman = Queen*). A disadvantage is that they are sensitive to bias in the data (most famously [40]).

Word2Vec

The most commonly used word embedding approach is *Word2Vec* [41]. *Word2Vec* uses neural networks (recurrent neural networks, RNN and feed-forward neural networks FFNN), either following a continuous bag-of-words (CBOW), or SkipGram architecture, where "the CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word." [41, p.5] The resulting vectors are of "modest dimensionality" between 50-100.

A *Word2Vec* embedding can only access vectors for terms that are part of the training vocabulary. It cannot infer vector representations for unseen words, and thus struggles with neologisms or in domain adaptation tasks, i.e. tasks in which the training data stems from a different domain than the application or test data.

FastText

Another approach that is used in this thesis (see Section 7.2.10) is *FastText* [42]. *FastText* produces embeddings on a character-level (*character n-grams*), which means that subword information, e.g. prefixes or stems, are included.

Through subword information, *FastText* can create a vector representation for unseen words. This makes *FastText* applicable for rare words or neologisms, as long as they use a similar morphological structure as words in the training language[21]. However, *FastText* is not very well suited for deriving semantic information. Although it can group related terms together even if they are used in different contexts, e.g. because one is a verb and the other is a noun (*build/building*), close vectors imply a *morphological* similarity rather than a *semantic* relationship.

`https://news.ycombinator.com/item?id=2775838`

[21] For the word *Quidditch*, FastText would probably be of little help, because it has a very unique structure that is uncommon for the English language.

Measuring Distance of Word Vectors

We can calculate the distance between two vectors v, w of dimension N in an embeddings space using the cosine similarity [43, Chapter 6, p.11]:

$$cos(v, w) = \frac{v \cdot w}{|v||w|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{2.5}$$

An alternative is the Euclidean Distance, defined as the *L2*-Norm of the difference between the vectors v and w:

$$||v - w||_2 = \sqrt{\sum_{i=1}^{N}(v_i - w_i)^2} \tag{2.6}$$

It has to be noted that two terms that are represented by close vectors are not necessarily semantically similar. A proximity entails that the terms are used similar *context*, which is usually also true for *antonyms*. For example, the context *He ... the chess game* is representative for both *won* and *lost*.

2.3.4 Knowledge Graphs and Semantic Nets

This work uses two lexical-semantic nets to retrieve synonym information: *WordNet* [44, 45] and *GermaNet* [46, 47].

WordNet[22] is an English database in form of a semantic net that models hierarchical relationships: *synonymy* (equivalence), *hyponymy* (*isA*-relationship), *meronymy* (*part-whole* relationship, e.g. leg-chair), *antonymy* (for adjectives, e.g. good/bad). *WordNet* includes relationships across different parts-of-speech, e.g. verbs and their corresponding adjective (*care/caring*).

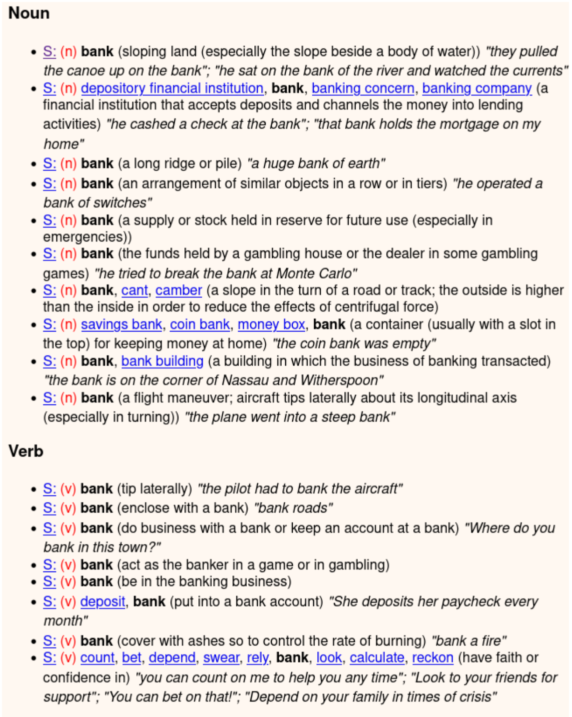
Different readings of the same lexical item are represented by different *synsets*. For each *synset* example sentences and descriptions (*glosses*) identify which reading the entry corresponds to. Figure 2.8 gives an example for a WordNet query for the term *bank* with its different synsets across parts-of-speech (noun/verb). The results are ordered according to their frequency (i.e. how common a reading is), beginning with the most frequent. From Figure 2.8 we can thus derive that the word sense *bank of a river* is more common than *bank as financial institution*. *WordNet* contains 155327 unique strings, and 117597 synsets.[23]

GermaNet is the German equivalent to *WordNet* [46, 47]. It contains currently[24] 167163 synsets, and 215000 lexical units. Both WordNet and GermaNet make use of word classes, i.e. semantic fields or

[22] https://wordnet.princeton.edu/

[23] February 13, 2024, https://wordnet.princeton.edu/documentation/21-wnstats7wn

[24] February 13, 2024, https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/



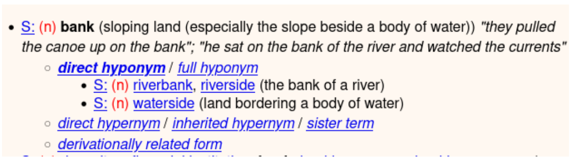
Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Verb

- S: (v) **bank** (tip laterally) *"the pilot had to bank the aircraft"*
- S: (v) **bank** (enclose with a bank) *"bank roads"*
- S: (v) **bank** (do business with a bank or keep an account at a bank) *"Where do you bank in this town?"*
- S: (v) **bank** (act as the banker in a game or in gambling)
- S: (v) **bank** (be in the banking business)
- S: (v) deposit, **bank** (put into a bank account) *"She deposits her paycheck every month"*
- S: (v) **bank** (cover with ashes so to control the rate of burning) *"bank a fire"*
- S: (v) count, bet, depend, swear, rely, **bank**, look, calculate, reckon (have faith or confidence in) *"you can count on me to help you any time"; "Look to your friends for support"; "You can bet on that!"; "Depend on your family in times of crisis"*

Figure 2.8: Example query “Bank” in WordNet



- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
  - ***direct hyponym*** / *full hyponym*
    - S: (n) riverbank, riverside (the bank of a river)
    - S: (n) waterside (land bordering a body of water)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

Figure 2.9: Example query “Bank” in WordNet, first synset, direct hyponym

categories, such as *Cognition/Kognition* to classify concepts.[25]

Graph-based Similarity Measures

Lexical-semantic nets like *WordNet* and *GermaNet* can be employed for measuring semantic similarity of concepts. The measures are usually referred to as *knowledge-based measures*. They can be grouped roughly into three groups [48]:

- Edge-counting approaches
- Feature-based measures
- Measures based on Information Content (IC)

[25] see http://www.sfs.uni-tuebingen.de/projects/ascl/GermaNet/germanet_structure.shtml#Tops

The most simple example of edge-counting is to find the minimal path between the two concepts [48, 49]. It carries the notion that two concepts a and b which are semantically similar are connected by a shorter path than two concepts which are not as similar.

$$Sim_{rada} = \min_{\forall i} |path_(a, b)| \tag{2.7}$$

A feature-based measure that is applied not only to knowledge-bases was introduced by Lesk, originally for the word sense disambiguation (WSD) task [50]. It computes the word overlap between two dictionary-like definitions (e.g. *synset glosses*). Thus, its application to semantic nets disregards the hierarchical structure completely and only focusses on the *glosses*. Consequentially, its performance is highly dependent on the quality of the *glosses*, and the use of controlled vocabulary, i.e. similar concepts are described with similar words. It is also sensitive to the length of the *definition* or *gloss*.

The notion of *Least Common Subsumer (LCS)* between two concepts is important for many knowledge-based similarity measures. It refers to the most specific concept that is ancestor to both concepts, i.e. both concepts have an "inhereted" *isA*-relationship to the LCS, and the path (number of *isA*-links) between the concepts and the LCS is shorter than to any other common ancestor. For instance, if the concepts *table* and *chair* have a common ancestor *furniture with four legs* which is a hyponym of *furniture*, then the LCS of *table* and *chair* is *furniture with four legs*.

An example for *Information Content (IC)*-based measures was introduced by Resnik [51] as

$$Sim_{res} = IC(LCS), \tag{2.8}$$

with the *information content* (IC) defined as:

$$IC(c) = -logP(c) \tag{2.9}$$

and $P(c)$ as the probability of the concept c in the corpus [52].

All similarity measures that utilize lexical-semantic nets in the style of *WordNet* have the disadvantage that the corresponding *synset* needs to be known. That means that a word sense disambiguation (WSD) step needs to be performed before the resource is queried, (or with help of the resource, e.g. by using Resnik's approach [51]). As Figure 2.8 illustrates, a query can potentially yield many results, and glosses and examples might yield too little information to reliably automatically disambiguate the *synsets*.

2.3.5 Topic Models

Finding common *topics*, or *semantic fields*, in a (large) collection of text is a useful method for document clustering and data exploration. The general idea behind topic models is to automatically uncover the underlying themes or subjects that are present in a large corpus of text data without the need for human annotation or manual categorization. The key assumption behind topic models is that each document in a collection is a mixture of different topics, and each topic is characterized by a distribution of words. In other words, documents are composed of a combination of different topics, and topics are represented by the words that tend to co-occur within them. We interpret a *topic* as a collection of terms that are related to each other, or "a distribution over a fixed vocabulary." [53, p.78] The terms do not necessarily have to be inherently semantically related, e.g. *gold*-*goose*, but share relevance to a topic t. In this case, a *golden goose* is an animal in a popular fairytale ([54], KHM 64). *Topic modeling* or *topic discovery* is usually realised as an unsupervised machine learning task, where *topics* are assigned to unannotated documents, which aims to find the *hidden* topic structure over a collection of documents [53].

Topic models can be applied to a variety of tasks. Intuitively, topics can be inspected to gain insights about the text corpus a topic model has been applied on. Topics can also serve as a foundation to cluster to documents with similar content. Furthermore, topic-based document retrieval can identify documents that are most representative of a specific topic. Topic models can also be applied as a subtask in text summarization, where words representing certain content in a document or document collection are used as a seed for natural language generation.

LDA

A common approach towards automatically deriving a *topic model* for a collection of documents is Latent Dirichlet Allocation (LDA) [53, 55]. LDA is a *probabilistic topic modelling* approach, in which a distribution of N topics over a set of M documents is learned from the *hidden topic structure* and the observed structure, i.e. the terms in the documents. LDA interprets a document as a combination of different (*latent*) topics with varying weights.

Following this document interpretation, LDA learns two distributions: The term-topic distribution, i.e. which terms refer to which topics and how are they weighted, and the topic-document distribution, i.e. which topics are contained in which documents.

For that purpose, it uses two hyper-parameters, the prior Dirichlet parameter α, which defines the document-topic density, and topic-word density β.

A schematic overview of the LDA process and its components is given in Figure 2.10.

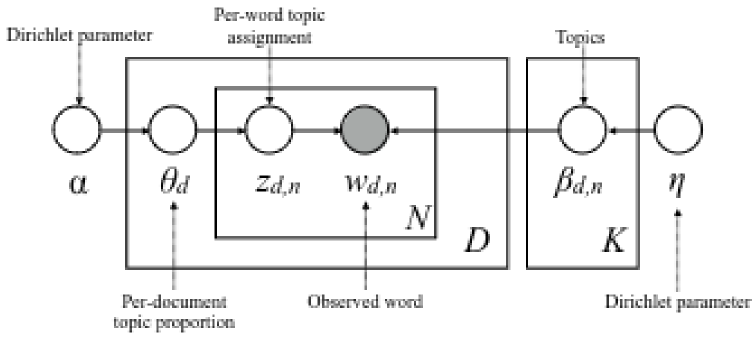


Figure 2.10: Schematic overview of LDA, Source: [4]

Perplexity and Coherence

In order to find the best hyper-parameter setting for the document-topic density α and the topic-word densitiy β, two measures can be employed: *perplexity* and *coherence*.

On a test set of M documents $D = w_1, ..., w_M$ with N topics, the perplexity, based on the probability of the unseen data given the learned topic model $p(w)$, is calculated as follows [55]:

$$perplexity(D) = exp\left\{ - \frac{\sum_{d=1}^{M} log(p(w_d))}{\sum_{d=1}^{M} N_d} \right\}. \tag{2.10}$$

On the other hand, coherence is a measure for the informativeness and interpretability of a topic. There are different measures that model coherence (see [56]). In this work, I use C_V, which is introduced by as "combin[ation of] the indirect cosine measure with the NPMI and the boolean sliding window" [56, p.405]. The cosine similarity is used as defined in equation 2.5.

The *context vectors* are composed of elements $v_{i,j}$:

$$v_{i,j} = NPMI(w_i, w_j)^{\gamma} = \left(\frac{log\left(\frac{P(\omega_i,\omega_j)+\epsilon}{P(\omega_i)\cdot P(\omega_j)}\right)}{-log(P(\omega_i,\omega_j)+\epsilon)} \right)^{\gamma} \tag{2.11}$$

where $P(\omega_i, \omega_j)$ are the probabilities that the terms ω_i, ω_j co-occur within the boolean sliding window, with ϵ added to avoid $log(0)$. [56] γ refers to the weighing of the NPMI value.

We can interpret *perplexity* as the measure of how well the model assigns topics to a text that it has not seen before, and *coherence* as the measure of how much sense a human user can make of the topics.

2.3.6 Inter-Annotator and Inter-Rater Agreement

For many research objectives in the natural language processing field, human judgements are crucial for the establishment of a gold standard against which automatic approaches can be compared. However, depending on the task at hand, e.g. verb lemmatization or named entity recognition (NER), the human subjects (*annotators* or *raters*) often disagree in their judgement. This discrepancy between annotators can have different reasons:

- Human error or malicious annotations (e.g. always selecting the first/same option)
- Incomplete, or inconsistent task description (e.g. missing information on edge cases in the annotation guidelines)
- Judgements are influenced by subjective factors, e.g. human sentiment or unconscious preference of one class over another
- Ambiguity in the items that are to be annotated (or rated)

For instance, the sentence "Thomas Cook had to evacuate the passengers" might refer to the person or the travel agency *Thomas Cook*. Hence, annotators might assign either *PER* or *ORG* labels if this sentence is presented in a named entity labelling task.

Therefore, the "level of disagreement" between raters or annotators is an important measure that has to be taken into consideration when human judgments and annotated labels are used in downstream tasks. Measures of this "level of disagreement" are ofter called inter-annotator or inter-rater *reliability*. However, this carries the notion that discrepancy stems from *unreliability* of the human subjects, which is not always the case, as listed above. The observed discrepancy between human judgments can be measured. In this work, two measures are used: Cohen's κ for judgements between two annotators, and Fleiss' κ for judgments of more than two annotators.

Cohen's κ is a measure for the level of agreement of two human annotators on categorical items [57]. He identifies two relevant values for the measure of agreement:

- "p_0 = the proportion of units in which the judges agreed,
- p_c = the proportion of units for which agreement is expected by chance." [57, p.39]

The measure is then presented as:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}.[57] \tag{2.12}$$

This means, that $\kappa < 0$ indicates that the agreement between the annotators is less than the expected chance agreement. Perfect agreement $\kappa = 1$ is achieved if the chance agreement is 0, and the annotators agree on all items.

Fleiss' κ extends Cohen's measure to n (two or more) annotators, on N items to be annotated with k categories. [58].

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}, \tag{2.13}$$

where the chance agreement between annotators is

$$\overline{P}_e = \sum_{j=1}^{k} p_j^2. \tag{2.14}$$

Let n_{ij} be the number of annotators who assigned the category j to item i. Then the proportion of all assignments to category j is defined as

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{2.15}$$

and the overall proportion of agreement between annotators is

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i. \tag{2.16}$$

which is the mean over the agreement of pairs of assignments, defined as

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{2.17}$$

Landis and Koch [9] provide a benchmark interpretation for the strength of agreement measured by κ statistics, as presented in Table 2.3.

Table 2.3: Landis and Koch [9] interpretation of the strength of agreement

Value	Strength of Agreement
< 0.0	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	(Almost) perfect

2.3.7 Performance Measures for Machine Learning Models

When machine learning models or other automatic approaches are employed, their performance is commonly measured against human judgements or other gold standard data. In supervised machine learning systems, this is commonly achieved by splitting the labelled data into a training and test set, often a 80 % - 20 % ratio. The model is then tested on the test data, the predicted labels are compared to the gold standard labels. With regard to a label l, a *true positive* is a correctly l-labelled item, a *true negative* is an item that is correctly not assigned the label l. A *false positive* is an item that is incorrectly assigned the label l. A *false negative* is an item that is incorrectly not labelled l, when it should be. *TP*, *TN*, *FN*, and *FP* are the corresponding cardinalities of the sets of *true positives*, *true negatives*, *false negatives* and *false positives* from the predictions that the model supplied for the test set.

The simplest measure for the performance is *accuracy*. For binary classes, the accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.18}$$

In multi-class tasks, it is the ratio of correctly predicted items out of all items.

On the other hand, the precision is a measure of how many of the predicted items are correct. A high prediction value means that if a model predicts a class, it is usually correct.

$$Precision = \frac{TP}{TP + FP} \tag{2.19}$$

In contrast to precision, the recall (sometimes *sensitivity*) is a measure for the portion of correctly identified items from all potential items (of the class in question). A high recall means that most of the item of a class are found.

$$Recall = \frac{TP}{TP + FN} \tag{2.20}$$

The F1-score combines precision and recall using the harmonic mean:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.21}$$

2.3.8 Term-Frequency Inverse Document Frequency

Term-Frequency Inverse Document Frequency (short *TF-IDF*) is a statistical representation of frequency and saliency of terms in a set of documents. Originating in document indexing and retrieval, the measure was proposed to discriminate between frequently occurring, but not discriminating terms, and terms that occur frequently across all documents in a set.

For instance, consider the seven core-canon *Harry Potter* books. Undoubtedly, the terms *Harry* or *Ron* are salient terms. However, they occur frequently across all books and hence provide little meaning, whereas the term *Quidditch* occurs in different frequencies across the books. While it occurs relatively equally in the first three books, the fourth book starts with the *Quidditch world championship*, which leads to an increased frequency. In the fifth book, the sport is cancelled for the school year, and hence gets not mentioned as often. Likewise, in the last book, *Harry* does not attend school and cannot play the sport. Therefore, mentions of the term *Quidditch* carry different saliency across the books. For indexing purposes, this means that "a great variation in term distribution is likely to appear. It may thus be the case that a particular term becomes less effective as a means of retrieval. [...] A frequently used term thus functions in retrieval as a non-specific term, even though its meaning may be quite specific in the ordinary sense." [59, p.13]

For that purpose *TF-IDF* is used as a weighting mechanism for terms in documents. It is calculated from two measures, the *term frequency* and the *inverse document frequency* (the *specificity* [59]).

The *term frequency* (*TF*) can be calculated either using simple counts of terms in a document, binary values (a term occurs or does not occur), or use a normalized or otherwise scaled frequency. The term frequency is given as the relative number of times a term t appears in a document d. For simplicity purposes, we assume here that $f(t,d)$ are absolute counts of the term t in d[26]:

$$TF(t,d) = \frac{f(t,d)}{\sum_{t' \in d} f(t',d)} \tag{2.22}$$

The *inverse document frequency* (*IDF*) is then the inverse number of documents d that contain the word, where N is the number of documents. *IDF* is scaled logarithmically.

$$IDF(t,D) = log\left(\frac{N}{1 + DF(t)}\right) \tag{2.23}$$

where the document frequency *DF* is the number of documents d in all documents D that contain the term t.

The two statistics are then combined (among others [43, Chapter 6, p.13]):

$$TF - IDF = TF(t,d) \times IDF(t,D) \tag{2.24}$$

[26]As for example used in the machine learning library sci-kit learn `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html`

Chapter 3

Related Work

This chapter first gives an overview on related narrative theories in Section 3.1. While the *hylistic* theory that introduces *hyleme* sequences for the formalization of mythological or other narrative content that is the foundation of a *Stoff* has been discussed in Chapter 2, Section 3.1 will frame the theory within selected other narrative theories, formalisms and cognitive theories.

In Section 3.2, related work that is methodologically relevant to the research objective in this thesis is discussed. The focus lies on methods that either try to solve a sub-problem of the research objective of this work, such as event modelling, or methods that are applied on similar data or similar genres to the ones studied in this work. It has to be noted that most of the related works use a wide definition of the term “narrative”, meaning any form of coherent chains of events.

The works cited and described in this sections are grouped by topics and then ordered roughly by significance to the research objective of this work.

3.1 Narrative Theories

3.1.1 Morphology of the Folk Tale

Vladimir Propp [60] proposed a model tailored to the Russian Magic Tales of Alexander Afanasyev [61]– The *Morphology of the Folk Tale*. In it, Propp proposed 31 invariant functions including a number of subfunctions that describe characteristic events in the plot of a folktale. These functions are grouped into five categories: *Preparation, Complication, Functions of the Donor, Struggle, Dénouement*. Propp himself defined initial function denoted by the Greek letter α not strictly as a function, but merely as setting the context of the tale, e.g. what is commonly expressed by phrases like “Once upon a time”. Proppian functions are defined by the action they express and by the archetypal characters that perform the action, their “spheres of action”. The *Dramatis Personae* are predefined in the theory as: *Hero*, *Villain*, *Dispatcher* (who relates the task to the

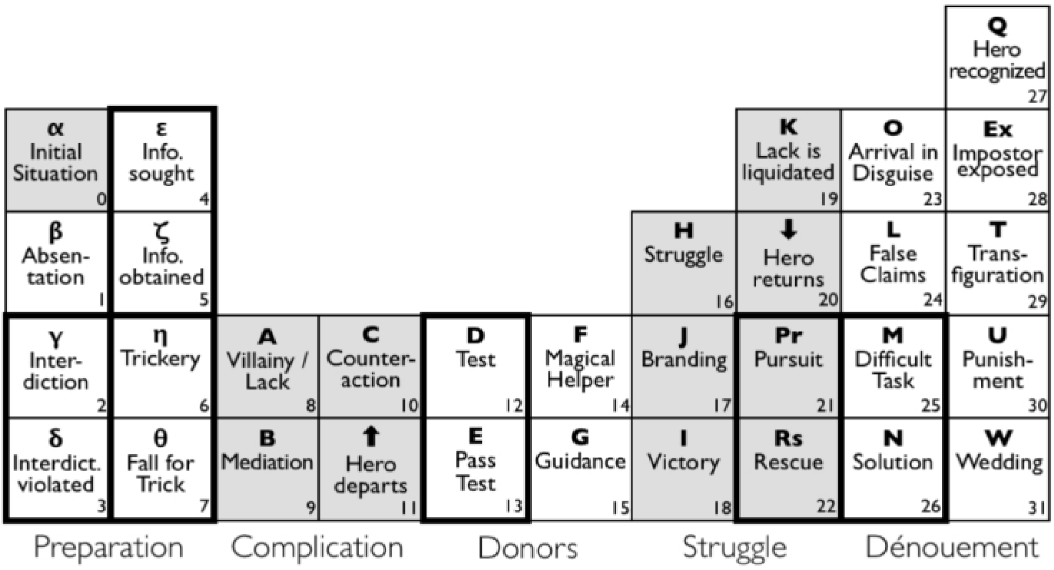


Figure 3.1: Propp's 31 invariant functions grouped into the five main categories, bold lines indicate pairings of functions that tend to appear together, Source: Antonia Scheidel [5]

Hero), *Helper*, *Donor* (who provides the *Hero* with a magical agent), *False Hero* (who poses as the *Hero*), *Princess* (and her father). Proppian functions appear strictly in order, but not all functions have to be present. Certain functions also relate to other functions, such as η *Trickery* and θ *Fall for Trickery*. Fig. 3.1 illustrates main functions, their grouping and their links to other functions. Function sequences build the frame of a tale. They are represented by a sequence of the literal representations of the functions.

Despite criticism of the Proppian formalism, it has been applied to tales from different cultural backgrounds, e.g. [62, 63, 64] and different genres of narration, e.g. [65, 66, 67]. Particularly interesting for the research objective of this work is how Burkert [10] applies the Proppian structure seamlessly to the myth of *Ištar's decent into the netherworld*. Burkert constructed the sequence of functions shown in Table 3.1[1]

Table 3.1: Burkert's [10] Proppian analysis of Ištar's descend into the netherworld

Function	Propp	Explanation/Realization in [10, pp.66]
β Absentation	A family member leaves the home.	Ištar focuses her sense towards the land of no return.
ϵ Reconnaissance	The villain seeks information about the hero.	The gatekeeper informs Ereškigal.
ζ Information gain.	The information is obtained.	
η Trickery	The villain tries to trick the victim.	
θ Fall for Trick	The victim falls for the trickery of the villain.	
A Villainy/Lack	The villain harms the victim or causes the loss of some sort.	Ištar is trapped in the netherworld, all life on earth is in danger.
B Mediation	The villainy is related/The hero is notified or instructed.	Ea received the message, and he creates assinu resp. kuluu; in the Sumerian version kurgarra and kalaturra.
C Beginning Counteraction	The seeker is ready.	
$\uparrow$ Departure	The hero departs from home	
G Guidance	The hero is led.	The emissaries fly like flies through the gates, they slip like lizards beneath the pillars.
H Struggle	The hero and his opponent compete with each other.	Instead of an open fight, the oponent is defeated by means of a ruse through the forms of civilised hospitality.
J Branding	The hero is marked/branded.	Ištar curses assinu.
I Victory	The villain is defeated.	
K Lack is liquidated	The missing object is won, the villainy is resolved.	
$\downarrow$ Return	The hero returns.	
Pr Pursuit	The hero is chased/pursued.	

[1]The explanations of the representations in the myth are taken from [10] where given and translated by the author.

The function *G Guidance* in Burkert's [10, pp.66] Proppian analysis of *Ištar's descend into the netherworld* is technically not correctly annotated. Since the *Guidance* function is a function of the *Donor*, it only applies if the *Hero* is guided towards a magical agent that helps him or her achieve the overall objective and defeat the villain. From the representation in [10] it seems as if the *Guidance* is interpreted as guidance towards the resolution of the *Lack*, and not a magical agent.[2]

3.1.2 The Hero's Journey

In his influential 1949 work, Joseph Campbell proposed a similar approach to Vladimir Propp's Morphology of the Folk Tale, which he calls *The Hero's journey* [20]. This archetypal journey is an objective for the Hero in a mythological context. Campbell suggests the term *Monomyth*. His *Helper* and *Mentor* characters are similar to Propp's archetypal characters of *Helper* and *Donor*. However, in contrast to Propp's approach, *the Hero's Journey* is more generalized and more applicable to different types of narratives, because it relies less on domain specific functions and character types, such as the *Donor*. In that, it aims to provide a general largely culture-independent framework for narratives. Campbell's theory has a stronger focus on the *Hero*'s perspective, such as the character's personal and psychological growth, while Propp's functions also include different aspects of antagonistic and auxiliary characters, such as the *False Hero*.

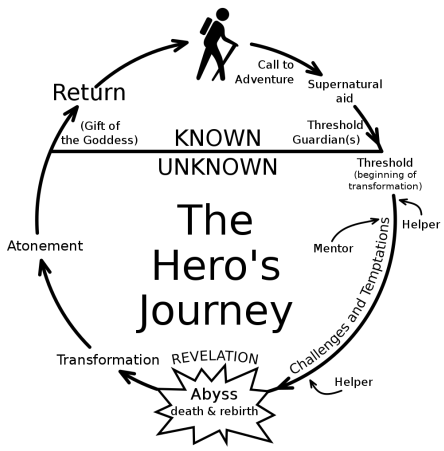


Figure 3.2: The Hero's journey, Source: Wikipedia, Hero's journey, Public Domain

3.1.3 Levi-Strauss

Claude Lévi-Strauss structuralist theory of mythology had a profound impact on various disciplines, including anthropology, linguistics, and literary studies. His most influential works are a collection of talks, *Myth and Meaning* [17], and the essay *Structural Anthropology* [14, 68].

[2]The aspect of Ištar's descend into the netherworld where the emissaries fly through the gates is not present in the respective *hyleme* sequence in the *hyleme* data set.

Lévi-Strauss' argues for the structuralistic analysis of myth through gross constituent units (*mythemes*), that are *bundles of relations*, i.e. different representations of semantically related relations. These *mythemes* make up the elementary structure of the myth, their context units are sentences. Thus, they are myth-inherent generic plot units. Variants of the same myth (*Stoff*) are grouped into permutation groups in order to account for inverse relations (opposites) existing between two variants. Lévi-Strauss' theory of mythology is funded on the concept of *Binary Opposition*, like *gods/men*, and *correlation*, varying elements appearing in similar contexts.

He argued that myths are not just random collections of independent stories but are actually narratives structured around two fundamentally opposing poles, such as *good/evil*. These oppositions represent deep-seated cognitive patterns that reflect the way human minds categorize and understand the world. For example, light vs. dark, life vs. death, nature vs. culture, tradition vs. innovation etc.

Using *mythemes* and their binary oppositions and correlations, he identified underlying universal patterns and structures in myths (and other related narratives). By isolating and categorizing these fundamental narrative elements, he sought to uncover the deep-seated cognitive processes that shape human storytelling and belief systems across cultures.

3.1.4 Narrative Clauses (Labov and Waletzky)

Labov and Waletzky [69] aimed to analyse the fundamental principles of narrative structures from natural speech. For that purpose they conducted 600 interviews with participants of different ethnicities and age groups, who spoke very informal English. In the interviews, participants where prompted to tell personal stories and relate past personal experiences. Labov and Waletzky then abstracted a framework of fundamental principles of narrative from the data on a clause level.

As a basis, they assume that in order for a narrative to stay the same, the related (i.e. told) events need to be semantically equal to the perceived (inferred) sequence of real events. They then define displacement as the potential range of positions that an independent clause can take before the semantic interpretation of events changes. For instance, *I am hungry-I eat the pie*, cannot be changed without changing the perception of events. *I am hungry-I turn on the television* can be changed, because it is perceived that being hungry can co-occur with turning the television on. The possible positions that a clause can assume is called the *displacement set* (DS). Based on the values of the displacement set of a clause, they define certain clause types:

- *Narrative clauses* have to remain in their position in the narrative.
- *Free clauses* can appear anywhere in the narrative.
- *Coordinate clauses* are succeeding clauses which can swap positions.
- *Restricted clauses* can appear in different positions, but are not unrestricted like free clauses.[3]

[3]Displacement in hyleme sequences can be applied, if the displacement results in a valid new sequence. The validity of

A *temporal juncture* occurs if two clauses are temporally ordered wih respect to each other. (*The cat eats the mouse-The mouse is dead.*). Labov and Waletzky define *narrative* as a sequence of clauses $c^1, ..., c^n$, which contains at least one temporal juncture, i.e. *a-then-b*. In a discourse, the temporal junction is often marked by lexical or grammatical forms, e.g. “and then”. They interpret a clause’s finite verb as the narrative head. The verbs tense are an indicator of the clause type.[4]

3.1.5 Affect States

If narrative structure is considered from a cognitive perspective, Lehnert [70] follows an approach of defining plot units using *affect states*, which are a distinction between *positive* +, *negative* - and *mental M* events, where *M* is the state of neutral or null emotionality. In combination with *causal links*, *actualization a*, *equivalence e*, *termination t*, and *motivation m*, *affect states* build 15 basic patterns, so called *standard affect configurations/primitive plot units*. If a plot unit concerns more than one character, cross-character causal links are used to connect the corresponding affect states. In turn, these basic plot units can be used to build more complex *general plot configurations*. Affect states do not always correspond directly to events, but they are closely related, because events often trigger affect states.

These plot unit configurations can be visualised as directed graphs, which allows comparison of abstract narrative patterns. An example of a basic pattern of a “primitive plot unit” and a general, more complex configuration is given in Fig. 3.3. The resulting abstract patterns can be placed in sequential order to form a representation of a narrative. Lehnert uses these patterns to assess the quality of human summaries. However, using graph comparison algorithms, they can also be used to assess semantic similarity of two or more plots.

Affect states have been automatically modelled by Goyal et al. [71, 72]. They present the AESOP

---

the sequence S after displacements is based on evidence from a *Stoff*. Not every perceivable sequence that is interpretable is automatically valid. Therefore, a sequence S after displacements (S') validity is not defined by retaining the original semantic value of the sequence (i.e. S and S' necessarily pertain to the same order of events), but validity within the scope of the *Stoff*. That being said: if we consider all hylemes in a variant, based on the scope of the *Stoff* they belong to, Labov and Waletzky’s categories can be applied to hylemes as follows:

- *Durative-constant* hylemes are *free* clauses, they can appear anywhere in a sequence.
- *Single-event* hylemes are *narrative (= fixed)* clauses, unless observed otherwise in the *Stoff*. Their potential range of displacement is defined by the evidence across sequences, e.g. in the *Orpheus Stoff* in Section 6.1 the hyper-hyleme z has a dispacement range of $0z1$ (can be moved one position up).
- If a *single-event* hyleme is observed in different positions in the narrative variants concerning a *Stoff*, it is *restricted*.
- *Single-event* hylemes can never be *free* clauses, because a *Stoff* is never complete, i.e. there is always the possibility that there are unseen variants of a *Stoff* of undetermined length (e.g. infinite possible slots).
- Two (or more) *durative-initial* hylemes h_1, h_2 are *coordinate* clauses. *Durative-initial* hylemes are always at the beginning of the sequence. If a sequence contains *durative-initial* hylemes after *single-event* hylemes, it is likely that a) the *Stoff* order of the sequence is annotated wrongly or b) the sequence actually describes more than one narrative or c) the hyleme type is not *durative-initial*, but an *intermediate* state (single-event).
- Two (or more) *durative-resultative* hylemes h_1 and h_2 *can be coordinate* clauses if they follow each other directly in at least one sequence, and h_1 does not invoke h_2, e.g. *Eurydice remains in Hades.-Orpheus is heartbroken.* If h_1 and h_2 are invoked by the same hyleme h_0, they are *coordinate* clauses.

[4]This does not apply to *hylemes*, because they are generally in present tense.

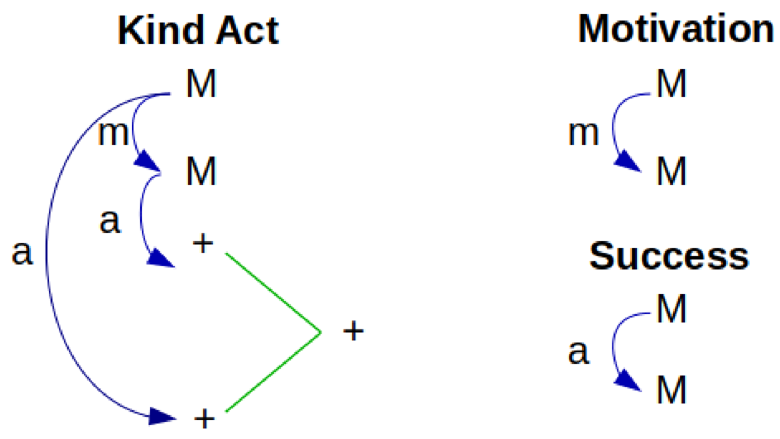


Figure 3.3: Left: General plot configuration “Kind Act”, Right: Basic plot units “Motivation” and “Success”

system, which generates plot unit representations in four steps: 1. affect state recognition, 2. character identification, 3. affect state projection, and 4. link creation. For that purpose, they introduce patient polarity verbs, i.e. verbs that invoke a certain polarity on their patient, e.g. “devour”. Goyal et al. evaluate their system on a set of Aesop’s fables.

3.2 Script and Event (Chain) Modelling

Scripts are a form of cognitive knowledge organization. They are structures through which humans can determine which actions, interactions and reactions are involved in specific situations, such as ordering food in a restaurant [73]. Internalised scripts are a fundamental tool that humans employ to experience their surroundings, but also for reading and listening comprehension. From the sentence “I asked the waiter to give me a glass of water” the recipient will be able to infer the contextual information that the utterance refers to a sequence of events which takes place in some form of a restaurant. The coherent chain of events that are mentally invoked by the utterances, e.g. the waiter writing down the order, the waiter leaving, and coming back with a glass of water, is called a script. Scripts undergo cultural and temporal change. For instance, the utterance “James called Mary” will most likely invoke a different script today than 100 years ago, and “Let us pray” invokes different scripts depending on the religious circumstances of the sender and receiver of the utterance. Whether or not a *restaurant*-script includes a *tipping-the-waiter* event similarly depends on the cultural background. Hence, scripts that are invoked through an utterance or referenced in a text are subject to the *sender*’s expectations and *receiver*’s internal knowledge system, as well as the general and circumstantial context and cultural background of both.

Schank and Abelson define a script as follows: “A script is a structure that describes appropriate sequences of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylized everyday situations." [73, p.210]

Furthermore, new references to objects in a script do not need to be introduced by an indefinite determiner, because they are already implied in the script, e.g. "When I arrived at the restaurant, *the* waiter took me to *the* table." Scripts are constructed from one point-of-view, e.g. *being a customer at a restaurant* induces a different script than *being a waiter at a restaurant*.[5] In computational narratology, scripts have been studied intensively. The approaches most related to our research objective are included in this section. However, it has to be noted that there are certain problems that arise when we try to model scripts for the fictional, especially the literary and the mythological/folkloristic domain. Mainly, the world-knowledge inherent to a fictional world often significantly differs to what is applicable in the real world. More importantly though, scripts may not be transferable across domains. For instance, the task of *brewing a potion* differs widely across cultures, belief systems, and temporal and geographical origins (and for the fictional domain also the corresponding fictional universes) of a narrative.[6]

Regneri et al. [74] have proposed an unsupervised method to learn events from different textual representations and their temporal order, so called Event Sequence Descriptions (ESDs) that make up a script. They employed Amazon Mechanical Turk[7] to collect data from non-expert users who were asked to describe stereotypical events such as "making scrambled eggs". By applying Multiple Sequence Alignment (MSA) algorithms to identify phrases relating to the same event, they construct a temporal script graph (TSG). Sequence descriptions that describe the same script are aligned in a matrix, where events that are omitted in one sequence are represented as a gap. Notably, they group phrases that relate to the same event but are not necessarily direct paraphrases. For instance, "wait for number to be called" and "wait for order" both belong to the same event of waiting for food while at a (fast food) restaurant. However, Regneri et al. did not distinguish between gaps that are the result of one sequence not mentioning an event and those gaps that the alignment produces because one event description summarizes an event where another event description breaks it down into sub-events. Nodes in the graph are merged so that the TSG has no gaps left, but the merge only considers nodes from different ESDs. This leads to some merges (clusters) overlapping semantically. In order to determine the weights for their cost function for event alignment, Regneri et al. [74] used scores based on WordNet (e.g. 100 for synonyms, 0 for lemmata without relation). They found that the predicate position held the greatest significance to the similarity. They found groups of verbs, such as support verbs like "get" to have smaller influence on the semantic similarity.

[5]For a deeper understanding on frames and scripts, I recommend the chapters III.C.7 (Semantic Primitives) and III.C.8 (Frames and Scripts) in [28]

[6]For a thorough discussion on the applicability of scripts in the literary domain, I would like to refer the reader to [22, p.106].

[7]`https://www.mturk.com/`

In a similar task, Regneri et al [75] aligned sentences describing an action with corresponding video sections relating to everyday kitchen-tasks. One of the subtasks of their study was to align different phrases describing the same or similar activities, such as "washing a carrot". For that purpose, they considered different verbalisations (e.g. "to wash" and "to rinse"). They created subsets of their data in which either the activity and the object matches, or only the object matches, or the activity and the verb describing it matches, but the objects differ from each other. In order to determine the semantic similarity of their text-based models, they employed the Jaccard coefficient (see Section 2.3.1) and a distributed vector-model. Regneri et al. [75] achieved the best results for semantic similarity of acitivity descriptions using a contextualized vector model that takes all constituent content words of a sentence into account.

Jans et al. [76] proposed an unsupervised method to identify scripts from a text, and predict the likelihood that an events belongs to a script. In order to achieve this goal, they use skip-grams for collecting statistics on the events present in a text. The prediction is subsequently achieved by means of a ranking function f. For that purpose, they compare PMI-based ranking functions [6], and bi-gram probabilities [77]. They tested their approach on the Reuters Corpus, Volume 1[8] and Andrew Lang's Fairy Tale Corpus[9]. Therefore, Jans et al.'s work is also related to the works discussed in Section 3.5. Across different evaluation measures, their bi-gram probabilities outperform Chambers and Jurafsky's approach (discussed below) that employs PMI [6].

Ostermann et al. [78] propose a method to map script events to their textual representation. For that purpose, they employ a machine learning approach based on crowd-sourced data related to scripts. They distinguish script-related events from unrelated events by employing a decision-tree classifier. Subsequently, they train a sequence labelling model for the event type labels.

Wanzare et al. [79] employ crowd-sourced texts on stereotypical situations for script structure inferral following a clustering approach. They construct paraphrase sets containing variations of event scenario descriptions, which are subsequently used for identifying event types. Their crowd-sourced alignments of variations of event descriptions are also the basis of a temporal script graph (TSG) which models the prototypical order of events in a script.

Scripts have recently been inferred using association rule mining (ARM) by Belyy and Durme [80]. In their work, they build upon previous approaches ([6, 76]). By applying ARM-based count statistics, they infer sets of interesting rules (i.e. event relations), while also successfully inferring missing events from sequences (*narrative cloze test*).[10]

The prediction of script events has been approached as a graph-based problem [82, 83], by incorporating information provided by knowledge graphs [84] (using a transformer model) or as a

[8] http://trec.nist.gov/data/reuters/reuters.html
[9] http://www.mythfolklore.net/andrewlang/
[10] A variant or extension of the narrative cloze test is the *story cloze test*, in which an appropriate ending for a four-sentence narrative needs to be chosen from a set of candidate endings. It can be applied as a measure for story understanding and script knowledge inferral [81].

problem of causal inference [85].

3.3 Event Extraction and Modelling

Many approaches exist which model event structure outside of scripts. In the following section, the approaches which are closest related to this work are presented.

Chambers and Jurafsky [6] interpreted scripts as narrative event chains. These event chains consists of narrative events associated with common actors. Multiple actors involved in a text describing the same activity evoke multiple individual event chains (e.g. *being a customer at a restaurant vs. being a waiter at a restaurant*). The structural representation of events can therefore be induced through the grammatical roles, e.g. agent (*orders(Character1,Food)* and patient *bringsInvoice(Character2,Character1)* in the events. Thus, their event representation includes verbs and characters (based on their semantic roles, similar to PAS). For the unsupervised inference of narrative events and narrative event chains, three subtasks need to be solved: event identification or induction, temporal ordering of events, and pruning the event space into event sets. Chambers and Jurafsky parse a text and identify verbal structures with shared arguments, from which they construct chains of events. They achieve a (partial) temporal ordering by employing *before*-relations, e.g. *sit-down*-before-*order*. They propose the *narrative cloze test* as a measure to determine how well an event prediction model infers missing script events from a given event chain. An example chain is given in Figure 3.4.

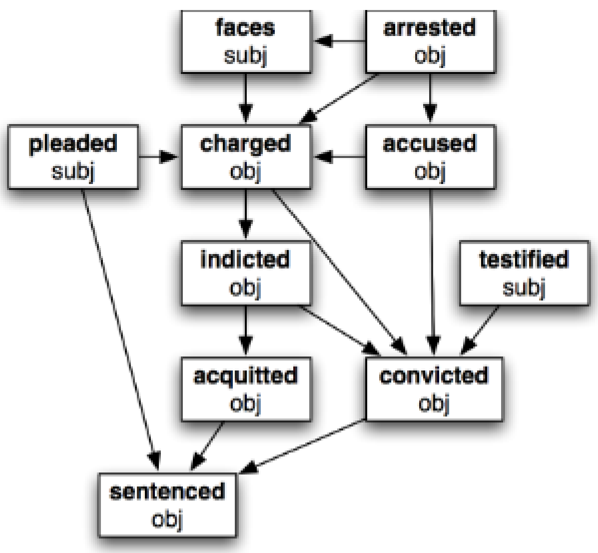


Figure 3.4: Example of an automatically extracted narrative chain (*Prosecution Chain*), Source: [6]

In a subsequent study, Chambers and Jurafksy [86] then proposed *narrative schemas*, as a set of typed narrative chains. This model includes all characters or actors involved in a narrative, not only the main protagonist. For that purpose, they utilize an unsupervised machine learning approach. They incorporate semantic roles into the argument structure of their event chains, subsequently merging multiple event chains into a coherent narrative schema.

A drawback of this approach is that they do not distinguish between events and non-events. For example *Mary loves her son Peter* and *Mary puts Peter to bed* are sentences in which verbs share common arguments (*Mary, Peter*), but the first sentence is a circumstantial statement describing the relationship between the participants, whereas the second sentence involves an event. Therefore, no temporal ordering can be assumed if the two sentences appear in a text.

Approaches towards event predictions that build upon Chambers and Jurafsky's prior work [6, 86] include using various configurations of neural networks [82, 87, 88, 89, 90] (and others), event embeddings [91, 92], event graphs [83], causal inference [93], association rule mining (ARM) [80], and knowlegde-base external information [84, 94].

Aldawsari and Finlayson [95] propose a method for supervised identification of events and sub-events. They distinguish explicit sub-event relationships, such as "attacked" and "wounded" and implicit relationships, such as "attacked" and "killed". They propose an elaborate set of features that combine lexical, semantic, and argument-related features with features modelling discourse and narrative structure. They consider different modes of narration of events and sub-events, e.g. the communication of an event-relationship in direct speech. In terms of narrative features, they theorize that information on the centrality of an entity mention is meaningful for detection event-sub-event relationship. However, they found co-reference in arguments that do not directly concern the events in question but are shared by the sentences describing the events ("non-major mentions"), also serve as useful features. Aldawsari and Finlayson observe that the syntactic features are the most important for the supervised model, followed by semantic features, discourse and narrative features, lexical features and arguments.

Glavaš et al. [96] propose a model of events and sub-events, including spatiotemporal information, for the annotation of events in the HiEve Corpus. They argue that the linear model of events as assumed by Chambers and Jurafsky [6, 86] is not fully suitable to capture relations between events, because events might have different types of temporal relations, their relation can be spatial and events have to be classified on a scale of granularity instead of assuming each event carries the same weight in a chain of events. For that purpose, their event model follows a hierarchical approach, with the possiblity that one event can be spatio-temporarily contained in another. They also annotate event co-references. In total, the annotated HiEve corpus contains 100 documents, and consists of 1354 sentences.

To understand the representation of event structure in Weblog stories, Manshadi et al. [77] have modelled events as predicate argument structures (PAS) (similar to the SPO-structure in *hylemes*, see Section 2.2). They developed a probabilistic model to represent event structure in largely ungrammatical texts from a large corpus of weblogs, which is then used for both event ordering and event prediction.

The task to align predicate argument structures (PAS) was introduced by Roth and Frank in 2012 [97]. For pairs of documents, an alignment aims to assign links indicating coreferent predicate arguments. As a baseline, Roth and Frank use predicate lemma overlap for PAS alignment. They base their approach on pairs of documents that are known for referring to the same event. However, their approach is not strictly focussed on event representation, as they make no distinction between coreferent predicate arguments that refer to non-events, e.g. commentaries. Therefore, their approach is more targeted at applications of machine translation, paraphrase detection or the study of textual entailment than plot modelling. As a gold standard, they manually annotated 70 text pairs from the GigaPairs corpus, preprocessed with a semantic parser to identify predicates. They perform predicate argument alignment using a graph-based clustering approach, where predicates are graph nodes and similarities are weighted graph edges. The resulting graph is subsequently clustered using a min-cut approach, and alignments are assumed for events that belong to the same cluster.

Another work that approaches event modelling through PAS was presented by Wolfe et al. [98] They approach the task of cross-document alignment of PAS through a pipelined system, PARMA (predicate argument-Aligner). With PARMA, Wolfe et al. provide a framework for predicate argument alignment that is based on a logistic regression model using lexical-semantic features, such as WordNet path distance and Tree Edit Distance of two candidate sentence dependency trees. They evaluate their alignment on the Extended Event Coreference Bank, Roth and Frank's predicate argument alignment corpus [97], and their own Multiple Translation Corpora. They report improved results compared to previous related works across all three corpora. In a later work, Wolfe et al. [99] improved the PARMA framework by including joint factors: fertility, i.e. the number of links to a single item, which may indicate an error, the predicate-centric and entity-centric structures of a predicate and its associated arguments, adding constraints and penalties to overlapping predicates and shared arguments that do not share a single predicate. They also included temporal ordering as a feature to penalize alignment of predicates which logically cannot be aligned, because their temporal order does not match.

A sub-task of event modelling is event co-referencing. Identifying if two utterances refer to the same event, a related event or an unrelated event is not trivial, as shown for instance by Hovy et al. [100] They define events as discourse elements (DEs) in a text, which may refer to the same event, have a partial co-reference (e.g. one battle among multiple battles) or are unrelated. Furthermore, the same event can be reported inconsistently, i.e. include additionally or contradictory information when reported from a different point-of-view. They introduce the notion of quasi-identity, where two mentions of events can be replaced with each other, because their relation is either membership (part-of) or sub-event, i.e. the event is part of a script [73]. Following these considerations, they annotated two corpora (Intelligence Community, IC and Biography Corpus) [100]. Notably, their annotated data contained on average 19.5 full coreferences of events per article, but only 7.2 sub-event relations and 2.7 member relations.

Events are also often the basis of personal stories, that are communicated by persons in various forms. Event extraction from transcripts spoken dialogues have been performed by Eisenberg and Sheriff [101]. Their event definition includes both actions and states of animate objects.

Event detection with a special focus of applying new event types has been pursued by Lai et al. [102]. They apply a few-shot learning approach on few examples of previously unseen event types. Their event schema also includes non-events (*NULL*-events), which they use as a filter before classifying events into finer-grained classes, such as *Business*, *Conflict*, *Movement*, or *Transaction*. They use different encoder models, *CNNs*, *LSTMs*, and *GCNs*.

Event extraction has also been explored on the basis of frame semantics [103]. Under the assumption that the tasks of frame-semantic parsing and event detection are structurally similar, but utilize different feature types, they follow a transfer-learning approach, and retrain the frame-semantic parser SEMAFOR for event extraction. Their adaptation of the model also considers an important difference between the tasks: the possibility to observe non-events. They observe that the inclusion of a non-event class biases the model towards that class. Which is an acceptable bias for the task, since non-events are more prevalent than events in real-life data. Other event classification efforts have shown similar results when considering non-events, e.g. [104] (see below).

3.3.1 Temporal Semantic and Event Annotations

Vauth et al. [104, 105], respectively Gius and Vauth [106] propose a new event model for the annotation of events in literary texts. In many NLP projects, event definitions, and the extraction of events, is often treated as a merely verb-centric task without much consideration what an event constitutes and what does not.[11] In contrast, their annotation schema provides a narratologically-grounded event model that is suitable for the NLP community, and distinguishes between different types of events, building upon prior narratological categories.[12] In their definition, they operationalize the concepts of narrativity and tellability as important discourse phenomena. They annotate verbal phrases using four categories *change of state*, *process event*, *stative event* and *non-event* (on a German data set). Using those annotations, they measure narrativity across different texts, e.g. Kafka's *Die Verwandlung*. Their automatic event classification tool, performs well except when distinguishing between *change of state* and *process* classes.

In the NLP community, the markup language TimeML [109, 110] provides widely a accepted schema for the annotation of temporal expressions and events in textual data, including temporal

[11] This is not the first criticism that discusses a lack of precision in the definition of the *event* concept in NLP. Along the same train of thought, the concept of *narremes* was refined by Baikadi and Cardona-Rivera [107].

[12] The difference between Gius' and Vauth's event model and the hyleme type annotations have been discussed in detail by the author in [108].

relations between events. An evaluation of four different automatic TimeML annotation systems can be found in [111].

A corpus of modernist novels and hypertext fiction (in English) has been annotated with temporal narrative labels by Kearns [112]. They annotate narrative phenomena that introduce anachrony to a narrative, such as ana- and prolepses, or changes in narrative level (and their degree). They also include subjective narration in the forms of stream of consciousness and free indirect discourse.

3.4 Narrative Sequences and Structures

A multi-features approach towards generating semantic sequences from texts is followed by Peng et al. [113] They include multiple features in their model, namely frame embeddings, named entities and sentiment. On these features, they train a neural model. They evaluate the model on a basic discourse parsing data set, and on the story cloze test and report good overall results. Another approach towards models of narrative sequences is based on identifying tropes [114] (in computer games). Deriving narrative structures (or rather textual representations thereof) can be seen as a sub-task in many other applications, such as screenplay summarization [115].

3.5 NLP approaches to Mythical and Folkloristic Content

A project closely related to this work, both in terms of data from the cultural heritage domain and the objective of comparison and alignment, has been undertaken by Reiter et al. [116, 117]. Reiter collected a corpurs of 38 folktales, selected based in their ATU[13] types, their length and straightforward narrative. Additionally, they included a corpus of 46 ritual descriptions as forms of narratives that consist of events and event participants. The ritual descriptions stem from the Nepalian Hindu cultural background. They use alignment algorithms, Needleman-Wunsch algorithm, Graph-based Clustering [97], and Bayesian model merging, to detect structural similarities in their data sets. In contrast to the other two alignment methods, Bayesian model merging can take multiple events into consideration. They use different similarity measures to identify potential candidates for alignment: FrameNet Similarity, WordNet Similarity, VerbNet Similarity, a distance similarity based on relative position of the event in the event chain, and a bag of lemmata approach called argument text similarity. They provide expert annotations for the alignment ritual descriptions.

Aldawsari et al. introduced the term *story fragment stitching* for the task to automatically align and construct story parts from different sources to form a comprehensive narrative [118]. To demonstrate this task, they reconstructed the story of Moses (as), from seven story fragments in the Holy Quran. They manually annotated 708 event mentions from ayats (verses) in different surahs (chapters) in Arabic and English, according to event categories based on a previous event analysis.

[13] Aarne-Thompson-Uther

They represent the event ordering in a directed graph approach and use weights derived from a transformer model (BERT) and TF-IDF weighting. They identify 43 events in the storyline, which appear in one or multiple ayats.[14] They approach outperforms their baseline (Needleman-Wunsch alignment) in terms of precision and F1-score, but yields a lower recall.

An annotated corpus of Aesop's fables in English and Chinese is presented by [119]. They provide labels for structural analysis, such as the setting, internal and external events and states, speech and moral. Additionally, they include relational labels, which describe the interplay between adjacent text spans, such as temporal connections, causal relations, or comparisons.

TrollFinder [120] is a corpus-based exploration tool for geolocations in Danish Folktales. Based on a digitized corpus of more than 30.000 folkloristic narratives, they derived places of provenience and mentions of geolocations following a gazetteer approach with topographic references. A simple keyword extraction based on word frequency and lemmatization is used for assignment of topics. Based on these meta data, they can gain interesting insights, for instance that tales involving witches are concetrated in certain areas.

With regard to African folktales Ninan and Odejobi [121] present a corpus of Yorùbá folktales with their English translations. Moreover, they also include structural information, information of characters, and the setting of the story (*prop*). For that purpose they propose and apply a special markup language, AFT (African Folk Tale) Markup. Both versions of the tales were annotated using this XML extension, making it a suitable resource not only for structural analysis, but also for applications like machine translation.

Finlayson [122, 123] follows a machine learning approach to automatically infer Proppian functions from text containing (manual and semi-automatically created) semantic annotations. They report high results for functions pertaining to *Villainy/Lack*, *Struggle/Victory*, and *Reward*.

Karsdorp and van den Bosch [124] study story networks, a model in which variants (retellings) of stories are connected (form a partial *Stoff* if you will). From the example of Dutch variants of the cautionary tale *Little Red Riding Hood* (KHM 26), and a corpus of chain letters, they conclude that newer variants are closer related to retellings than older versions.

Some works engage with folkloristic content through motif or tale type indices. Yarlott and Finlayson [125] critize existing motif indices as inconsistent and incomplete, and the notion of *motif* as ill-defined for the applicability in natural language processing and computation linguistics. They operationalize the definition of *motif* as "a set of closely-related variants of a non-commonplace, specific narrative element that is repeated across tales of the same type."[125, p.5][15] They attempt to automatically extract motifs from folktale texts, without limiting themselves to TMI motifs. They

[14] The resulting series of events can be transformed into a hyleme series with minor adjustments (e.g. changing past tense statements into present tense and changing the structure of individual statements without changing the content).

[15] The motivation behind this definition resembles the investigations in Chapter 6.2.3 where different textual representations of hyper-hylemes across variants of the same myth are compared.

do not report results.[16]

The MOMFER project is a search enginge for the Thompson Motif Index (TMI) [126]. It allows semantic queries of the index by using motif titles, descriptions, and WordNet categories. They present three case studies on how to employ the search enginge: the representation of monsters in folktales of different geographical origin, a case study on the representation of color in the TMI, and the representation of gender in motifs.

There are a number of approaches towards structural annotations, based on XML, for Proppian annotations [127, 128, 129, 130, 131, 132]. Malec [133] subsequently used annotations in the Proppian fairy tale Markup Language (PftMl) to automatically annotate and classify Russian magic tales.

Many prior studies examine *characters in stories*, we focus here only on the most relevant works that identify characters in folkloristic and mythological narratives. Valls-Vargas et al. [134] identify Proppian characters from unannotated texts. For that purpose, they use action matrices which encode Propp's restrictions on actions that are performed by a archetypal character, e.g. the *Donor*. The action matrices consist of verbs referring to actions which are performed between two Proppian character roles (rows/columns), e.g. *Donor gives Hero*. From a reference role action matrix for each character role, they then infer the most suitable role for a given character in a text.

The contribution by Jahan et al. [135] stands out for their theoretical-groundedness. They consider many facets that are relevant for annotating characters across domains (ProppLearner corpus [136] for folktales, OntoNotes for varying domains including news texts, and The Corpus of English novels). Their considerations include how to approach anthropomorphic characters, such as speaking rivers, and generalized referential constructs, such as "all Americans". Their character definition includes animacy and relevance to the plot and pertains to both main and minor characters. Jahan et al. annotated 170 texts across domains, and identified in total 1,347 character chains from all co-reference chains. They automate the character identification by including a animacy detection system from their prior studies [137] and intricate feature modelling that encompasses semantic and syntactic dimensions. Their system performance is notably proficient for the task, with a overall weighted F1-score of 90 %.

Gianitsos et al. [138] propose a method for classifying Ancient Greek texts into prose or verse. To that end, they propose a special feature set explicitly tailored for Ancient Greek. The feature set includes mainly grammatical and syntactic features. Their model achieves an impressive 98.9 % F1-score for the classification task.

This work is also marginally related to the discipline of Computational Assyriology. This often refers mainly to the study of primary sources, e.g. the development and application of NLP tools, but also digitization or reconstruction of artifacts. Efforts concerning Computational Assyriology

[16] To our knowledge, the work towards automated motif extraction is still underway. No peer-reviewed results have been published as of date February 13, 2024.

across all fields of application have existed for multiple decades. For a comprehensive overview I would like to refer the reader to the work of Aleksi Sahala [139, p.31-72].

3.5.1 Automated Storytelling

The performance of computational tools for story generation, a sub-task of natural language generation, and computational creativity has often been applied as one of the measures for capabilities and limitations of artificial intelligence, from practical, theoretical and philosophical standpoints [140].[17] There are many approaches to automated storytelling with a focus on folkloristics. Peinado [142] used a manually crafted ontology of Propp's 31 invariant functions as a framework, while Gervas used Proppian functions as story grammar [143].

3.6 Semantic Similarity and Relatedness

Semantic *similarity* is a form of relation between two or more concepts[18], judging their distance in terms of meaning. For instance, the concepts *road* and *street* are semantically similar. Semantic relatedness on the other hand refers to a pragmatic relationship – or a cognitive association – between two concepts that might have a different meaning, e.g. *street* and *car*. Semantic *proximity* is occasionally used when both similarity and relatedness are concerned. Semantically similar terms, e.g. *street* and *road*, are also semantically related, but the opposite is not necessarily true. There is a vast number of approaches towards semantic similarity, and a substantial amount of work towards semantic relatedness. Semantic similarity annotations, synonymy detection, development of robust semantic similarity measures, or automatic identification of semantically similar concepts in text are some of the core endeavors undertaken by the NLP and Computational Linguistics (CL) community.[19] Therefore, in this section only the most common approaches are presented along with selected research efforts.

Firstly, I discuss annotation studies (i.e. human judgements) on semantic similarity/relatedness. The first data set containing judgments on the semantic similarity of (English) word pairs was presented by Rubenstein and Goodenough [37]. They asked 51 participants to judge the semantic similarity ("degree of synonymy") of 65 common English noun pairs on a scale from 0.0-4.0, with 4.0 absolute synonymy. They used the resulting data set to prove that words that are similar in meaning appear in similar contexts, and infer that overlap of context words implies similarity of a concept pair. In that, they proved the applicability of context based similarity measures used for information retrieval. They also showed that the similarity of context is a strong indicator for concepts which are close to synonymous, like *gem-jewel*, but less so for medium or low similarity.

[17] For a brief overview of the history of automated storytelling, please refer to [140], and the more recent survey [141].

[18] In the context of semantic similarity/relatedness, I use the term *concept* instead of *word*, because words can refer to multiple concepts (homonymy), e.g. *bank* (financial institute/river bank).

[19] A simple search for the term "semantic similarity" in the ACL anthology yields 9480 results (semantic relatedness: 4580, semantic proximity: 172) [February 13, 2024].

The second data set concerning semantic similarity was developed by Miller and Charles [38] two and a half decades later.[20] They favour the expression *semantic similarity* as a continuous variable over *degrees of synonymy*, because synonymy is subjective and the meaning of a sentence can change even if a word is replaced by a synonym. Similarity judgements rely on context, which can be defined pragmatically as "words around a target word" (collocations), or structurally. The structural considerations for context include semantic (e.g. semantic roles and selectional preferences[21]) , syntactic (e.g. which types of object a word assumes), pragmatic (e.g. the relation between a question and a subsequent answer), and stylistic (e.g. technical or colloquial terms) conditions. All these considerations have to be taken into account when semantic similarity of concepts is determined by means of a contextual representation. Miller and Charles [38] first reproduced Rubenstein and Goodenough's (R&G) experiment [37], using 30 noun pairs (10 each originally rated high similarity, medium and low similarity) and 38 participants. Then they performed further experiments on six noun pairs (two for each level of synonymy in R&G), where 24 new participants where prompted to sort example sentences with gaps into two groups for each noun pair, each group containing sentences that take one of the nouns for the gap position. Lastly, in a third experiment, subjects were asked to write sentences for each word in a word pair. Those sentence pairs were then used in the same manner as in the second experiment. They essentially interpret semantic similarity between two nouns $n1$ and $n2$ as a function of context, in that a similar concept $n1$ is more often mistakenly placed in the gap position of the sentence that was originally filled by $n2$.

They came to a few fundamental conclusions: animate object pairs seem to have privileges over other pairs, because relatively many contextual variations allow a substitution, e.g. "___ live a life in poverty" (slave, monk) [sic]. They raise the possibility that the "relation between semantic similarity and contextual similarity varies between semantic fields." [38, p.22] Secondly, antonymy relationship (*good/evil*) is a special case, because the two concepts are fundamentally different, but the contexts are very similar. However, they argue that the strong association between antonyms is not due to interchangeability, but through association by contiguity and that one antonym is usually more plausible than the other (e.g. "The food was ___. I will never go to that restaurant again." (*great/horrible*)). Both of these works, [37] and [38], are remarkable in that they laid the foundation for what modern studies understand as distributional semantics, although Miller and Charles argued for a more intricate feature set than most frequently used.

Since the early 2000s, different and larger data sets were constructed, using different scaling systems, e.g. 0-5, or 0-10. Through the spread of internet access to homes and institutions, researchers were able to recruit more test subjects (crowdsourcing workers), allowing the creation of bigger and more comprehensive data sets. At the same time, it became easier to evaluate the results of human judgments through the use of computational tools. The largest data set (10,000

[20] It is conceivable that during the Cold War era related works on data sets existed in the former Eastern countries that the community is not aware of.

[21] e.g. "Molly eats X" implies that the direct object X (patient) is something edible, like "a can of tuna".

items) for the English language is the *Sentences Involving Compositional Knowledge* (SICK) data set [144]. In a crowdsourcing study, the workers were asked to annotate pairs of concepts according to their semantic relatedness on a 5-point scale. Additionally, the authors asked workers to provide semantic relations between the concepts, such as *entailment*, *contradiction*, and *neutral relation*. This data set aims to provide a benchmark for the evaluation of semantic relatedness methods.

One of the semantic relatedness data sets for the German language is presented by Gurevych [145]. It contains 65 concept pairs (nouns), rated by human subjects (native speakers of German) on a scale from 0 to 4. They then apply a information content (IC) metric to compute semantic similarity on the frequency of word stems, mainly to account for compound words in the German language. They however raise the concern that a proper word sense disambiguation (WSD) might be needed to fully establish robust semantic relatedness measures. In general, they confirm that IC-based metrics are applicable to morphologically complex languages, like German.

The availability of semantic similarity data sets varies from language to language, which is especially problematic for under-resourced (UR) and low-resource (LR) languages. While some concept pairs might be applicable cross-culturally, others might be influenced by cultural knowledge and circumstances, e.g. in some cultures the term pair *Sunday/prayer* might be closer related than *Friday/prayer* and vice versa. For a exhaustive survey on semantic/relatedness data sets, measures and methods, I recommend the works of Chandrasekarana and Mago [146].

A multi-lingual data set for semantic similarity is Multi-SimLex [147]. It contains 1888 similarity ratings for concept pairs in 12 high- and low-resource languages (e.g. Welsh and Kiswahili). They pair concepts represented by words from different word classes, such as noun-adjective pairings. For their manual annotation, they reach a satisfying inter-annotator agreement across all languages ($0.667 \leq \rho \geq 0.812$). Furthermore, they present a construction pipeline for similar projects in other languages, and elevate the construction of data sets by multi-lingual concept alignment, resulting in 66 additional cross-lingual evaluation data sets.

Kiritchenko and Mohammad [148] reflected on the construction of data sets for semantic similarity, in that they investigated different rating schemes. Through experiments, they came to the conclusion that ranking schemes which are commonly applied when asking human workers for judgments, are inferior to a best-worst rating approach. Based on this conclusion, Asaadi et al. [149] produced the BigBiRD data set, which can be used as a benchmark for bi-gram semantic relatedness. For a survey of methods pertaining to the measurement and prediction of semantic relatedness, please refer to [150].

Applications and approaches towards semantic similarity and relatedness can be roughly grouped into four categories: knowledge-based efforts [146, 151, 152], corpus-based approaches [146], vector-based [146, 151] and deep learning methods [153].

Semantic similarity can also be extended towards the similarity of sentences (in contrast to concept or n-gram pairs), and further towards document similarity. Examples of data sets concerning sentence similarity are the SemEval data sets (2014, Task 1) [154] and (2017, Task 1) [155]. A data set for document similarity ("text snippets") was used in SemEval-2014 (2014, Task 10) [156].

3.7 Narrative Similarity and Story Similarity

Chaturvedi et al. [157] investigate story similarity based on a corpus of 577 movie remakes. For that purpose they take not only plot units into consideration, but also find similarities in character representations and their (social) relationships. They employ the notion of a *story kernel*, and a *character alignment kernel*. The *story kernel* consists of a bag-of-words (BOW) representation of verbs in the narrative. The *character kernel* consists of a 1-1 alignment of characters in two narratives with each other based on similarity. Character attributes are derived from adjectives and actions of the character (predicate agent and patient relationship).

Elson [158] has proposed Story Intention Graphs (SIG) to model narratives. A SIG model is a connected graph of entities and relations, which include both actions and agency, i.e. the intentions behind and goals of an action. For that purpose, they annotated a corpus of 70 SIG encodings, called Drama-Bank. As an annotation unit, they chose sentences or clauses. Temporal order of the narrative is modelled via *followedBy* relations. From the SIG encodings, they derive general patterns, similar to general plot configurations in Figure 3.3. Patterns can indicate tropes (or motifs). If one of those patterns can be detected in two candidate SIG encodings, the underlying stories share an analogy.

Saldias and Roy [159] attempted to match personal stories, using annotations of clauses with Labov's model of personal narratives. The annotations include labels for action clauses, orientation clauses, and evaluation clauses.

Human judgements of story similarity were investigated by Fisseni and Löwe [160]. They wanted to investigate which factors of narratives humans take into consideration when comparing stories, and concluding narratives to be equivalent. For that purpose, they performed three experiments in a classroom settings with German literature and language students. The students were presented with two short variants of stories that show dissimilarity with regard to one controlled dimension, e.g. change in narrative order or style (offensive language). In one experiment, they asked subjects to rate the similarity/difference of versions of the tale *The three feathers* [54] (KHM 63), which is presented in different variations across the different Grimm editions. The variants differ in narrative order, granularity, and with two possible endings. They come to the conclusion that untrained expert annotators do not favour structural information over other dimensions when assessing similarity of narratives. When asked to provide multiple dimension on which the similarity between two concrete stories is judged, the structural similarity becomes less important to annotators (self judgement).

In a crowdsourcing task, Nguyen et al. [161] derive human judgements of narrative similarity. For that purpose they ask domain experts and layman to judge the perceived similarity of texts from a Dutch folktale collection. Their results indicate that experts and non-experts judge story similarity by featuring in different aspects. They construct the annotation task from narratives of different story types and genres, in varying pairings (e.g. same story type but different genres or vice versa). Annotators are then asked to provide a rating on how similar they perceive the narratives to be, ranging from 1 (no similarity) to 5 (the same or almost the same). Hence, they do not treat story similarity as a binary classification (same/not the same or similar/different). They define dimension of narrative similarity, such as characters, plot, genre, theme or style. Their results indicate that plot, genre and themes are the most important aspects that crowd worker annotators take into consideration (self-declared). For experts, story types and motifs play a higher role in judging similarity, and secondarily they factor in plot, characters and themes.

3.8 Applications of MSA Algorithms to Natural Language

Multiple sequence alignment (MSA) algorithms, such as the Smith-Waterman algorithm, or the Needleman-Wunsch algorithm (see Section 2.3), have been used to process and investigate natural language in various forms. Besides alignment of scripts by Regneri et al. [74], which has already been discussed above, Fay [162] uses a sequences of temporarily ordered events on which he performs MSA using the Needleman-Wunsch approach. Fay approaches the problem of comparability of story events through the associated subject and objects of each event statement. He constructs a match graph representation of two candidate stories, pairing objects from stories A and B together in various configuration. He then uses MSA to determine the best alignment.

Many other applications of Natural Language Processing employ multiple sequence alignment techniques, e.g. in sentence-level paraphrase detection [163], constructing dictionaries of semantic expressions and their possible natural language realisations [164], for ordering of prenomial modifiers to facilitate natural language generation [165], text summarization [166], to detect similarities in pronouncation habits [167], or the identification of discourse relations [168].

Chapter 4

Exploratory Study of Hyleme Data

Figure 4.1: Word Cloud of 200 most frequent terms in the German hyleme data

This chapter introduces two hylistic data sets (in German and English), and compares them with two other data sets, which contain narratives, but do not follow hyleme structure.[1] The first comparison set is the 1857 version of Grimm's *Kinder- and Hausmärchen* [54] (KHM). It was chosen because it consists largely of stories with a straightforward narrative. The language of the data is antiquated and largely unstandardized German. It contains 2986 sentences.

The second data set used for comparison is a sample of German Wikipedia articles, listed in Table 4.1.[2] It contains 22 Wikipedia page texts from various topics, such as people, history, geographic locations and politics. The texts are in standard German. The Wikipedia data set and the hyleme data set are similar in size (6327 hylemes vs. 6380 sentences).

This chapter has three purposes: Firstly and most importantly, the hyleme data sets and their properties, e.g. the most common verbs/hyleme predicates, are introduced. Regarding those properties, they are compared to other data sets. Secondly, the performance of popular Natural Language Processing (NLP) tools is compared across the data sets. It is determined how well

[1]For better readability, the terms *hyleme* and *hylistics* are are left unitalicized in this and the following chapters.
[2]The sample has been collected from links of the main Wikipedia landing page on January, 18th, 2023.

the tools perform for the purpose of automatic processing hyleme data, standardized and non-standardized language. Lastly, the German hyleme data is annotated with hyleme types, i.e. *durative-constant*, *durative-initial*, *durative-resultative*, and *single-event* (see Section 2.1). Based on the gold standard annotation, a hyleme type classifier is trained for both data sets.

Table 4.1: Wikipedia Data Set

Page	Sentences
Ready Teddy	162
Belarus	907
Swjatlana Zichanouskaja	59
Matteo Messina Denaro	108
Cosa Nostra	1015
Sizilien	629
Boris Pistorius	198
Bundesministerium der Verteidigung	234
Tamar (Georgien)	45
Georgische Bagratiden	70
Königreich Georgien	287
James Cook	364
Hawaii	621
Anthony Giddens	259
Strukturationstheorie	139
Warschauer Ghetto	327
Orkantief Friederike	152
Gefängniskirche Tegel	55
Sanaz (Sängerin)	46
Lucile Randon	70
Mousse Boulanger	51
Deutsche Besetzung Polens 1939–1945	582

4.1 German and English Hyleme Data

The German hyleme data set contains 6315 hylemes in 228 sequences from the mythological domain. The sequences describe the narrative of mythical stories and rituals from a diverse set of geographical and temporal backgrounds. The sequences were extracted mainly concerning myths which relate to one of three fields: Classics, Ancient Near Eastern Studies, and Religious Studies (Bible and other religious texts). Figure 4.3 illustrates the distribution of the hyleme sequence according to their backgrounds. The German data set was extracted by domain experts in the fields of Classics CS, Ancient Near Eastern Studies ANES, and Religious studies over a three year period as part of the DFG-funded Myth Research Group 2064 STRATA. For easier comparison of the hyleme sequences, the German hyleme data set is presented in a “hyleme database”, which allows domain experts to enter hylemes and hyleme sequences with their properties (such as source and myth title) and has different fields for hyleme components, i.e. hyleme subject, predicate, object and determinations.[3]

The English hyleme data set contains hylemes extracted from Henry Callaway's *Nursery Tales and Histories of the Zulus* [7].[4] It contains 5176 hylemes in 384 sequences. While the German data set contains many variants of profoundly different myths and a few examples of variants of the same myth, the English data set contains folktale variants of related material. Each sequence in the English data set describes a part of a folktale narrative. Since most of the myths from the German data set are from different temporal, geographical and cultural contexts, the German data set can be used for comparison of variants of the same myth, but yields only little room for exploratory myth comparison and alignment. The English data set on the other hand contains practically no variants of the same folktale, but the different tales share enough similarity to allow for interesting exploratory hyleme comparison and alignment. The tales were collected by Callaway over a limited time frame (approx. first half of the 19th century), from a limited set of narrators, while the time span of the German hyleme data ranges from the beginning of written history (approx. 3100 BCE), to 875–1075 CE. The compilation of the English data set was undertaken by the author as the first approach towards the hylistic analysis of folklore material. While the German hyleme data set contains mostly mythological topics and narratives, folktales follow a different structure. However, the two domains share narrative similarities, such as descriptions of ritual practises or supernatural phenomena. In that, the results of this section and the following analyses in the later chapters of this thesis are comparable. For the study of hyleme structures, the data set is interesting because of similarities between different tales. Furthermore, the specific domain of folktales of the amaZulu is so far understudied by the computational folkloristics community. The data set is therefore an excellent tool for further studies on plot analysis as well as the linguistic analysis of historical/vernacular isiZulu.

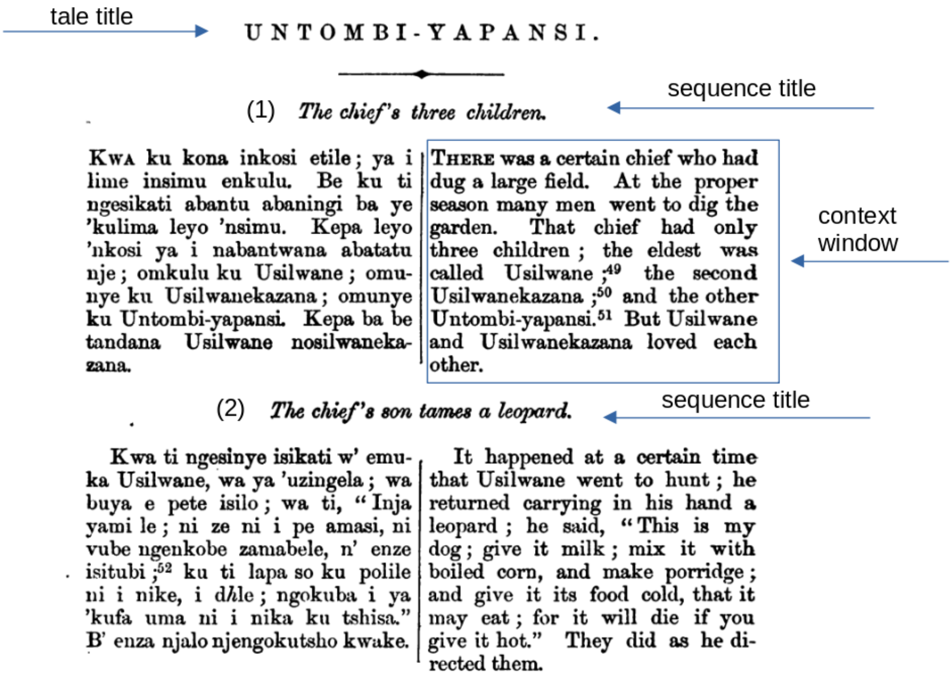
The English hyleme data is based on 30 tales from Callaway's collection. A few sequences include folkloristic background information, and less narrative, e.g. the description of the honey bird. In cases where these descriptions are important for understanding aspects of the tales, they were converted into hyleme structure, containing mainly *durative-constant* hylemes. The folktale collection includes English transcriptions of the Zulu vernacular language.[5]

The hyleme extraction was performed on the English text. The Zulu text was consulted, where the English text needed further specification or when in doubt about translations. Since some tales are relatively long (> 700 hylemes), it was decided that the context window for one sequence should be set to the paragraphs between Callaway's subheadings (see Figure 4.2), where available. The subheadings often contain descriptions, or a one-sentence-summary, of the content of the following part of the tale. In hylistic terms, some of these subheadings can be interpreted as

[3]The database was developed by the Lower Saxony State and University library. The software initially exhibited several unexpected behaviours that required resolution. It was subsequently refined and extended under guidance of the author.

[4]The original text (introduction and commentaries) may contain terms and phrases which might be disturbing to some users. The derived hyleme data set does not contain any harmful terms.

[5]A digitised version of the folktale collection is available in the Five Hundred Year Archive: `https://fhya.org/AdditionalResources/file/id/250151?subquery=Callaway`



UNTOMBI-YAPANSI.

(1) *The chief's three children.*

KWA ku kona inkosi etile; ya i lime insimu enkulu. Be ku ti ngesikati abantu abaningi ba ye 'kulima leyo 'nsimu. Kepa leyo 'nkosi ya i nabantwana abatatu nje; omkulu ku Usilwane; omunye ku Usilwanekazana; omunye ku Untombi-yapansi. Kepa ba be tandana Usilwane nosilwanekazana.

THERE was a certain chief who had dug a large field. At the proper season many men went to dig the garden. That chief had only three children; the eldest was called Usilwane;[49] the second Usilwanekazana;[50] and the other Untombi-yapansi.[51] But Usilwane and Usilwanekazana loved each other.

(2) *The chief's son tames a leopard.*

Kwa ti ngesinye isikati w' emuka Usilwane, wa ya 'uzingela; wa buya e pete isilo; wa ti, "Inja yami le; ni ze ni i pe amasi, ni vube ngenkobe zamabele, n' enze isitubi;[52] ku ti lapa so ku polile ni i nike, i dhle; ngokuba i ya 'kufa uma ni i nika ku tshisa." B' enza njalo njengokutsho kwake.

It happened at a certain time that Usilwane went to hunt; he returned carrying in his hand a leopard; he said, "This is my dog; give it milk; mix it with boiled corn, and make porridge; and give it its food cold, that it may eat; for it will die if you give it hot." They did as he directed them.

Figure 4.2: Excerpt from Henry Callaway's Nursery Tales and Histories of the Zulus [7]

hyper-hylemes. Heading (1) in Figure 4.2 is not a hyper-hyleme since it does not follow the hyleme subject-predicate-object structure, as it does not contain a verb. Heading (2) would be suitable as a form of hyper-hyleme, because it follows the hyleme structure and contains a brief description of the content of the following section without only alluding to it (in contrast to, e.g. "Uthlakanyana practises hypocrisy [...]"). The hyleme sequences pertaining to the subheadings can be combined to create longer sequences for entire tales.

Table 4.2 summarizes the four data sets and their particularities.

Table 4.2: data sets used in this chapter

Data set	Language	Standardized	Domain	Hyleme Structure
Wikipedia	German	yes	Encyclopedia	no
Grimm	German	no	Folktales	no
Hyleme (German)	German	yes	Myths	yes
Hyleme (English)	English	yes	Folktales	yes

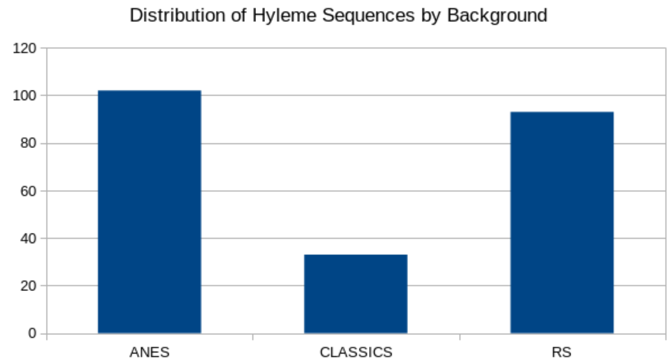


Figure 4.3: Distribution of German hyleme sequences by topic, Ancient Near Eastern Studies (ANES) 102, Classics 33, Religious Studies (RS) 93

4.2 Hyleme Predicates

For the purpose of modelling, comparing, and subsequently aligning hylemes, a robust method for the detection, lemmatization and comparison of predicates is needed. Therefore, the following section discusses the distribution of verbs in hylemes and the performance of the spaCy verb lemmatizer for the hyleme data sets in comparison with the two comparison sets (Grimm/Wikipedia).

Table 4.3 shows the distribution of verbs in the hyleme data sets and the two comparison data sets. For each German hyleme, the hyleme predicate was exported from the hyleme database. For each predicate, the occurrence in the hyleme data was counted. Subsequently, the list was manually corrected. Where a predicate was misspelled, the occurrences were counted to the right predicate. Multi-word constructions were corrected, where a predicate incorrectly contained a predicate and a predicate determination, often a subject complement, e.g. "stark sein" (engl. "to be strong"). For the English data set, the hylemes were parsed and the hyleme predicate was extracted using the spaCy[6] lemmatizer on the token with the dependecy tag ROOT.

Both comparison data sets (Grimm and Wikipedia) have been processed using spaCy. The corresponding verb lemma lists were constructed by filtering tokens according to their parts-of-speech tag (*VERB* or *AUX*) and subsequently lemmatizing them, using the spaCy lemmatizer (*token.lemma_*). In a second step, the resulting frequency lists were manually corrected. Non-German words and words falsely labelled as verbs (e.g. interjections) were removed. Incorrect lemmata were corrected and added to the right lemma. Two examples for common mistakes of the lemmatizer are given in Table 4.4. Since the spelling in the Grimm corpus is largely unstandardized, the number of incorrect lemmata in the Grimm corpus was higher than in the Wikipedia sample. For example, the lemma *gehen* was labelled correctly 98 times, in 133 instances it was given incorrectly. Due to the antiquated language and unstandardized spelling in the Grimm corpus,

[6] version 3.5.3 `https://spacy.io/`

the lemmatizer struggled with the past tense of certain verbs, such as “gieng” as the simple past of “gehen” (engl. “to go/walk”). As a unit of analysis, we use different segments for the data sets. The hyleme data sets are analyzed on the basis of individual hylemes, the comparison sets are analyzed on a sentence-level. The number of total verb occurrences and distinct verbs after manual correction are reported in Table 4.3. According to definition, hylemes have exactly one hyleme predicate (finite verb). Therefore, for the hyleme data sets, the number of segments and verbs is the same. The Grimm data set is about half the size of the German hyleme data set and the Wikipedia data set. Additionally, it has an average of 3.6 verbs per segment which is more than the other two data sets (1.0 in hyleme data sets, 1.35 in Wikipedia). This is due to the period and style of the genre of the Grimm corpus, as well as the tendency of the German language to combine multiple main clauses with conjunctions. The following example from the fairy tale “The Shroud” (KHM 109) [54] illustrates this:

> “Es hatte eine Mutter ein Büblein von sieben Jahren, das war so schön und lieblich, daß es niemand ansehen konnte ohne ihm gut zu sein, und sie hatte es auch lieber als alles auf der Welt.”

Translation:

> “There was once a mother who had a little seven-year-old boy. He was so handsome and lovable that no one could look at him without liking him, and she loved him above everything else in the world.” [169]

This sentence contains at least four hylemes: *A mother has a [son]. The boy is seven-years old. The boy is handsome and lovable. Everybody likes the boy.*[7] *The mother loves the boy more than anything in the world.*

Table 4.3: Verb statistics after manual corrections

	Hyleme German	Hyleme English	Grimm	Wikipedia
Total segments	6315	5176	2986	6380
Total verbs	6327	5176	10883	8644
Distinct verbs	1174	443	1142	1396

From the four data sets, the Wikipedia set has the highest number of distinct verbs (1396). The German hyleme data set and the Grimm data set have a similar number of distinct verbs (1174 resp. 1142). The English hyleme data has only 448 distinct verbs. This is due to a variety of reasons: All hylemes in the English data set were extracted by the same domain expert. Therefore, language particularities of the extractor, e.g. favoring one expression over a synonym, are more prevalent. Additionally, the data set was constructed based on narratives from one domain. The folktales have a limited set of actions compared to the other domains. Furthermore, the English hylemes were extracted with the purpose of hyleme alignment in mind. Without formally using a controlled

[7] This hyleme might be extracted as *“Nobody can look at the boy without liking him.”* depending on the preference of the hyleme annotator to stay close to the original source or paraphrase for simplicity.

vocabulary, the extractor tried to express the same or similar actions using the same predicates, e.g. "tell someone to do something". The prevalence of communication actions in the tales, as illustrated in Sequence (2) in Figure 4.2, also influences the verb variety. Lastly, since the hylemes were extracted from the English translation of the vernacular Zulu, the tendency of simplifying the original action-bearing predicate in the translation might also limit the number of distinct verbs in the resulting hyleme sequences.

From the 5176 hyleme predicates in the English data set, the spaCy model correctly identified 5153, resulting in 99.56 % correctly predicted lemmata. In 17 of the 23 mispredictions, the spaCy model predicted an out-of-vocabulary word (proper name) as the hyleme predicate. For the German hyleme data, the spaCy lemmatizer delivered 287 wrong lemmata, yielding 93.87 % correct lemmata. The mispredictions include several re-occurring types of errors, including wrong tokens, often out-of-vocabulary proper names, but also lemmata derived from a third person present tense of verbs with vowel change, e.g. "läufen" instead of "laufen" (engl. "to walk/run"), from "er läuft" (engl. "he walks/runs"). In verb lemmata ending with *-n*, the model often assumed *-en*, e.g. "opferen" instead of "opfern" (engl. "to sacrifice"). Additionally, the lemmatizer predicted third person present tense forms as the verb lemma, e.g. "schläft" ("sleeps") instead of "schlafen" ("to sleep"). However, the spaCy lemmatizer (trainable lemmatizer)[8], which is the default lemmatizer for German since spaCy version 3.3, only performs lemmatization on tokens (using edit trees). It therefore consistently predicts lemmata of separable verbs[9] ("aufstehen", engl. "to get up/stand up") as the lemma of the base word ("stehen", engl. "to stand"), whenever the separable verb prefix ("auf-") is separated. When separable verb lemmata are included, the the performance of the spaCy lemmatizer drops to 81.85 % correctly identified lemmata. However, the correct lemmata can be easily reconstructed, given that the dependency tag *SVP* and the base lemma were correctly identified, by appending the base lemma to the separable prefix.

Despite including subword information in the prediction[10], the lemmatizer still most often predicts lemmata which include wrong infixes ("auf*ge*stehen"). Since hylemes are predominantly in present tense, the latter behaviour does not affect the performance of the lemmatizer on the German hyleme data.

The frequency of the distinct verbs after manual correction are shown in Figure 4.4. All four data sets show a similar trend, with few high-frequency verbs and a high number of low-frequency verbs. All four verb lemma statistics were analyzed according whether or not they follow a power-law distribution (Zipf's Law). This is a common behaviour of data concerning almost all natural language phenomena (see for instance [170]). A Kolmogorov–Smirnov test did not confirm a power-law distribution (p = 0.0). Therefore, the nullhypothesis could not be rejected.

[8] `https://spacy.io/api/edittreelemmatizer`

[9] Consistently in this case means that out of 859 identified *SVP*-dependency tags, the lemmatizer only reconstructs separable verbs lemmata correctly 12 times.

[10] `https://explosion.ai/blog/edit-tree-lemmatizer`

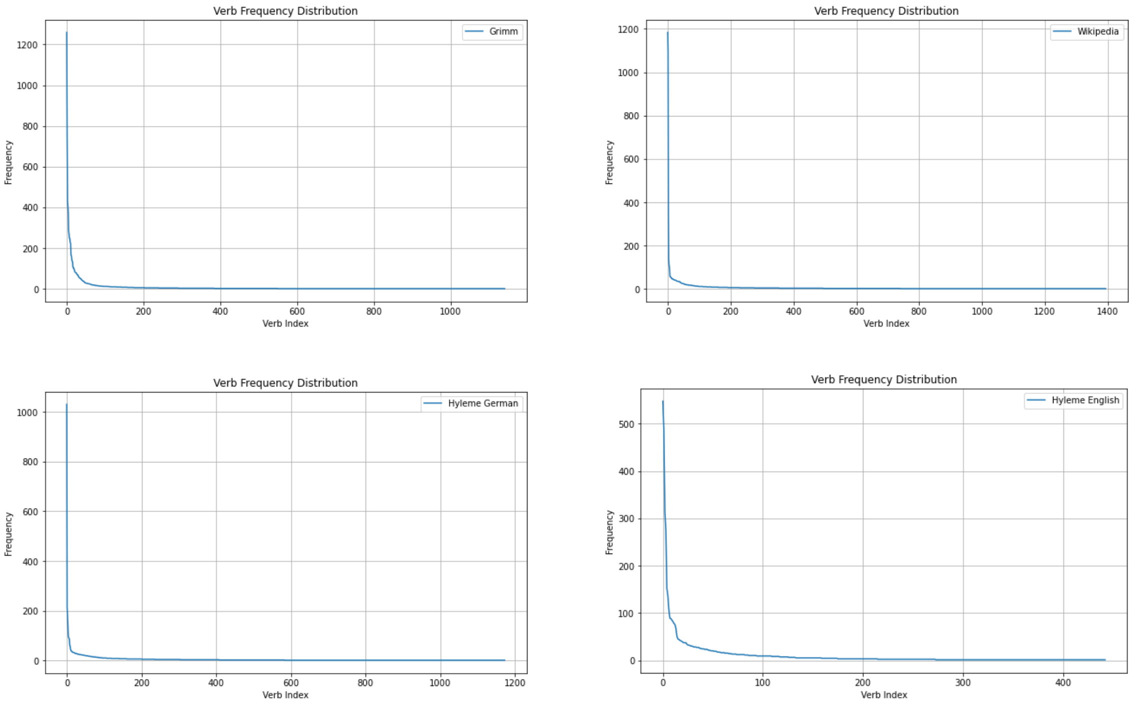


Figure 4.4: Verb Frequencies in different data sets

The top 30 % of verb occurrences in German hyleme predicates are shown in Figure 4.5. Table 4.2 compares the top 30 % of verbs in all three data sets. The verb lemmata *sein, haben, and werden* in the Wikipedia and Grimm data set are very frequent, which is also due to the fact that those auxiliary verbs are part of perfect and future tense constructions, such as *"Ich habe geschlafen."* (engl. "I (have) slept.") or *"Ich werde einkaufen gehen."* ("I will go shopping.") In Table 4.2 those verbs are marked with *.

In contrast, hylemes are always in present tense, therefore the verbs *haben* and *sein* are part of present tense constructions, such as "Die Höhle ist riesig." (engl. "The cave is huge.") Even though perfect and future tense constructions are not part of the occurrences of the verb lemma *sein* in the hyleme data, *sein* is still the lemma with the most occurrences in the data set (16.3 %). While *sein* is the lemma with the highest occurrences in the other two data sets, it occurs less frequently than in the German hyleme data (although it is part of the perfect tense constructions in those data sets). In the Grimm data set 11.2 % of the verbs are forms of *sein*, in the Wikipedia data set 13.7 %. The high occurrence of hylemes that use forms of *sein* is mainly due to the use of the verb in *durative-constant* hylemes. Interestingly, forms of *to be* make up only 9.39 % of the hyleme predicates in the English data set. This might be due to the extractor's choice to split long communication events, as in Sequence (2) in Figure 4.2 into multiple hylemes, instead of one. This approach aids hyleme

Table 4.4: Example corrections in the comparison data sets

Data set	Correct Lemma	Incorrect Lemma
Wikipedia	untertauchen (1)	untergetaucht (2)
		unterzutauchen (2)
Grimm	gehen (95)	gieng (49)
		Gieng (42)
		giengen (29)
		gehn (5)
		geh (1)
		Gienge (1)
		gehst (2)
		gienge (2)
		Geh (2)

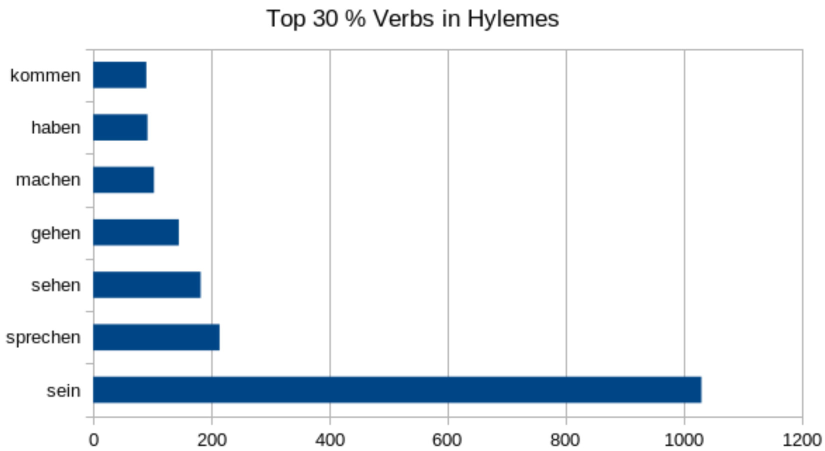


Figure 4.5: Top 30 % Verbs in Hylemes

comparison, and makes alignment more robust.[11] Therefore, the communication predicates, *tell, ask,* and *say* make up a higher portion of the hylemes than if the events where extracted as one hyleme.[12]

[11]e.g. "Uthlakanyana tells the cannibal to shake the bag. Uthlakanyana tells the cannibal to bring the bag." instead of "Uthlakanyana tells the cannibal to shake the bag and bring it." The former extraction allows hyleme alignment on either of the hylemes.

[12]Some NLP, resp. event modelling, approaches do not consider communication actions events. However, in the folkloristic and mythological context utterances can have a high significance. This is also conveyed in other narrative theories, such as Propp's *Morphology of the Folk tale*, where functions like *Interdiction* β and *Violation of Interdiction* γ (e.g. Red Riding Hood being told not to leave the house, but leaving anyway.) are of utmost importance. The fact that the $\beta\gamma$ pattern is present in folktales across cultures was illustrated by the author in [171].

Table 4.5: Top 30 % Verbs in different data sets

Hyleme German			Hyleme English			Grimm			Wikipedia		
Verb	Total	%	Verb	Total	%	Verb	Total	%	Verb	Total	%
sein	1030	16.3	tell	548	10.54	sein*	1259	11.2	werden*	1184	13.7
sprechen	214	3.4	be	486	9.39	haben*	746	6.9	sein*	1096	12.7
sehen	182	2.9	ask	316	6.10	wollen	432	4.0	haben*	349	4.0
gehen	145	2.3	say	273	5.27	sprechen	406	3.7			
machen	103	1.6				kommen	372	3.4			
haben	92	1.5									
kommen	90	1.4									

4.3 Named Entities

Apart from verb lemma extraction, the correct identification of named entities in hylemes is a crucial step for performing hyleme comparison and alignment. Additionally, the modelling and subsequent comparison of background information relies heavily on named entity, e.g. the hylemes "Umkxakaza is the princess" and "Untombinde is the princess" can only be used to model character representations of "Untombinde" and "Umkxakaza" if the two names are properly identified.

In the next section the distribution of named entities are discussed. For that purpose, the data sets are labelled with (named) entities using the named entity recognising (NER) functionality from the spaCy pipeline using the language models *de_core_news_sm*, *de_core_news_md* and *de_core_news_lg* (and the corresponding English models). The spaCy pipelines employs the WikiNER model [36] for the named entity recognition subtask. Since the NER tags *PER* and *LOC* are the most relevant for hyleme data, we focus the error analysis on those two classes. For the purpose of this comparison, we define the named entity classes *PER* and *LOC* as follows:

- Location *LOC*: A name that is representative of a real, fictional or mythological geographic location or space, e.g. Belarus, Mount Olympus, the Netherworld, Haltern am See.
- Person *PER*: A name that is representative of a real, fictional or mythological person or rational, self-aware entity, e.g. Commander Data, Pliny the Elder, Zeus, He who must not be named.

Firstly, we discuss the error analysis of the named entity recognition for the comparison data sets.

In the Grimm data set, the NER classifier found 3167 entities, of which 778 were true named entities according to the definitions above (precision ca. 0.25), from a set of 112 distinct names. The rest were false positives across all parts-of-speech, e.g. *Herz (PER), kohlschwarz (LOC), dir (PER), Wozu (PER), Oho!" (PER), Geh (PER)*. Of the 788 entities that were identified, 182 were contained in an incorrect span, either by including a punctuation mark (*Hänsel. (PER)*) or tokens before or after the named entity (*dir Gretel (PER), Katz von Kehrewitz, wie (PER)*. In extreme cases, the span included entire utterances *(Bremen, du verstehst dich (LOC)), Trude „warum bist du so bleich?" (PER)*. Of the 711

PER-labelled entities, 39 were initially mislabelled as LOC.

From the Wikipedia data set, all bibliographic information was removed, e.g. Further Reading, Sources or Bibliographies. On the basis of the descriptive texts, the classifier found 4855 named entities, of which 4171 entities were true named entities (precision ca. 0.86) with 2094 distinct names. The rest consisted of either tokens misclassified as named entity, e.g. *Neo-Rockabilly-Szene (LOC)* or *Teleskopmasten (LOC)*, or a true named entity from a class not PER or LOC, e.g. *Belavia Belarusian Airlines. (PER)*, which should be classified as *ORG*. Additionally, if multiple entities appear separated by commas or in lists, they were occasionally (50 times out of 4855 entities labelled) recognized as one, e.g. *Palermo, Catania, Messina, Syrakus (LOC)*. Those instances were separated into individual named entities. The WikiNER model was trained on data that was labelled in a semi-supervised approach. The label set included *LOC*, *ORG*, *PER*, *MISC* and *NON* (Non-entity). Those broader classes included fine-grained NE types. Particularly interesting for the Wikipedia data set that is used for the comparison is that the *ORG* type includes Band names (*BAND*), and the *OTHER* type includes *NORP*[13]. The Wikipedia data set contained person names and locations in different languages, e.g. Italian, Polish, Hawaiian, Georgian and Belarussian. In 355 out of 4115 instances, the name has been identified, but the span included either less or additional tokens, e.g. *Louis Breeland, der (PER)* or *Lee Allens Tenorsaxophonsolo (PER)*. Tokens with *PER* or *LOC* labels of ethnic groups or nationalities, such as "Belarussen" (engl. "Belarussians"), "Italiener" (engl. "Italians") were manually corrected to *NORP*, following [36]. Fictional named entites and names of deities were kept, because they play important roles in the hyleme data and the Grimm comparison data set. Sometimes, the output contained multiple errors like an incorrect span and a wrong label. These instances were counted towards both errors. In total, there were 1066 instances that contained at least one error. A detailed list of the errors can be found in Table 4.6. After data clean-up, the Wikipedia data set contains more LOC- than PER-labelled tokens (2410 resp. 1433).

The performance of the NER classifier is compared across data sets. For that purpose, a small set of randomly selected sentences/hylemes for each of the data sets were annotated with Named Entities (*PER, LOC, MISC, ORG, and Null (no entity)*). The spaCy models for German, *de_core_news_sm*, *de_core_news_md*, and *de_core_news_lg*, are compared against the gold standard of 99 annotated sentences/hylemes per German data set. For the English hyleme data set, the models *en_core_web_sm*, *en_core_web_md*, and *en_core_web_lg* were used. The gold standard annotation follows the named entity classes how they were used by [36].

Tables 4.7-4.18 show the performance of the NER classifiers across data sets. For the Wikipedia data set, we see that the model *de_core_news_lg* performs best. *PER*-labels are detected with a precision of 0.95, *LOC*-labels with a precision of 0.68.

The Grimm data set contains very few true named entities. This is mainly due to the tendency

[13]Nationality, Organizational, Religious, and Political affiliations

Table 4.6: Error analysis for Named Entity Recognition in Wikipedia data set

Error	Count	Example (pred. label)
token is not NE	388	*Pfefferminzgeschmack* (PER)
span too wide or too narrow	355	*Cap Gris-Nez,* (LOC)
multiple entities in prediction	49	*Niedersachsen, Nordrhein-Westfalen, Sachsen,* (LOC)
entity split over two predictions	8	*Ostküste* (LOC)
		Kanadas (LOC)
incorrect label	299	
- LOC, correct: PER	18	*Little Richards (LOC)*
- PER, correct: LOC	9	*Großrussland (PER)*
- PER, correct: MISC	20	*Terraferma* (PER)
- LOC, correct: MISC	38	*Nord Stream Pipeline* (LOC)
- PER, correct: NORP	3	*Staufer* (PER)
- LOC, correct: NORP	89	*Sizilianer* (LOC)
- PER, correct: NON	11	*Scheibenzüngler* (PER)
- LOC, correct: NON	40	*Bufotes boulengeri* (LOC)
- PER, correct: ORG	13	*Ronnex Records* (PER)
- LOC, correct: ORG	58	*Bezirksregierung Weser-Ems* (LOC)

to refer to characters in Grimm's tales by general descriptions, such as *the boy* or *the tailor*. The gold labels do not contain any instances of *LOC*-entities. In total, only 34 named entities were annotated. Consequently, the classifier struggles to find and label instances correctly. Nevertheless, the smallest spaCy model, *de_core_news_sm*, outperforms the other two models, with a precision of 0.56 for the *PER*-labelled instances.

In contrast to the Grimm data set, the hyleme data contains a relatively high number of named entities. This is due to the fact that hyleme sequences are not continuous texts, and hence do not use co-references ("the boy"-"he"). Instead, names are usually repeated in each hyleme. Compared to the Wikipedia data set however, the hyleme data contains a lot of named entities that are not part of the vocabulary of WikiNER (e.g. *Innana*, *Marduk*). The best-performing model for the German hyleme data is the large spaCy model, *de_core_news_lg*, which has a precision of 0.79 for the *PER*-labelled instances. However, the recall of 0.43 is lower than for the medium spaCy model, *de_core_news_md* (0.48). For the *PER* label, it yields an F1-score between 0.82 and 0.85, with the spaCy model *de_core_news_sm* performing best, albeit predicting the *ORG* occasionally, although it is not present in the data.

While the spaCy pipeline performed well on the English hyleme data set for the verb lemmatization task, it fails to perform named entity recognition on the English hyleme sequences entirely. This is mainly due to consistent mislabelling of the names as *MISC*. The alert reader will have noticed that the named entities mentioned so far, e.g. in Sequence (1) and (2) in Figure 4.2, start with the letter *U*. This feature in Callaway's text is due to the translation from the vernacular Zulu. Proper names of entities and nouns referring to persons fall under noun class I (singular, describing people). The *u* (usually lowercase) is the corresponding prefix. Hence, this prefix applies to non-Zulu names as well. In modern English texts that include Zulu names, the *u*-prefix is often dropped. Consider the following example:

- "Abacwaningi abane baseSADiLaR bethamele inkomfa yeALASA yangonyaka wezi-2022. Kusuka kwesobunxele kuya kwesokudla: u-Andiswa Bukula, uRooweither Mabuya, uBenito Trollip, uMuzi Matfunjwa."[14]
- Transl.: "Four SADiLaR researchers attending the ALASA 2022 conference. From left to right: Andiswa Bukula, Rooweither Mabuya, Benito Trollip, Muzi Matfunjwa."

Callaway's tales additionally use named entities that do not fall under noun class I, mainly mythological or animal-like characters. Those names include other prefixes, e.g. *isi-* in "Isikqukqumadevu". Therefore, a named entity dictionary for the English data set was created as a gold-standard, and for the application in downstream tasks. In total, the dictionary includes 68 named entities, which consists of 50 *PER* entities, four *LOC* entities, and 14 *MISC* entities, mainly referring to peoples or tribes and vernacular terms or names for animals.[15] The gold standard named entities for the German hyleme data contains 419 unique entities, of which 272 are *PER* entities, 105 are *LOC* entities, and 42 are *MISC*.

[14]Example and translation taken from the SaDiLAR newsletter, p.14, April 2023.

[15]If an animal shows human-like behaviour, it is labelled as *PER*. If an animal has a name, but behaves appropriately like an animal, it was labelled *MISC*.

Table 4.7: Classifier WikiNER against Wikipedia gold labels (spaCy model de_core_news_sm)

	Precision	Recall	F1-Score	Support
LOC	0.67	0.60	0.63	84
MISC	0.33	0.31	0.32	55
Null	0.96	0.96	0.96	1790
ORG	0.30	0.38	0.34	39
PER	0.80	0.90	0.85	72
accuracy			0.91	2040
macro avg	0.61	0.63	0.62	2040
weighted avg	0.92	0.91	0.92	2040

Table 4.8: Classifier WikiNER against Wikipedia gold labels (spaCy model de_core_news_md)

	Precision	Recall	F1-Score	Support
LOC	0.66	0.70	0.68	84
MISC	0.29	0.47	0.36	55
Null	0.98	0.95	0.97	1790
ORG	0.32	0.36	0.34	39
PER	0.83	0.90	0.87	72
accuracy			0.92	2040
macro avg	0.62	0.68	0.64	2040
weighted avg	0.93	0.92	0.92	2040

Table 4.9: Classifier WikiNER against Wikipedia gold labels (spaCy model de_core_news_lg)

	Precision	Recall	F1-Score	Support
LOC	0.68	0.64	0.66	84
MISC	0.33	0.42	0.37	55
Null	0.97	0.97	0.97	1790
ORG	0.46	0.33	0.39	39
PER	0.95	0.97	0.96	72
accuracy			0.93	2040
macro avg	0.68	0.67	0.67	2040
weighted avg	0.93	0.93	0.93	2040

Table 4.10: Classifier WikiNER against Grimm gold labels (spaCy model de_core_ news_sm)

	Precision	Recall	F1-Score	Support
LOC	0.00	0.00	0.00	0
MISC	0.00	0.00	0.00	2
Null	0.99	0.98	0.98	2469
ORG	0.00	0.00	0.00	3
PER	0.56	0.48	0.52	29
accuracy			0.97	2503
macro avg	0.31	0.29	0.30	2503
weighted avg	0.99	0.97	0.98	2503

Table 4.11: Classifier WikiNER against Grimm gold labels (spaCy model de_core_ news_md)

	Precision	Recall	F1-Score	Support
LOC	0.00	0.00	0.00	0
MISC	0.00	0.00	0.00	2
Null	0.99	0.98	0.98	2469
ORG	0.00	0.00	0.00	3
PER	0.30	0.28	0.29	29
accuracy			0.97	2503
macro avg	0.26	0.25	0.25	2503
weighted avg	0.98	0.97	0.97	2503

Table 4.12: Classifier WikiNER against Grimm gold labels (spaCy model de_core_ news_lg)

	Precision	Recall	F1-Score	Support
LOC	0.00	0.00	0.00	0
MISC	0.00	0.00	0.00	2
Null	0.99	0.97	0.98	2469
ORG	0.00	0.00	0.00	3
PER	0.35	0.28	0.31	29
accuracy			0.96	2503
macro avg	0.27	0.25	0.26	2503
weighted avg	0.98	0.96	0.97	2503

Table 4.13: Classifier WikiNER against German hyleme gold labels (spaCy model de_core_news_sm)

	Precision	Recall	F1-Score	Support
LOC	0.14	0.23	0.17	13
MISC	0.00	0.00	0.00	5
Null	0.93	0.93	0.93	653
ORG	0.00	0.00	0.00	0
PER	0.70	0.40	0.51	77
accuracy			0.85	748
macro avg	0.36	0.31	0.32	748
weighted avg	0.89	0.85	0.87	748

Table 4.14: Classifier WikiNER against German hyleme gold labels (spaCy model de_core_news_md)

	Precision	Recall	F1-Score	Support
LOC	0.22	0.54	0.31	13
MISC	0.00	0.00	0.00	5
Null	0.96	0.87	0.92	653
ORG	0.00	0.00	0.00	0
PER	0.73	0.48	0.58	77
accuracy			0.82	748
macro avg	0.38	0.38	0.36	748
weighted avg	0.92	0.82	0.86	748

Table 4.15: Classifier WikiNER against German hyleme gold labels (spaCy model de_core_ news_lg)

	Precision	Recall	F1-Score	Support
LOC	0.19	0.38	0.26	13
MISC	0.00	0.00	0.00	5
Null	0.94	0.91	0.93	653
PER	0.79	0.43	0.55	77
accuracy			0.84	748
macro avg	0.48	0.43	0.43	748
weighted avg	0.91	0.84	0.87	748

Table 4.16: Classifier WikiNER against English hyleme gold labels (spaCy model en_core_ web_sm)

	Precision	Recall	F1-Score	Support
MISC	0.05	1.00	0.10	3
Null	0.98	0.98	0.98	938
PER	0.00	0.00	0.00	52
accuracy			0.93	993
macro avg	0.35	0.66	0.36	993
weighted avg	0.93	0.93	0.93	993

Table 4.17: Classifier WikiNER against English hyleme gold labels (spaCy model en_core_ web_md)

	Precision	Recall	F1-Score	Support
MISC	0.05	1.00	0.10	3
Null	0.99	0.98	0.99	938
PER	0.00	0.00	0.00	52
accuracy			0.93	993
macro avg	0.35	0.66	0.36	993
weighted avg	0.93	0.93	0.93	993

Table 4.18: Classifier WikiNER against English hyleme gold labels (spaCy model en_core_ web_lg)

	Precision	Recall	F1-Score	Support
MISC	0.02	0.33	0.03	3
Null	0.98	0.98	0.99	938
PER	0.00	0.00	0.00	52
accuracy			0.93	993
macro avg	0.33	0.44	0.34	993
weighted avg	0.93	0.93	0.93	993

4.4 Topic Models

As a last method of data processing and exploration, this section wants to investigate topic models in hyleme data. Topic models are statistical models that aim to discover latent (hidden) patterns (*topics*) within a collection of documents or texts (see Section 2.3).

For the purpose of comparing the hyleme data sets, the hypothesis is that the German hyleme data has a greater variety of topics than the English data set. Therefore, the analysis of topic models in hyleme data is applied to the individual sequences (as documents).

For our purposes, the popular and widely accepted Latent Dirichlet Allocation (LDA) [55] approach is used. First, the hyleme data is pre-processed. So-called stopwords, words that are very frequent but have low semantic value, e.g. determiners, are removed. The remaining tokens in the hyleme sequences are lemmatized. LDA assigns a probability to a document-topic pairing, and probabilities for words to be part of a topic.

4.4.1 Hyper-parameter tuning

As introduced in Section 2.3, the LDA topic modelling algorithm uses three (hyper-)parameters. The number of topics, k, and the two scalar concentration parameters: the document-topic density α, and the word-topic density β. Essentially, a high α value assumes that documents contain more topics, whereas a higher β implies that topics consist of more words, which are specific to that particular topic. Additionally, the distribution of topics among the documents can either be *symmetric* or *asymmetric* [172].

To evaluate, which number of topics and which hyper-parameter setting for α and β are the most suitable, two measures are commonly applied. The first is the *perplexity*, which shows how well a topic fits a certain sample. A lower perplexity usually indicates a better fit of the samples to the respective models. However, if only the log-perplexity is considered, the topic models are at risk of overfitting, i.e. finding over-specific topics that are too small and can be only applied to very few documents (hyleme sequences) to yield good results. Therefore, the number of topics was first limited to a realistic range of $2 \leq k \leq 100$. This range assumes a maximum of 100 different topics in a collection of 228 documents for the German hyleme sequences, respectively 384 documents in the English hyleme data set. For the training, I use a chunksize of 100 documents and 10 passes through the corpus per training iteration.

The second evaluation measure that can be applied to topic models is the *coherence* score. In contrast to the perplexity, coherence is a measure of the informativeness and interpretability of a topic. The higher the coherence score, the more consistent the topics. Figures 4.6 and 4.7 show the coherence score and the perplexity for topics numbers $2 \leq k \leq 100$ with different hyper-parameter settings for the German data set. For visualisation purposes, we plot coherence scores for $0.01 \leq \alpha < 1$, and symmetric/asymmetric prior values (S/AS), in Figure 4.6a and Figure 4.6c (with holding $\beta = 0.31$)

and $0.01 \leq \beta < 1$, and symmetric prior values (S), (with holding $\alpha = 0.31$) in Figure 4.6b and 4.6d. The perplexity scores for the different settings follow accordingly in Figure 4.7. Additionally, all hyper-parameter settings are tested against the full corpus, and a 75 % corpus, which leaves a 25 % validation set.

In Figure 4.6, we see that the coherence values for $\alpha = 0.31$ and $\alpha = 0.91$ are highest on the full corpus, but $\alpha = 0.91$ does not perform well on the 75 % corpus. In Figure 4.6b and Figure 4.6d, we see that $0.31 \leq \beta \leq 0.91$ perform best, with $\beta = 0.31$ being less prone to fluctuations. For $\beta = 0.01$ and a symmetric β, the coherence score is lower for both corpora. For all hyper-parameter settings, the range $20 < k < 40$ shows the largest increase in coherence.

In Figure 4.7 the log-perplexity score is plotted against the number of topics. It becomes apparent, that the log-perplexity is lower for higher numbers of topics, in all hyper-parameter settings. The lowest perplexity is achieved by $\alpha = 0.91$, while $\beta = 0.01$ performs consistently low for both corpora (75 % and full corpus).

For the German hyleme data set, the best performing (hyper-)parameter setting yields a coherence of 0.65 and a perplexity of -7.45, with $\alpha = 0.31$, $\beta = 0.31$ and $k = 96$ (on the full corpus). If we limit $20 \leq k \leq 40$, the best performing LDA model yields a coherence of 0.58 and a perplexity of -7.44, with $\alpha = 0.31$ and $\beta = 0.91$ on $k = 36$ topics (on the full corpus). The evaluation for English hyleme data set (plots in Appendix A.1) yields similar results. The coherence scores are highest for $\alpha = 0.61$ and $\alpha = 0.91$. For the 75 % corpus split $\beta = 0.31$ and $\beta = 0.91$ performs best, but $\beta = 0.91$ leads to a strong fluctuation in coherence. For the full corpus, $\beta = 0.91$ outperforms $\beta = 0.31$ for $2 < k < 70$, for $k > 70$, $\beta = 0.31$ performs best.

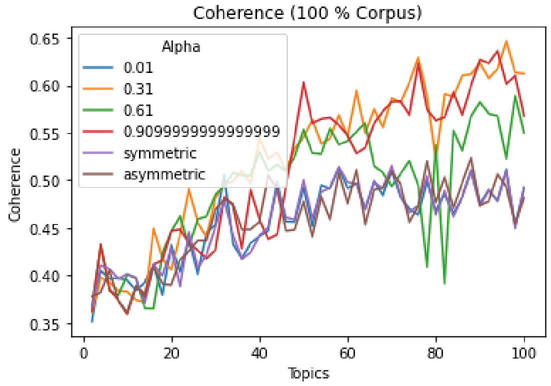
The log-perplexity for the different hyper-parameter settings shows similar behaviour to the German data set, ranging between -5.6 and -6.6 for both corpora for the values of α. The lowest log-perplexity for both corpus splits is achieved with $\beta = 0.91$.

The best performing (hyper-)parameter setting for the English hyleme data set yields a coherence of 0.68 and a perplexity of -6.57, with $\alpha = 0.91$, $\beta = 0.31$ and $k = 82$ (on the full corpus). For $20 \leq k \leq 40$, the best-performing hyper-parameter setting $\alpha = 0.91$, $\beta = 0.61$, with $k = 32$, yielding a coherence score of 0.55 on the 75 % corpus split.

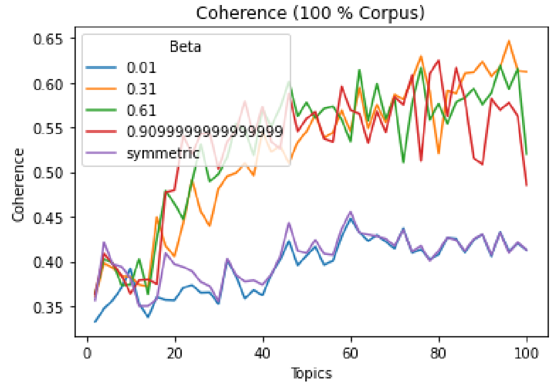
4.4.2 Results

Lastly, to further determine the interpretability of the topics, the LDA models can be investigated according to which words they contain and the sequences that can be associated with them. This can be achieved for instance by means of visualisation. For that purpose, we illustrate the topic models using the python library pyLDAvis. [173][16] Figures 4.8 and 4.9 shows the best performing topic models from the model with $k = 36$ and $k = 32$ for the both data sets, and an example

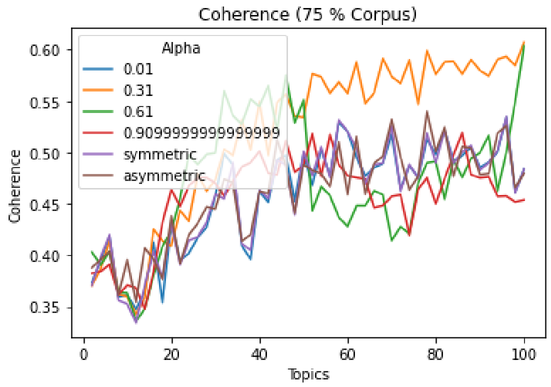
[16] `https://pyldavis.readthedocs.io/en/latest/readme.html`



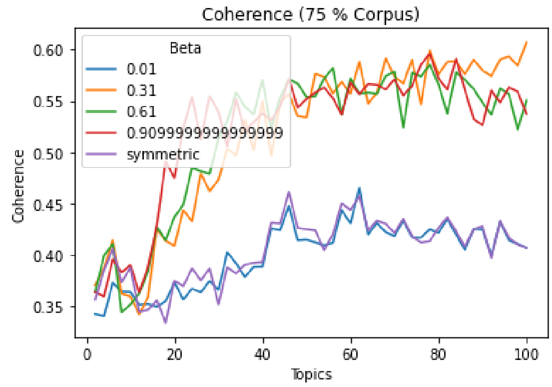
(a) Coherence for $0.01 \leq \alpha < 1$, *S/AS*, $\beta = 0.31$, 100 % corpus



(b) Coherence for $0.01 \leq \beta < 1$, *S*, $\alpha = 0.31$, 100 % corpus



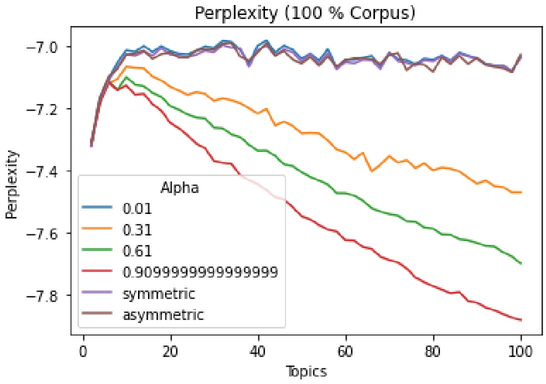
(c) Coherence for $0.01 \leq \alpha < 1$, *S/AS*, $\beta = 0.31$, $\alpha = 0.3$, 75 % corpus



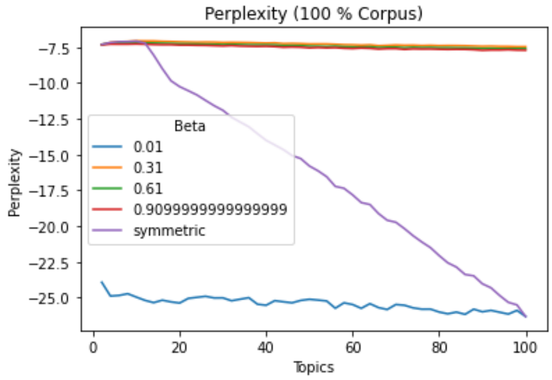
(d) Coherence for $0.01 \leq \beta < 1$, *S*, $\alpha = 0.31$, 75 % corpus

Figure 4.6: Coherence with different hyper-parameter settings, $2 < k < 100$ (German hyleme data set)

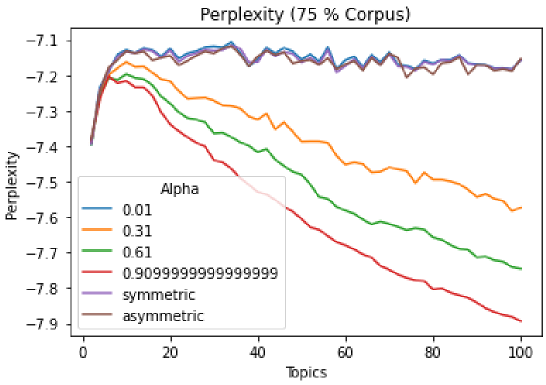
model selected. From the German example model, we select topic number 1, from which we can inspect the most frequent terms. The lemma "gott" (engl."god") is the most frequent term. We can see that the estimated term frequency of the word within the topic (red) is approximately the same as the overall frequency (blue). That means that it does not appear as often in other topics, the lemma is very distinctive for topic number 1. The second most frequent term is "erde" (engl. "earth"). This topic includes religious names and other terms that indicate it mainly applies to the religious hylemes in the data set. For the German hyleme data, the topics are centered around a nucleus, with many similar topics sharing common terms, and four larger topics (which together contain 66.7 % lemmata in the corpus). The four topics are very distinctive, with topic number one referring to largely religious terms, e.g. *god*, *noah*, *ark*, etc. Topic number 2 refers to ancient near eastern terms, e.g. *dumuzi, innana, demon* and *netherworld.* Topic number three includes topics from judea-religious background, such as *JHWH*, or *israel*. Lastly, topic number 4 describes ancient near



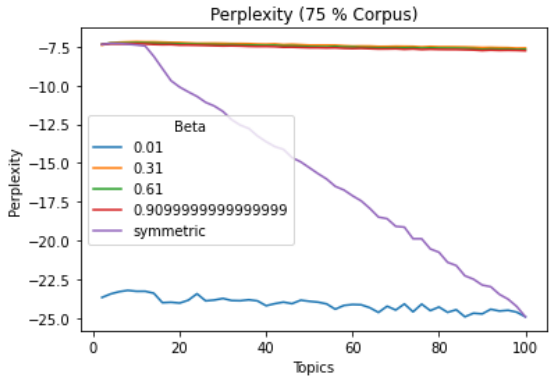
(a) Log-perplexity for $0.01 \leq \alpha < 1$, *S/AS*, $\beta = 0.31$, 100 % corpus



(b) Log-perplexity for $0.01 \leq \beta < 1$, *S*, $\alpha = 0.31$, 100 % corpus



(c) Log-perplexity for $0.01 \leq \alpha < 1$, *S/AS*, $\beta = 0.31$, $\alpha = 0.3$, 75 % corpus



(d) Log-perplexity for $0.01 \leq \beta < 1$, *S*, $\alpha = 0.31$, 75 % corpus

Figure 4.7: Log-perplexity with different hyper-parameter settings, $2 < k < 100$ (German hyleme data set)

eastern topics, concerning *enlil, enki, kingdom, and me-instruments of power*[17].

For the English hyleme data set, we see a more uniform topic distribution, although some topics are larger and more distinctive than others. We see that most topics are centered around a common vocabulary. But some distinct topics stand out, such as topic number 1, which contains terms concerning the interaction between an old woman and her son-in-law, who sends her away to find a stream where no frog cries (*Ugungqu-kubantwana* [7, pp.164]). From example topic number 1, we can also see the effect that a semi-standardization of the hyleme data has. Terms like *tell*, which is used to describe a certain communication event (as described above), are very frequent across all topics (blue bar).[18]

[17] On the me-instruments of power, see [174]

[18] The two model visualisations are available for inspection as interactive charts under: `https://teaching.gcdh.de/hyleme/topic-models/`

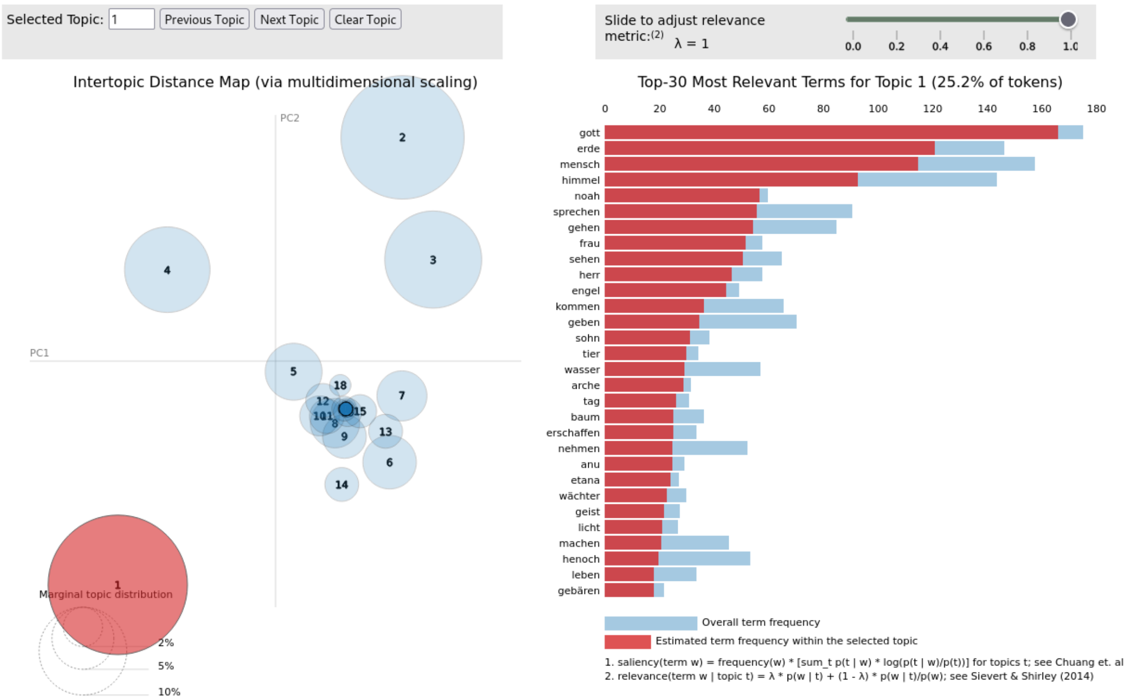


Figure 4.8: Best performing topic model, with example topic 2 selected, for the German hyleme data set

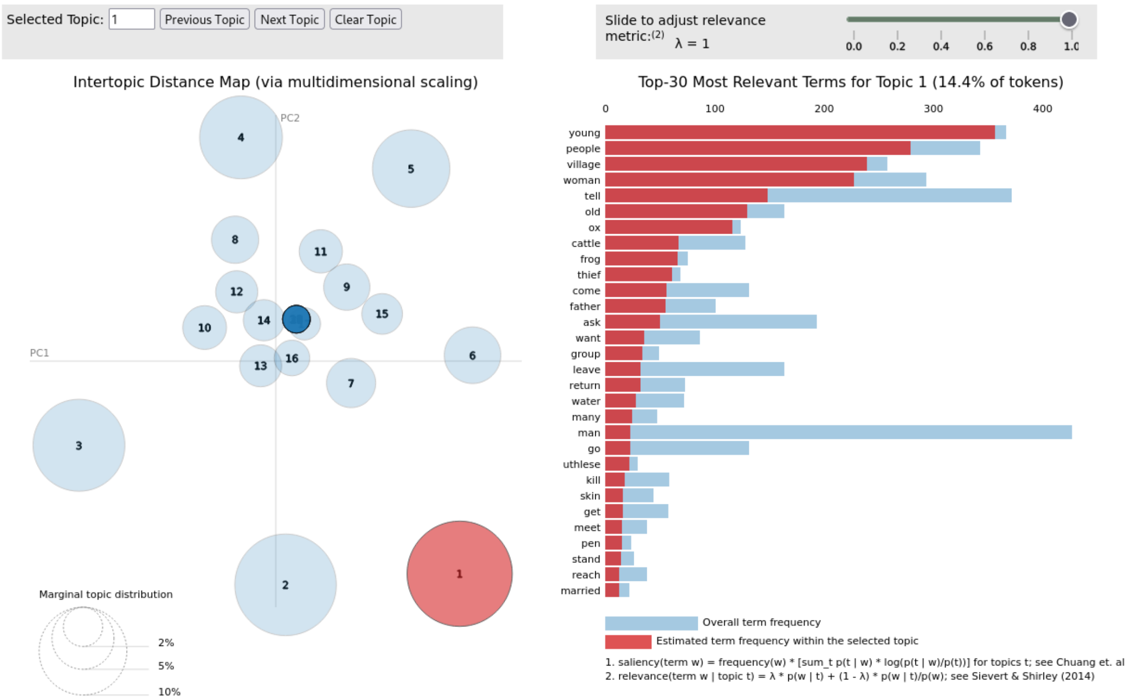


Figure 4.9: Best performing topic model, with example topic 2 selected, for the English hyleme data set

4.5 Temporal Semantics and Hyleme Types

After the previous sections investigated the hyleme data sets and their specific characteristics, I want to now investigate the temporal semantics of hyleme data. As introduced in Chapter 2.1, hylemes have different types: *single-event* hylemes, *durative-constant* hylemes, *durative-initial* and *durative-resultative* hylemes. While the English data set was annotated with hyleme types at the time of the data set construction, the German data set did not have any associated hyleme types. Therefore, we first describe and evaluate the annotation task, before the data sets are inspected according to their temporal semantic characteristics.

4.5.1 Hyleme Type Annotation

Six annotators annoated 6315 German hylemes according to their hyleme type. Since *durative-initial* and *durative-resultative* hyleme types are context-sensitive, i.e. they are invoked or their truth value is changed by previous or following hylemes, annotators always annotated entire sequences.

All the annotators possessed prior experience with the hylistic theory, mainly extracting hyleme sequences, and underwent annotation training in an initial annotation meeting. During this meeting, each annotator received a set of sequences to annotate individually, which were subsequently discussed as a group. To aid the annotators in making accurate choices for event categories when uncertain, they were provided with a set of example statements and guiding questions presented in a flowchart.

Furthermore, annotators who had explicit knowledge in the respective disciplines, such as Classics, were requested to consult the original sources for guidance whenever doubts arose. For instance, the English statement “Orpheus brings back the dead (from the netherworld)” could be interpreted as either *single-event* or *durative-constant*. However, by referring to the original Greek source and identifying the use of the imperfect form, it was determined that the correct annotation should be *durative-constant* [175]. In a subsequent meeting, any questions that arose during the annotation process were thoroughly discussed. Table 4.19 gives an overview of the annotators’ backgrounds, their affiliation with the STRATA project, and the number of items they annotated as either first or second annotator. Each hyleme sequence was annotated twice. Eleven different annotator pairs, with varying first and second annotators, annotated the German hyleme data set. In most cases, the inter-annotator agreement (Cohen’s κ) for the task ranges from substantial ($0.61 \leq \kappa \leq 0.80$) [9] to almost perfect agreement ($0.81 \leq \kappa \leq 0.99$). The agreement is reported in Table 4.20. Annotator pairs A2-A4 and A4-A5 have perfect agreement over the shared annotations. Pair A2-A5 has a relatively low value of κ = 0.4. This is due to one particularly long sequence containing 114 hylemes (two shared sequences between A2 and A5 in total). Many hylemes in this sequence pertain to the descriptions of a mythical house, e.g. “The vault of the house is a rainbow”. These were annotated as *durative-constant* by one annotator, while the other interpreted these descriptions as results of some action in the sequence (i.e. the building of the house), and therefore annotated them as *durative-resultative*. Consequently, event type annotations of all descriptions of the house in that sequence are mismatching (consequential error). This results in a low overall κ for the annotator pair A2-A5.

In cases where the first and second annotation did not match, the gold standard was established through a dedicated discussion in the second annotation meeting. Additionally, the judgement of the annotator from the discipline relevant to the sequence, e.g. Classics, was used to resolve the discrepancies. The performance of the annotators against the gold standard, along with the total number of annotated items, is presented in Table 4.21.

Annotator	Background	Level of Education	Affiliation	Gender
A1	ANES	B.A.	Student Assistant	
A2*	CS/CL	M.Sc.	PhD student	
A3	Classical Studies (CS)	B.A.	Student Assistant	
A4	ANES/DH	Doctoral Degree	Intern	
A5	ANES	B.A.	Student Assistant	
A6	ANES	M.A.	PhD student	

Table 4.19: Annotators' background (* = the author)

Table 4.20: Inter-annotator agreement (Cohen's κ) between pairs of annotators

Pair	No. of items	Cohen's κ
A1-A2	4552	0.848930
A1-A3	398	0.874665
A1-A4	299	0.929306
A1-A5	149	0.733025
A1-A6	96	0.631285
A2-A3	187	0.918325
A2-A4	90	1
A2-A5	127	0.402008
A3-A4	136	0.866710
A3-A5	239	0.811959
A4-A5	42	1

In contrast to annotators A1-A4, who received dedicated annotation instructions as described in Section 2.3.6, A5 and A6 did not receive additional training beyond being a member of the research group. A5 and A6 assigned hyleme types when they were instructed to enter hyleme sequences into the hyleme database. Additionally, A5 was asked to assign hyleme types to previously entered sequences where hyleme types were missing by another member of the research group.

The final gold labels are an important foundation for the next analyses. The distribution of the labels for both hyleme data sets is shown in Table 4.10. The majority of the data consists of *single-event* hylemes. Of the *durative* hylemes, the *durative-constant* hylemes are the largest group. Three hylemes had to be excluded from the data, because their types could not be determined (e.g. "The kur-ĝara and gala-tur ...?").

Table 4.21: Cohen's κ of annotators against Gold standard

Annotator	Gold	No. of items
A1	0.939978	5494
A2	0.914271	4956
A3	0.951389	960
A4	0.953625	567
A5	0.705362	557
A6	0.631285	96

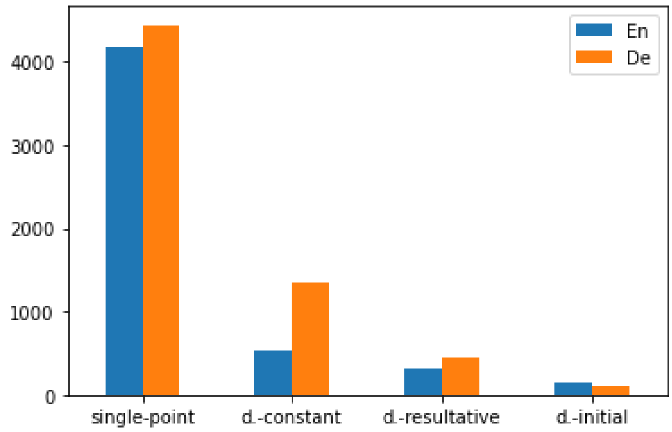


Figure 4.10: Distribution of the hyleme types in the final data sets (gold standard annotation)

4.5.2 Hyleme Type Classification

Based on the gold labels of the annotation as described in the previous section, three hyleme type classifiers were trained for each of the data sets. The resulting models can be used to pre-classify new hylemes, or to classify statements from other sources in comparison data sets in subsequent analyses.

The separation of the data into *durative* and *single-event* hylemes is an important first step for the subsequent analyses of the hyleme sequences, since *single-event* hylemes correspond to actions, whereas *durative* hylemes correspond to states, habituals and background information.

The task to automatically classify hyleme types is not trivial. Especially, automatically distinguishing the three types of *durative* hylemes is challenging. This is due to multiple reasons. Firstly, the three classes are unbalanced, with more *durative-constant* hylemes, and very few *durative-initial* hylemes, as shown in Figure 4.10. Additionally, *durative-initial* hylemes can be quite similar to *durative-constant* hylemes in terms of vocabulary and grammatical structure. Their hyleme type values are often context-sensitive. As an example, consider the sentence *Harry Potter lives in No. 4 Privet Drive* in the three sequences in Tables 4.22, 4.23, and 4.24.

Table 4.22: Context-sensitivity Example 1, *durative-initial*

	Hyleme	Type
1	**Harry Potter lives in No. 4 Privet Drive.**	durativ-initial
2	Harry Potter receives his letter from Hogwarts.	single-event
3	Harry Potter is a student at Hogwarts.	durativ-resultative

Table 4.23: Context-sensitivity Example 2, *durative-resultative*

	Hyleme	Type
1	Harry Potter is the son of James and Lily Potter.	durativ-constant
2	Voldemort kills James and Lily Potter.	single-event
3	Hagrid brings Harry to the home of the Dursleys.	single-event
4	**Harry Potter lives in No. 4 Privet Drive.**	durativ-resultative

Table 4.24: Context-sensitivity Example 3, *durative-constant*

	Hyleme	Type
1	**Harry Potter lives in No. 4 Privet Drive.**	durativ-constant
2	Harry Potter lives in the cupboard under the stairs.	durativ-constant
3	The Dursleys and Harry go to the zoo.	single-event
4	Harry accidentally frees a boa constrictor.	single-event

For the classification task, a multinomial naive bayes model was selected. For that purpose, the hyleme data set was split into a training and test set with a split of 75 %-25 %. The hyper-parameters were selected by performing a grid search. In particular, the grid search established whether the feature vector is best constructed using a bag-of-words or TF-IDF vectorizer. As a result of the grid search, the hyper-parameters were set as: Laplace smoothing parameter $\alpha = 0.01$, bag-of-words features, and an ngram range of 3.

Firstly, we analyze the results for binary classes *single-event* and *durative*, which includes *durative-initial*, *durative-constant*, and *durative-resultative* hylemes. For that purpose, all three labels were subsumed under the coarse class *durative* for training. The binary classifier performs well on *single-event* hylemes, and reasonably on *durative* hylemes. The performance of the classifier on the hyleme data sets (*DE* = German, *EN* = English) is reported in Table 4.25. The performance of the binary classifier for German is better than for the English hyleme data, due to a higher number of examples for the *durative* classes in the training data.

Table 4.25: Results of the binary hyleme type classifier

Data Set	Label	Precision	Recall	F1-Score	Support
DE	durative	0.83	0.75	0.79	427
EN		0.77	0.64	0.70	243
DE	single-event	0.91	0.94	0.92	1151
EN		0.92	0.96	0.94	1051
DE	accuracy			0.89	1578
EN				0.90	1294
DE	macro avg	0.86	0.85	0.85	1578
EN		0.84	0.80	0.82	1294
DE	weighted avg	0.89	0.89	0.89	1578
EN		0.89	0.90	0.89	1294

Secondly, we investigate how the classifiers perform if trained on just the different types *durative* hylemes. For that purpose, all *single-event* hylemes were removed from the training and test sets. The majority of the German test set consists of *durative-constant* hylemes (0.69 %) and *durative-resultative* hylemes (24 %). The distribution in the English training set is similar, with 54 % *durative-constant* hylemes, and 32 % *durative-resultative* hylemes. Since the English data set consists of less *durative* hylemes, the durative classifier does not perform as well as on the German data set. The results are reported in Table 4.26.

Lastly, we present two classifiers for the classification of fine-grained classes for each data set. The first classifier was trained on the entire training and test sets including fine-grained *durative* classes. The second classifier combines the first two models (binary and durative-only) in two steps. Firstly, the test data was classified into coarse classes, *single-event* and *durative*. Secondly, the hylemes that were classified as *durative* were re-classified by the classifier that was trained only on *durative* hylemes. Table 4.27 shows the performance of both approaches. The classifier that was trained on the fine-grained training data slightly outperforms the two-step approach in terms of overall

Table 4.26: Results of the durative hyleme type classifier

Data Set	Label	Precision	Recall	F1-Score	Support
DE	durative-initial	0.50	0.23	0.32	30
EN		0.42	0.36	0.43	33
DE	durative-constant	0.81	0.90	0.85	294
EN		0.70	0.81	0.75	136
DE	durative-resultative	0.62	0.51	0.56	103
EN		0.61	0.52	0.56	81
DE	accuracy			0.76	427
EN				0.66	250
DE	macro avg	0.64	0.55	0.58	427
EN		0.61	0.56	0.58	250
DE	weighted avg	0.74	0.75	0.75	427
EN		0.64	0.66	0.65	250

accuracy for the German data set. For the English data set, the two-step approach works slightly better. Confusion matrices for both classifiers and both data sets are reported in Figures 4.11 and 4.12. We can see that the classifiers favour the *single-event* class.

Table 4.27: Comparison between classifier trained on all data (left), and two-step classifier (right)

Data Set	Label	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
DE	durative-initial	0.50	0.17	0.25	0.43	0.20	0.27	30
EN		0.41	0.19	0.26	0.69	0.31	0.42	36
DE	durative-constant	0.72	0.67	0.69	0.70	0.69	0.69	294
EN		0.69	0.52	0.60	0.66	0.65	0.65	128
DE	durative-result.	0.55	0.45	0.49	0.51	0.44	0.47	103
EN		0.49	0.38	0.43	0.67	0.52	0.59	79
DE	single-event	0.90	0.95	0.93	0.91	0.94	0.92	1151
EN		0.91	0.97	0.94	0.92	0.96	0.94	1051
DE	accuracy			0.85			0.84	1578
EN				0.87			0.88	1294
DE	macro avg	0.67	0.56	0.59	0.64	0.57	0.59	1578
EN		0.63	0.52	0.56	0.73	0.61	0.65	1294
DE	weighted avg	0.84	0.85	0.84	0.84	0.84	0.84	1578
EN		0.85	0.87	0.85	0.87	0.88	0.87	1294

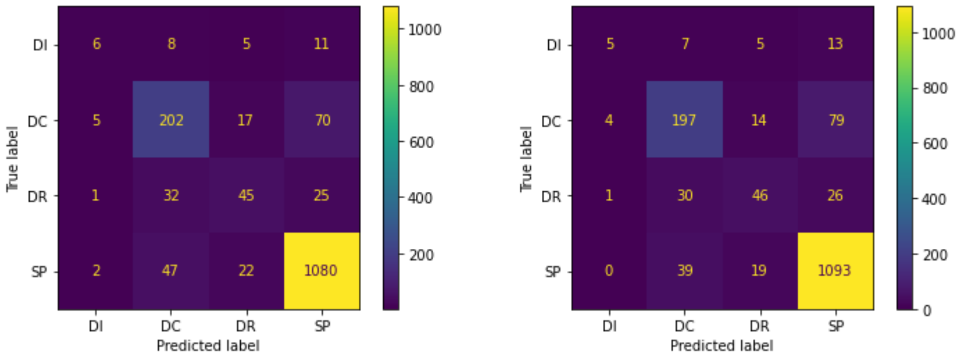


Figure 4.11: Confusion matrices for the classifier trained on all classes (left) and the two-step classifier (right) for the German data set, DI = *durative-initial*, *durative-constant*, DR = *durative-resultative*, SP = *single-event*

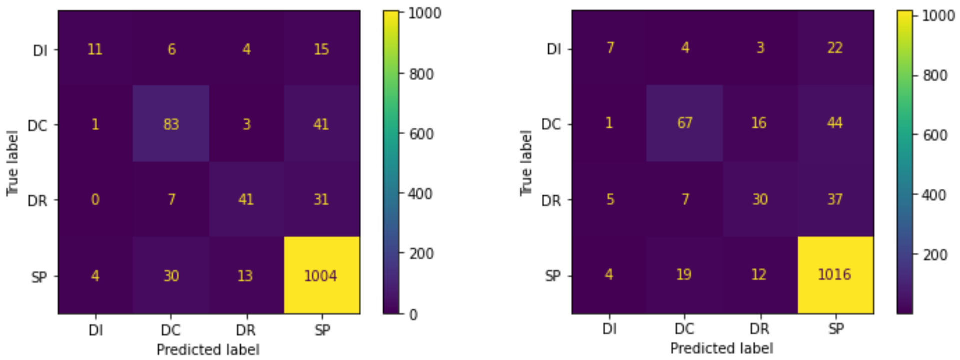


Figure 4.12: Confusion matrices for the classifier trained on all classes (left) and the two-step classifier (right) for the English data set, DI = *durative-initial*, *durative-constant*, DR = *durative-resultative*, SP = *single-event*

4.6 Summary

This chapter introduces the hyleme data, and compares them to two other data sets. The German and English hyleme data sets show particularities to other standardized and unstandardized language data. The German hyleme data set has a strong focus on the predicate *sein* (engl. *to be*). In the English hyleme data, predicates describing communication events are more prevalent. On the English hyleme data, the spaCy NLP pipeline, especially the dependency parser and the lemmatizer, work well, with the lemmatizer yielding 99.56 % correct lemmata for hyleme predicates. For the German data, the lemmatizer does not perform as well, especially if separable verbs are concerned. However, since separable verbs can be easily reconstructed given a correct lemma of the base verb, the lemmatizer yields overall good results (93.87 %).

Regarding the named entities, the spaCy component fails to perform named entity recognition properly on the English hyleme data set. This is mainly due to the names in Callaway's tales being completely out of vocabulary for the boolean models and following Zulu syntax, including *u*-prefixes. For both tasks, lemmatization and named entity recognition (NER), the spaCy pipeline struggles with rare or out-of-vocabulary terms. Due to the diversity of myth variants and topics in the German hyleme data set, it shows a greater variety of named entities, both places (*LOC*) and persons (*PER*). The English hyleme data, which consists of hyleme sequences describing Callaway's tales, does not include as many named entities across all classes. In order to achieve reliable named entity recognition on the both hyleme data sets, entity dictionaries for both data sets were manually constructed. These ensure reliable results for further analyses, including the hyleme comparison (see Chapter 7.2) and the modelling of background information (see Chapter 5.1).

For both hyleme data sets, topic modelling is used as a method of data inspection. After hyper-parameter tuning, we can identify coarse and fine topics that coincide with hyleme sequences of specific origins or that apply to semantically similar hyleme sequences.

Lastly, this chapter reports on the annotation study for the hyleme types in the German hyleme data set, which were not initially included. After initial training, the annotators reached an overall satisfying inter-annotator agreement κ. When investigating hyleme type distribution, we see that in both gold standard data sets the actual distribution of labels, i.e. *single-event* hylemes (i.e. actions) are more prevalent than *durative* statements. The gold standard data is then used to train classifiers for both data sets. Due to context-sensitivity and data sparseness, the classification task is not trivial. For that reason, a coarse classifier is presented, which distinguishes between *durative* and *single-event* hylemes. This classifier yields overall satisfying results. Fine-grained classifiers suffer from data sparseness, especially for the *durative-initial* class.

The annotation study and the classifier were presented at the 17th Linguistic Annotation Workshop (LAW-XVII), co-hosted at the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, July 9-14, 2023.

Chapter 5

Domain-specific Knowledge Engineering

In Chapters 2 and 4 different types of hylemes were introduced and annotated. For the comparison and alignment of hyleme sequences, a separation between *single-event* hylemes (actions) and *durative* hylemes (states and habituals) is crucial. Action-bearing information is central for modelling narrative plots. On the other hand, states, background information and habituals can be used to compare the domains underlying a myth (or other narrative) variant. Background information is often not directly comparable, because characters or circumstances can be described and represented differently. In the mythological context, similarity is often expressed through the same relations, attributes or roles of characters. As an example, Table 5.1[1] shows *durative-constant* hylemes from two invocations of the Mesopotamian demon *Lamaštu*.

Table 5.1: Durative-constant Hylemes derived from the two invocations of *Lamaštu*, implicit information in square brackets

Invocation 3	Invocation 7
Lamaštu is the daughter of Anu.	*Lamaštu* is the daughter of the Anu of the sky.
[*Lamaštu*] is scary.	*Lamaštu* is scary.
[*Lamaštu*] is not a goddess.	*Lamaštu* is a goddess.
[*Lamaštu*] is a she-wolf.	*Lamaštu* has the teeth of a donkey.
	Lamaštu has the face of an enormous lion.
	Lamaštu's cheeks are pale yellow.
	[*Lamaštu* loves the mountains.]

They contain three different types of information: 1. Relationships between two entities in the domain, e.g. "Lamaštu is the daughter of Anu.", 2. Assignment of an entity to a category, e.g. "Lamaštu is a she-wolf.", and 3. Attributes of an entity, e.g. "Lamaštu is scary." Hylemes describing an entity in multiple narrative variants can be contradictory, e.g. *goddess* vs. *not a goddess*. Approaches that combine information from different sources and myth variants often fail to model

[1] The hylemes in this example have been extracted by a domain expert of Ancient Near Eastern Studies and translated by the author.

those differences and contradictions. For instance, the Wikidata entry for Lamaštu[2] lists instances of both classes, both *demon* as well as *goddess*.

This chapter demonstrates how background information inherent to narratives can be modelled, re-used and compared. For that purpose, this work follows a semantic domain modelling approach. For each narrative variant the domain information communicated in *durative-constant* hylemes is used as ground-truth from which a minimal ontology is constructed. These minimal ontologies can be used for domain comparison, and serve as a re-usable source of *variant-specific* domain information.

5.1 Modelling Background Information in Minimal Ontologies

To construct an ideal minimal ontology, the information contained in *durative-constant* hyleme serve as *axioms*, or ground-truth. Characters or entities in the hylemes are modelled as individuals in the ontology. Class assignments of entities are derived from predicative nominal subject complements (e.g. "Lamaštu is a she-wolf.") Entity attributes are communicated as predicative adjective subject complements (e.g. "Lamaštu is scary."), or as prepositive adjectives (e.g. "The scary Lamaštu is a she-wolf.") They are modelled as data properties in the ontologies. Relationships (modelled as object properties) between class individuals or between classes are conveyed either by verbs that are not forms of *to be*, or by more complex constructions, or implied, e.g. by inference of relations and *isA-relationships*. For instance, the hyleme "Lamaštu is the daughter of Anu" implies:

1. *Daughter(Lamaštu)* (Class Assignment)
2. *isDaughterOf(Lamaštu, Anu)* (Relation)
3. *Father(Anu)* (Class Assignment)
4. *isFatherOf(Anu, Lamaštu)* (Relation)
5. *Father isA Parent isA Ancestor isA Person* (Class Hierarchy)
6. *Daughter isA Child isA Descendant isA Person* (Class Hierarchy)

The resulting OWL2 ontologies are shallow in the sense that they only model basic class hierarchy, in order to facilitate comparability of the domains. Additionally, alternative spellings of names or known aliases are supplied as a data property to individuals, e.g. "Lamaštu" and "Lamashtu". Aliases are an important property for the matching of characters in multiple hyleme sequences.

Where available, corresponding Wikidata ID or Pleiades ID[3] (for geolocations) are provided for the ontology individuals. Ontology labels and individual attributes are available in German and English. The ideal domain ontologies are carefully, manually crafted in an iterative process, while

[2] https://www.wikidata.org/wiki/Q767220
[3] https://pleiades.stoa.org/

being reviewed by a domain expert (Classics (CS) and Ancient Near Eastern Studies (ANES)), with *durative-constant* hylemes cross-referenced against original sources where possible.[4]

5.2 Example Ontologies

In this work, two sets of manually modelled ontologies are presented as ideal minimal ontologies. The first set of ontologies is derived from a variety of hyleme sequences pertaining to the Mesopotamian deity *Dumuzi* and his death. The sequences describe variants of different myths.

Some statements are present in some narratives, but not in others, for instance "Dumuzi is a shepherd." If the statement appears in a sequence S_1, we can assume the axiom *shepherd(Dumuzi)* to be *true* in S_1. However, if the statement is not part of the set of *durative-constant* hylemes in a different sequence S_2, no assumption can be made about the truth value for the second myth variant that S_2 describes. More specifically, ¬ *shepherd(Dumuzi)*, can only be assumed to be *true*, if an explicit hyleme *Dumuzi is not a shepherd* is part of the hyleme sequence S_2.

Eight hyleme sequences pertaining to *Dumuzis Death*, derived from eight sources (see Table 5.2), are the basis for the domain ontologies. They are available in TTL-format. Examples from the myth variant *Dumuzi-Durtur-Eršema* are shown in Figure 5.1 and 5.2.

Table 5.2: Sources of the hyleme sequences used for extraction of background information (*Dumuzi's Death*)

Title	Source
Death of Dumuzi	Kramer (1980) [176]
Song of Innana and Dumuzi (J)	ETCSL Nr. 4.08.10
Innana-Dumuzi Lament (CUNES 53-08-060)	Cohen (2014) [177]
Dumuzi and Geštinanna	ETCSL Nr. 1.4.1.1
Dumuzi Lament (ASJ 7, 1–9)	Alster (1985b) [178]
Innana-Dumuzi-Eršema (BM 15821)	Cohen (1981) [179], Ershemma No. 165
Innana-Dumuzi-Balaĝ (BE 30/1, 7)	Fritz (2003) [180, p.131-132]
Dumuzi-Durtur-Eršema (CT 15, pl. 20-21)	Cohen (1981) [179], Ershemma No. 88

The second set of ontologies is derived from 12 myth variants of the *Orpheus and Eurydice* (see Table 5.3). A controlled vocabulary was constructed to facilitate mapping different representations of relations and classes to the same concepts and relations. Each SKOS concept in the controlled vocabulary has a German and English label, a shallow hierarchical structure (skos:broader), and where applicable each entry is connected to the open data source Wikidata (skos:exactMatch or skos:narrowMatch). Lastly, each entry has a definition (in German) that aids the assignment of the correct class or relation for the ontologies. As an example, Figure 5.3 shows the individual-centric view of the minimal ontology for the hyleme sequence derived from Apollodorus' library.

[4]Additionally, the shallow ontologies, and their comparison was presented and discussed among the members of the STRATA myth research group in November 2022.

Figure 5.1: Individual *Dumuzi* data properties, Tool: Webprotégé

```
<hatEmotionaleBeziehung> rdf:type skos:Concept, owl:ObjectProperty ;
  skos:prefLabel "hatEmotionaleBeziehung"@de ;
  skos:altLabel "hasEmotionalRelationship"@en ;
  skos:narrowMatch <https://www.wikidata.org/wiki/Q1334052> ;
  skos:narrowMatch <https://www.wikidata.org/wiki/Q736922> ;
  skos:related <Person> ;
  skos:definition """Emotionale Verbindung zwischen Personen (z.B. Eltern-Kind, Ehefrau-
      Ehemann)"""@de ;
  skos:topConceptOf <Orpheus> .
```

Listing 5.1: SKOS concept *hatEmotionaleBeziehung*/*hasEmotionalRelationship* in *Orpheus* Controlled Vocabulary

```
<Herrscher> rdf:type skos:Concept, owl:Class ;
  skos:prefLabel "Herrscher"@de ;
  skos:altLabel "ruler"@en ;
  skos:broader <Person> ;
  skos:exactMatch <https://www.wikidata.org/wiki/Q1097498> ;
  skos:definition """Machthabendes Oberhaupt eines Volkes oder Territoriums"""@de ;
  skos:inScheme <Orpheus> .
```

Listing 5.2: SKOS concept *Herrscher*/*ruler* in *Orpheus* Controlled Vocabulary

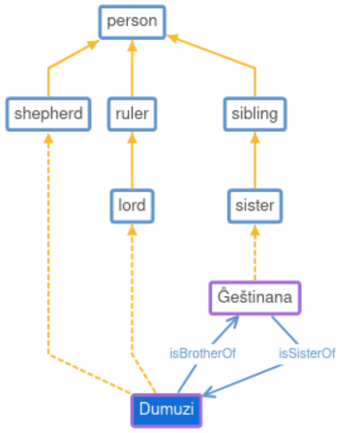


Figure 5.2: Individual-centric visualisation for the hyleme sequence "Dumuzi-Durtur-Eršema", Tool: Webprotégé

Table 5.3: Sources of the hyleme sequences used for extraction of background information (*Orpheus*)

Abbreviation	Source
AB_32	Apollodorus Library 3,2
BCP_312	Boethius Consolatio Philosophiae, 3,12
FM_3	Fulgentius Mythologiae, 3
HC_311	Horace Carmina, 3,11
HF_7	Hermesianax Fragment 7
KD_45	Konon Diegeseis
MA_1	(Marcus) Manilius, Astronomica, 1
MA_5	(Marcus) Manilius, Astronomica, 5
MV_i	Mythographus Vaticanus, i
OM_10	Ovid Metamorphosis, 10
P_9	Pausanias, 9
PS_179	Plato Symposium, 179d

The controlled vocabulary and the minimal ontologies are available to be re-used, under a creative-commons license (CC-BY 4.0)[5].

[5]https://gitlab.gwdg.de/franziska.pannach/hylva

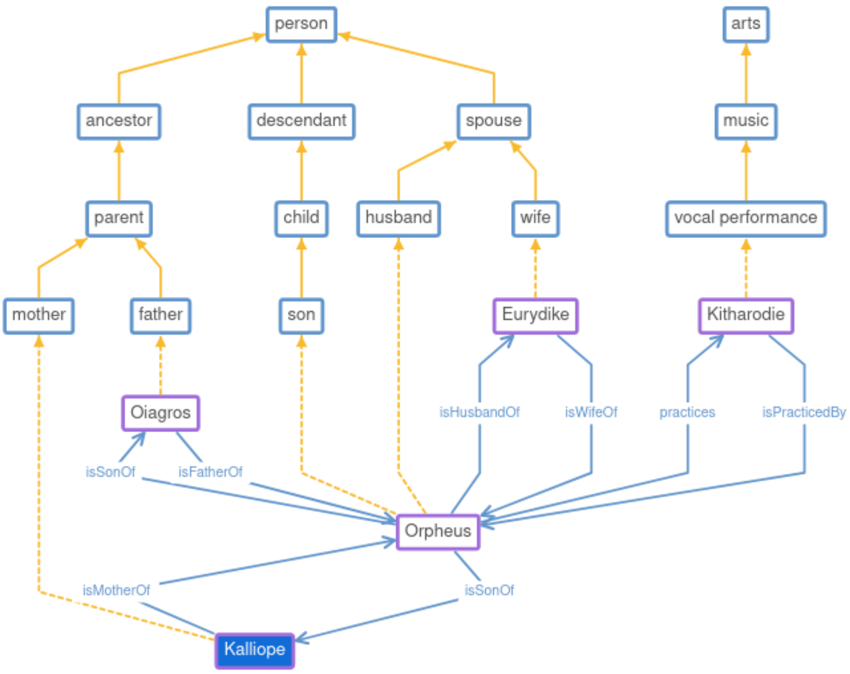


Figure 5.3: Individual-centric visualisation of example ontology for the hyleme sequence Apollodorus' library, Tool: Webprotégé

5.3 Ontology Comparison

The shallow domain ontologies can be used for inter-myth and intra-myth comparison. When comparing variants of the same myth the objective is to find the most similar ontologies, or to discover interesting deviations. Ontologies of different myth variants allow insights into similarities between narrative settings of related or unrelated material.

In this work, overall domain comparison is achieved by applying two measures: 1. The *class overlap*, which can answer questions like "Which kinds of entities appear in the myth variant?", 2. The *individual overlap* determines which characters, locations or other entities are part of the myth variants.

Ontology individuals are matched by their labels in German or English, their aliases or identifier (Wikidata ID or Pleiades ID). Class overlap is defined as essentially the Jaccard distance (see Section 2.3) of the sets of ontology classes:

$$CO = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}, \tag{5.1}$$

where C_1 and C_2 are the sets of classes of ontology O_1 and O_2 respectively. The overlap of the ontology individuals (IO) is measured accordingly.

The results of these two comparisons are presented in distances matrices in Figures 5.4 and 5.5. For the *Dumuzi* ontologies the highest class overlap appears between the pair "Innana-Dumuzi-Eršema (BM 15821)" and "Innana-Dumuzi Lament (CUNES 53-08-060)" with a value of $CO = 0.55$. For individual overlap the pair "Dumuzi-Durtur-Eršema (CT 15, pl. 20-21)" and "Innana-Dumuzi-Balaĝ (BE 30/1, 7)" have the highest value of 0.5.

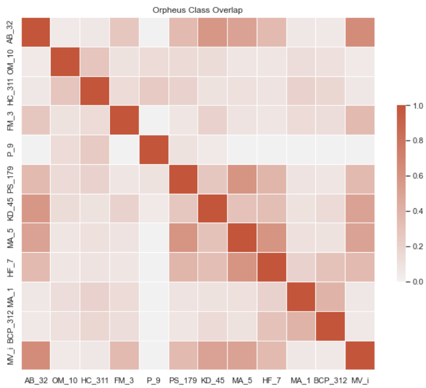
Figure 5.4 illustrates the class and individual overlaps within the context of different variants of the *Orpheus* myth. Notably, a distinct dissimilarity emerges between variant P_6 (Pausanias 9) and the remaining variants, regarding both class and individual overlap. This result is in accordance to information available about the sources: Pausanias alludes the narrative of *Orpheus and Eurydice* in his travel report without re-telling the myth in its entirety [181].[6]

When examining the similarity of ontologies at the individual level (IO), a notable highlight is the pairing of Plato and Hermesianax, demonstrating a substantial overlap score of 0.6. Similarly, a class overlap score of 0.65 is produced by comparing the domain ontologies for the hyleme sequences extracted from the Mythographus Vaticanus (approx. 875–1075 CE) and Apollodorus' Library (1./2. century CE).

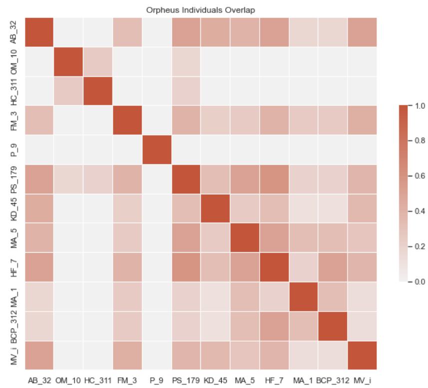
Figures 2 and 3 (depicted as 5.4c and 5.4d respectively) provide a focused visualization of optimal

[6] `http://www.perseus.tufts.edu/hopper/text?doc=Paus.+9.30.6&fromdoc=Perseus%3Atext%3A1999.01.0160`

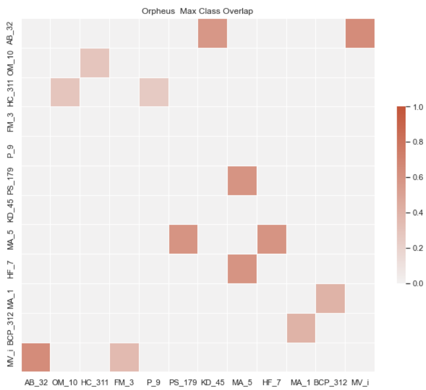
inter-variant matches, (excluding self-matches) along the diagonal of the matrix. Remarkably, the observed semblance in domain descriptions appears to transcend temporal and geographical attributes of the source materials, such as their origin in Roman or Greek cultural contexts. The similarity between Mythographus Vaticanus and Apollodorus' versions is interesting because it does not follow the results of Karsdorp and van den Bosch [124], who found a closer relation between a retelling and a their more recent predecessors of a story. This does not necessarily mean that one of the two results is wrong. Instead, one could argue that Karsdorp's corpus of Dutch retellings of the tale *Little Red Riding Hood* contains variants of a narrative that cover a smaller time frame, compared to variants of the myths concerning *Orpheus*, which spread across approximately one and a half millennia, and are thus more prone to loss and rediscovery. In this context, the success of a version (in that it becomes the source for subsequent variations) might be related among many other things to the robustness of distribution channels, e.g. how often a variant was written down and distributed along with other works of the Greek or Roman poet, as well as the popularity of the source at the time. We can assume that the more popular a version is, the more copies existed, the higher the chance for survival of that version.



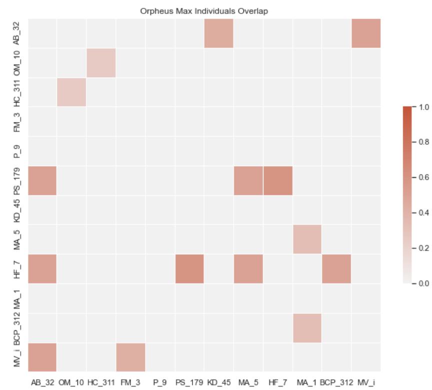
(a) Class overlap for variants



(b) Individual overlap for variants



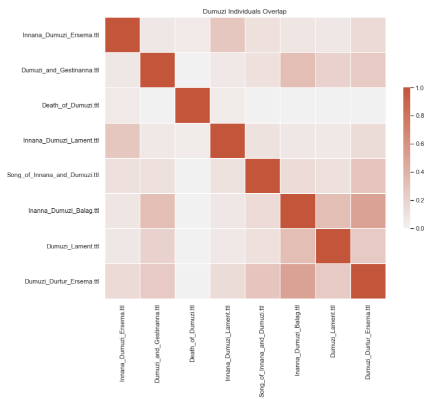
(c) Maximum class overlap



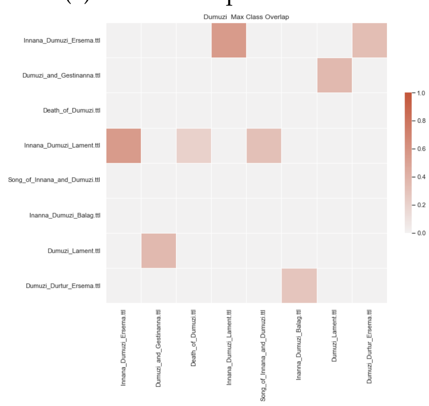
(d) Maximum individual overlap

Figure 5.4: Overlap for variants of *Orpheus* journey to the netherworld

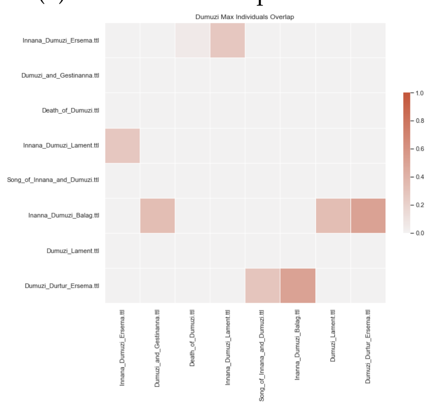
(a) Class overlap for variants



(b) Individual overlap for variants



(c) Class overlap for variants



(d) Individual overlap for variants

Figure 5.5: Overlap for variants of *Dumuzi*'s death

5.4 Automatic Ontology Construction

As indicated in Section 5.1, certain grammatical patterns can imply specific knowledge representations for the minimal ontologies. The patterns can be identified by investigating the dependency trees of the predicate argument structures (PAS). For that purpose, all noun chunks in a hyleme text are identified. For each noun chunk, the dependency of the *root* token of the noun chunk is identified. The resulting sequence of dependency tags, e.g. SB-PD (subject, predicative), is indicative for certain knowledge representations. The most common dependency abbreviations (TIGER scheme [11]) are given in Table 5.4. The most common patterns identified in the data are presented in Table 5.5. Together, they account for 699 hylemes in the set of German *durative-constant* hylemes, ca. 52 %.

Table 5.4: Important dependency tags and their meaning, derived from [11]

Abbreviation	Explanation
AG	Attribute, Genitive
MO	MOdifier
NK	Noun Kernel
DA	DAtive
OG	Genitive Object
OA	Accusative Object
OA2	Second Accusative Object
PD	PreDicative
SB	SuBject
SVP	Separable Verb Prefix

Examples for the patterns that have an associated knowledge representation pattern are given in Figures 5.6 to 5.11. Some patterns may include additional information important for the minimal ontologies. For instance, adjectives usually correspond to attributes, e.g. in Figure 5.9 “verwerflich” (engl. “reprehensible”) is an attribute to the ontology individual “secret”. In cases where a hyleme contains a direct and an indirect object, at least two relations can be inferred (relation(Ind1,Ind2), where Ind2 is the direct object and relationMod(Ind1,Ind3) where Ind3 is the indirect object).

Table 5.5: Most common dependency patterns and associated knowledge representation pattern (German hyleme data set)

Pattern	Associated Knowledge Representation	No. of Hylemes
SB-NK	*predMod(Ind1,Ind2)*	185
SB-PD	*Class(Ind1)*	154
SB-OA	*relation(Ind1, Ind2)*	125
SB	*attribute(Ind1)*	114
SB-PD-AG	*Class(Ind1)* and potentially *relation(Ind1, Ind2)*	68
SB-OA-NK	*relation(Ind1, Ind2)* and *relationMod(Ind1, Ind3)*	53

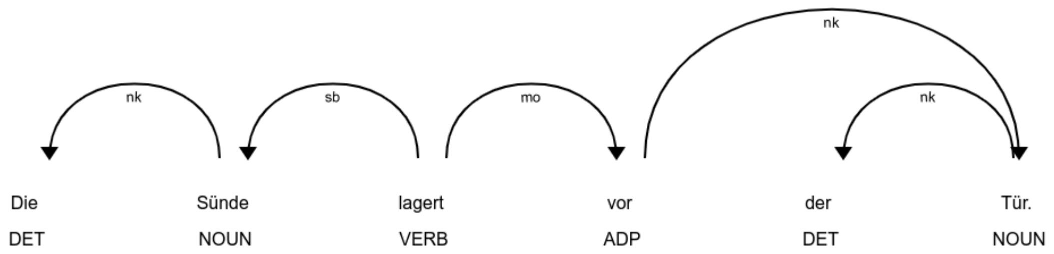


Figure 5.6: Pattern: SB-NK, Associated Knowledge Representation: *lagertVor(Sünde, Tür)*, Engl. "The sin lies in front of the door." (*liesInFront(Sin, Door)*

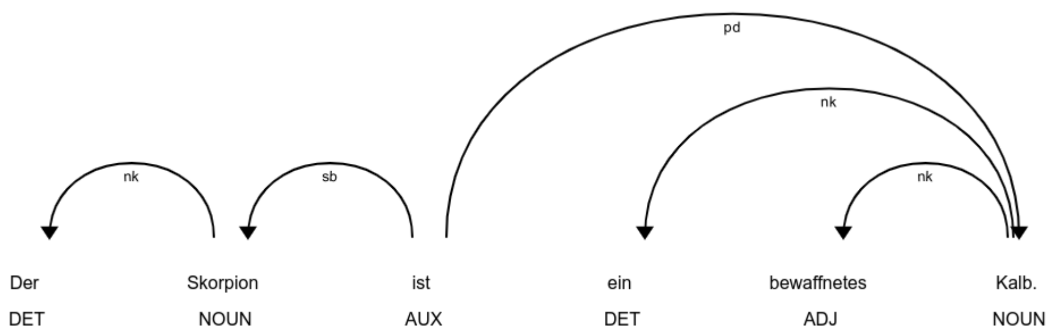


Figure 5.7: Pattern: SB-PD, Associated with Knowledge Representation: *Kalb(Skorpion)*, Engl. "The scorpion is an armed calf." (*Calf(scorpion)*)

However, not all *durative-constant* hylemes can be transformed into ontology classes, properties or individuals this way. For hylemes containing out of vocabulary words, dependency trees are often incorrect. Additionally, hylemes sometimes contain multiple relations, a pattern might not be associated to one distinct knowledge representation pattern, or the value of the knowledge representation can be context-sensitive. For instance, the statement "The scorpion is an armed calf" can imply either that there is a specific group of calves, which are armed, or that the scorpion is a calf which is coincidentally armed. In the former case, the class *calf* would have a subclass *armed calf*. In the latter case, the class assignment for the individual *scorpion* would be *calf* and the individual would have an attribute *armed*. To complicate matters further, in this particular case, the hyleme does not refer to a class assignment at all, but to a name variation of *Marduk*, from the

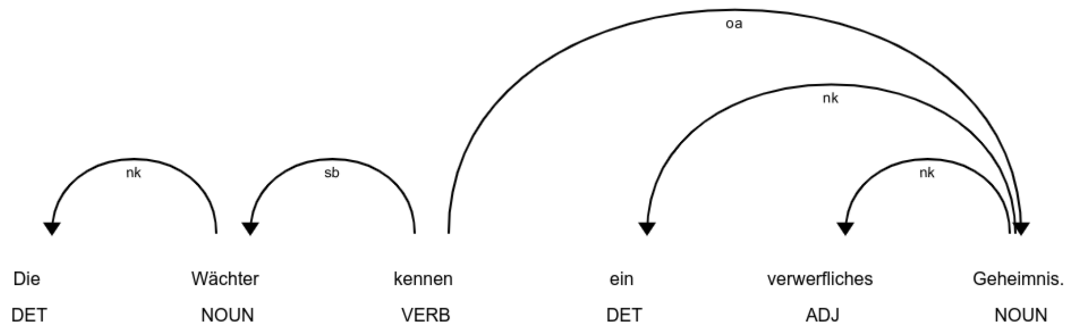


Figure 5.8: Pattern: SB-OA, Associated with Knowledge Representation: *kennen(Wächter, Geheimnis)*, Engl. "The guardians know a reprehensible secret." *know(Guardians, Secret)*

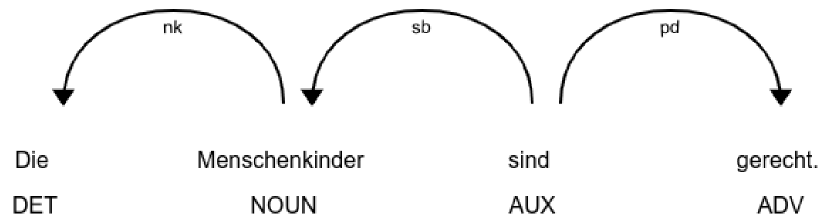


Figure 5.9: Pattern: SB-PD, Associated with Knowledge Representation: *gerecht(Menschenkinder)*, Engl. "The children of men are righteous." (*righteous(children of men)*)

Sumerian cuneiform sign AMAR[7] (engl. "calf").[8]

Domain and range values of relations and selectional preferences of classes are domain-sensitive and do not always coincide with real-life restrictions. For instance, as in the example "The scorpion is an armed calf.", an entity might belong to multiple subclasses of *Animal*. In the same manner, some relations can be the inverse of another relation (e.g. *isDaughterOf(Ind1,Ind2), isParentOf(Ind2,Ind1)*). That these restrictions may differ from myth to myth, and even between variants of the same myth, adds even more complexity to the domain-modelling task. This is an additional reason why a single-ontology approach would not be feasible.

Therefore, extractions of ontology classes and relations can be *assisted* with extraction rules based on common patterns, but they always have to be cross-referenced with domain-knowledge and manually corrected and completed, since domain ontology engineering is an inherently community

[7] `https://en.wiktionary.org/wiki/%F0%92%80%AB`

[8] I thank Prof. Christian Chiarcos for this helpful comment.

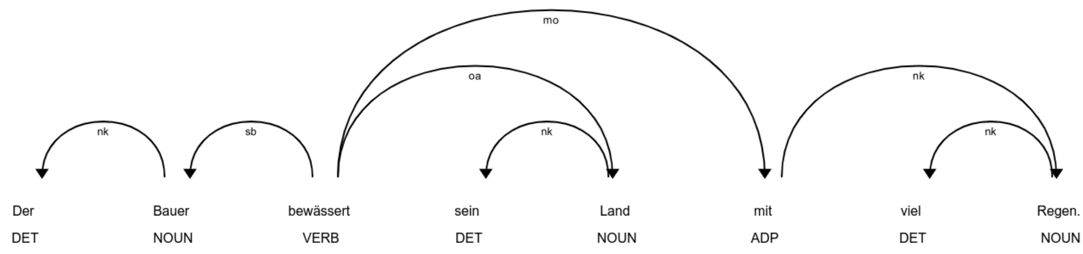


Figure 5.10: Pattern: SB-OA-NK, Associated with Knowledge Representation: *bewässert(Bauer, Land)*, *bewässertMit(Bauer,Regen)*, *WirdBewässertDurch(Land, Regen)*, engl. "The farmer waters his land with a lot of rain." (*waters(farmer, land)*, *watersWith(farmer, rain)*, *isWateredBy(farmer, land)*)

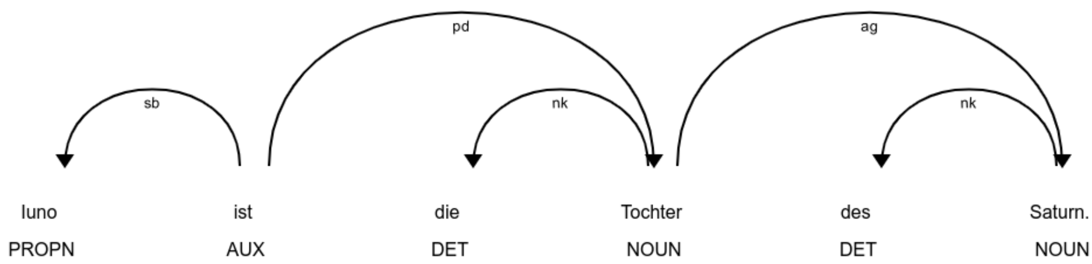


Figure 5.11: Pattern: SB-PD-AG, Associated with Knowledge Representation: *Tochter(Iuno)*, *istVaterVon(Saturn, Iuno)*, *istTochterVon(Iuno, Saturn)*, *Vater(Saturn)* , Engl. "Iuono is the daughter of Saturn." (*Daughter(Iuno)*, *isDaughterOf(Iuno, Saturn)*, *isFatherOf(Saturn, Iuno)*, *Father(Saturn)*)

driven process.

5.5 Discussion

From a narratological standpoint, it is sensible to treat *durative* and *single-event* hylemes differently, because the hyleme types represent different aspects of a narrative, i.e. the background information and the plot. Minimal ontologies are a suitable tool to ensure that variations in the available background information is processed and factored in the similarity estimation of myth variants. However, since the extraction of hyleme sequences follows a semi-structured approach, the development of a controlled vocabulary may be needed if hylemes describing the background information, states and habituals, are too different.[9]

A major drawback of this approach is that the manual construction of the minimal ontologies is

[9] The construction of a controlled vocabulary for the representation of the Greek ferryman of the underworld Charon as a foundation for modelling background information and to facilitate the construction of related hyleme sequences is currently under-way as part of a Master's thesis related to this project.

time-consuming. Even when the extraction process can be semi-automated by following extraction rules, many tasks, such as the linking of entities to open knowledge graphs, the interpretation of domain-specific edge cases, and the modelling of domain and range properties to remain a manual task undertaken by an informed scholar. Additionally, this process obviously only applies to sequences that contain *durative* hylemes. For hyleme sequences which consist only of *single-event* hylemes, the background information can not be modelled in this manner. However, since the hyleme sequences are derived by domain experts, one could argue that if the hyleme sequence does not contain this kind of information, it is most likely not present in the source material (e.g. no statement from which the truth value of *Dumuzi is a shepherd* can be derived).

5.6 Summary

In this chapter, we present a knowledge engineering approach funded on the hylistic category of *durative-constant* hylemes. Carefully, manually crafted shallow domain ontologies allow insights into inter- and intra-myth comparison, based on ontology classes and individuals. Using those shallow or minimal domain ontologies, similarities and differences between myth variants can be investigated. In this chapter, two sets of OWL2 domain ontologies are presented. The first set presents domain information on myths pertaining to *The Death of Dumuzi*. The second set contains ontologies on myth variants of *Orpheus and Eurydice*, for which a dedicated controlled vocabulary was developed. Ontology comparison can be quantified using simple overlap measures, namely *individual overlap* and *class overlap*. This chapter also gives an insight into the grammatical structures of *durative-constant* hylemes. Certain grammatical patterns can be exploited as a basis for deriving knowledge representation of background information, but eventually the task depends on informed scholars and domain experts.

The content of this chapter has been presented as posters at the First Meeting of the Digital Ancient Near Eastern Studies, DANES 2023, 19–21 February 2023, Israel (*Dumuzi*) and the Fourth Biennial Conference on Language, Data and Knowledge (LDK 2023), Vienna, 12–13 September 2023 (*Orpheus*)[10].

[10]This part of the project was undertaken with assistance of a master student of Digital Humanities who holds a doctorate degree in ANES and a student assistant with a background in Classical Studies (CS).

Chapter 6

Modelling Event and Plot Structure

"His Eurydice does not complain, since her only cause of complaint would be that she is loved too much."

C.M. Bowra, *Orpheus and Eurydice* [175]

The background knowledge and states of narrative texts have been annotated and evaluated in Section 4.5, which is an important step in the alignment process. Since events can now be separated from background information, this chapter aims to demonstrate approaches on modelling events and plot structure. In this work, the term *event* describes a voluntary or involuntary action, reaction, experience, perception, or feeling of an animate or inanimate agent. The start and end of an event takes place within the scope of a narrative variant, e.g. a myth variant. In hylistic terms, events are communicated by *single-event* hylemes.

This chapter provides a case study on the myth of *Orpheus and Eurydice*. It aims to illustrate how an ideal alignment of hylemes in hyleme sequences can be achieved in a perfect scenario, i.e. many variants describing the same *Stoff* and information regarding higher level hylemes, so called *hyper-hylemes.* For the case study, the mythical *Stoff* is modelled as a set of sequences that can be described with a regular grammar that was explicitly derived from the relevant hyleme sequences in the German hyleme data set. The original sequences were derived by domain experts in the field of Classics.

In Section 6.1, after an introduction of the myth and the related data in the German data set, the regular grammar is introduced and tested against two new variants. The sequences that can be produced using the grammar are subsequently compared using sequence algorithms in Section 6.2. Section 6.3 proposes an optimal alignment for the sequences produced from the grammar. Two sequence alignment approaches are applied to the resulting sequences: The Needleman-Wunsch algorithm produces a global alignment, which can be compared to the gold standard of the optimal alignment. The Smith-Waterman algorithm identifies patterns in the sequences, which can yield interesting results regarding the re-use of sequence components. The chapter ends with a summary in Section 6.4.

6.1 The myth of *Orpheus and Eurydice*

For this case study, one myth is selected from the hyleme data, the *Stoff* of *Orpheus and Eurydice*. This myth has been adapted, remixed, modernized and re-used countless times in ancient sources and modern variants. Marlow calls it "one of the immortal stories, with love, music, death and tragedy interwoven in it" [182, p. 361]. Its ongoing popularity makes it the ideal example to study the event modelling and alignment techniques presented in this chapter. Furthermore, while the variants have core concepts and events in common which make them part of the same *Stoff*, they show interesting divergences that are worth studying with the methods used in this work. For instance, we might want to investigate Orpheus' failure. Is it present in a variant? And if so, how does it manifest? Does he lose his wife, because he disobeys an instruction to not turn around? Or does he fail because he was set up to fail from the beginning?

The study of this myth and comparatistic studies pertaining it have a long tradition. For instance, Bowra attempts to reconstruct an early Greek poem by comparing versions of Ovid's *Metamorphoses*, 10, 1-73, Pseudo-Vergil, *Georgics* 4. 454, *Culex* and others [175]. Most of these variants are included in the hyleme data.

In the hyleme data, the myth of *Orpheus and Eurydice* is represented by no less than 18 different variants, mainly from Greco-Roman sources. The sources corresponding to the variants and their sequence IDs are reported in Table 6.1. The sequences range in length from one hyleme to 27 hylemes. To model the event structure in the sequences, we filter out all *durative* hylemes and consider only *single-event* hylemes. This results in two empty sequences, and two single hyleme sequences, i.e. sequences that originally contain no or only one *single-event* hyleme and one or more *durative* hylemes..

In addition to the relatively large number of sequences describing the myth of *Orpheus and Eurydice*, it is also suitable for a case study on hyleme alignment because the story can be broken down into broader, recognisable elements which appear across sequences. In other words, we know that a certain set of plot elements needs to be present for us to consider a sequence a variant of the myth. In this regard, the hylistic analysis of the present example ties seamlessly with other narrative theories, such as Proppian analyses, as introduced in Chapter 2. Not all of these plot elements have to be present in every variant. Instead, we can think of them all as parts of the *Stoff*-cloud. In the following section, the core plot elements are presented, and their interplay is modelled as a (right) regular grammar.

Table 6.1: Sources in the data set used for sequence modelling and comparison

ID	Abbreviation	Source
1	HF_7	Hermesianax Fragment 7
7	SHF	Seneca, Hercules furens
15	OM_10	Ovid Metamorphosis, 10
34	FM_3	Fulgentius Mythologiae, 3
40	MA_5	(Marcus) Manilius, Astronomica, 5
54	IB_11	Isocrates Busiris, 11
55	SHO	Seneca Hercules oetaeus
57	MV_i	Mythographus Vaticanus, i
86	VC	Pseudo-Vergil Culex, 68-295a
90	AB_32	Apollodorus Library 3,2
107	KD_45	Konon Diegeseis, 45
118	HC_311	Horace Carmina, 3,11
140	D_4254	Diodorus, 4,25,4
152	P_9	Pausanias, 9
169	MA_1	(Marcus) Manilius, Astronomica, 1
172	VG_4	Vergil Georgics 4, 453–527
176	BCP_312	Boethius Consolatio Philosophiae, 3,12
177	PS_179	Plato Symposium, 179d

6.1.1 Modelling Plot Variants using Regular Grammar

We can assign hylemes to core plot elements, like *Orpheus'* descend, or the breaking of the taboo, that re-appear in multiple sequences. However, they might be realised in different hylemes that would not be comparable as such. For instance, the taboo is communicated in the data in various forms, e.g. “Orpheus nimmt die Bedingungen an.” (engl. “Orpheus accepts the conditions.”), “Orpheus bekommt Anweisungen bezüglich Eurydike.” (engl. “Orpheus receives instructions concerning Eurydice.”), or “Orpheus darf sich auf dem Weg bis zur Ankunft in seinem Haus nicht umdrehen” (engl. “Orpheus is not allowed to turn around on the way until he reaches his house”). A direct semantic comparison of these hylemes would not necessarily yield a match. While some sequences use linguistic variations to express the same event, plot elements can also be represented by different amounts of hylemes.

Furthermore, in some cases the predicate argument structures (PAS) are only similar in the context of a *Stoff*. When a (lexically) identical hyleme appears in a completely different myth, it might not convey the same meaning in the broader context. For instance, “Orpheus accepts the conditions.” could also refer to him having to leave Eurydice behind at the end of some of the variants. In that case, it cannot be aligned with any of the taboo-conveying hylemes. In this example, we see different aspects of semantic similarity that needs to be taken into consideration when hyleme matching is performed. Some hylemes are similar in the sense of lexical variation, for instance paraphrases or slight differences “Orpheus accepts the conditions.”-“Orpheus accepts the conditions for the ascend.” In other cases, hylemes are not similar but convey the same narrative effect. Sometimes one hyleme entails the other one, e.g. "Orpheus receives instructions."-"Orpheus is not allowed to turn around." Thirdly, some hyleme matches require significant insight into the *Stoff*, the sources, including possible variations in the translations, e.g. "The fate goddesses replace Eurydice's used up (life) threads."-"Eurydice comes back to life."

This challenge can be further illustrated if we group different realisations of hylemes to each other by representing hylemes by the GermanNet word category (see Section 2.3) of the hyleme predicate lemma. Some examples of different hylemes and corresponding word categories are given in Table 6.2. In Table 6.6, sequences 7 and 55 are compared, including GermaNet word categories of the hyleme predicates.

Table 6.2: Example Word Categories of Predicate Lemmas, transl. by the author

Hyleme	Hyleme Predicate Lemma	GermanNet Synset ID	Word Category
Orpheus wappnet sich mit seiner Kithara. *Orpheus prepares himself with his kithara.*	wappnen	s54812	Kognition *cognition*
Orpheus segelt zu Charon. *Orpheus sails to Charon.*	segeln	s57989	Lokation *location*
Orpheus spielt auf seiner Kithara. *Orpheus plays his kithara.*	spielen	s53565	Gesellschaft *society*
Orpheus gewinnt verschiedene Götter für sich. *Orpheus wins over several gods.*	gewinnen	s56160	Kommunikation *communication*
Agriope bekommt zarten Lebensatem. *Agriope receives a tender breath of life.*	bekommen	s52558	Allgemein *general*

GermaNet word categories are relatively broad, which should simplify the task to group actions communicated by hyleme predicate together. However, matching hylemes by the GermaNet word class associated with the hyleme predicate (verb) lemma would be a challenging task to automate. Firstly, the disambiguation of the corresponding synsets can be difficult. Secondly, similar events can be described using verbs that belong to different word classes, as seen in the hylemes "Orpheus singt." (engl. "Orpheus sings.") which belongs to the word class Communication and "Orpheus schlägt die Lyra an" (engl. "Orpheus strikes his lyra."), which belongs to the word class Gesellschaft (Society). For the purpose of achieving an ideal event model and subsequently an optimal alignment, this chapter therefore follows an approach based on a manual modelling of *hyper-hylemes*.

In order to achieve a gold standard for the case study, we group hylemes together into hyper-hylemes, which are core plot components reappearing across sequences. The hyper-hylemes were derived manually by inspecting the candidate sequences. Each hyper-hyleme can be represented by one or multiple hylemes and appears in one or multiple sequences. By applying them, we match events that are conveyed in single hylemes, e.g. "umschauen" (engl. 'to look around', 'to look behind oneself') with variants that are communicated over multiple hylemes, e.g. "umdrehen' (engl. 'to turn around')'+"anschauen" (engl. 'to look at sb').

The hyper-hylemes are represented below by smaller case letters, which are used to construct the sequences in the regular grammar. They are the the core items on which the alignment is achieved. Not every myth variant includes all of the hyper-hylemes presented in this work, but all hyper-hylemes occur more than once. The number of occurrences is given in square brackets. Parantheses indicate variations of the content.

Similarly to the Proppian functions, the hyper-hylemes appear in sequential order, with the exception of z which can appear either before or after b. We define the vocabulary Σ as:

1. v: Aristeus chases Eurydice (in order to rape her.) [2]
2. t: (Following a snake bite,) Eurydice dies. (*first loss*) [6]
3. e: Orpheus decides to enter the netherworld and/or prepares for the descent. [7]
4. r: Orpheus descends to the netherworld. [14]
5. u: (Using his beautiful music,) Orpheus convinces the inhabitants of the netherworld to release Eurydice. [11]
6. z: Orpheus receives Eurydice. Eurydice comes towards Orpheus. [4]
7. b: The inhabitants of the netherworld set conditions for the ascent of Orpheus and Eurydice. (Namely, they do not allow him to look at her until they reach the surface.) [9]
8. h: Orpheus and Eurydice begin their ascent. [9]
9. d: Orpheus neglects the conditions (by turning around and looking at Eurydice). [11]
10. a: Orpheus loses Eurydice again. (*second loss*) [9]
11. f: The result of Orpheus' renewed loss, e.g. Orpheus' despair. [5]
12. m: (The Thrakian women) kill Orpheus. [3]

A word cloud visualisation of the textual representation of each of the hyper-hylemes is shown in Appendix A.2. Using those hyper-hylemes, a regular grammar can be constructed for all variants of the myth of Orpheus and Eurydice that were studied here.

The grammar consists of a number of production rules, presented in Table 6.3, which can be used to create all 18 sequences from the vocabulary (terminals) of hyper-hylemes as presented above. ε is used to produce the empty sequences. We write the grammar as an extended right-regular grammar for readability and interpretability. However, the grammar can be transformed into a strict right-regular grammar by introducing new non-terminals. Furthermore, we can represent the rules of the extended right-regular grammar in a deterministic finite automaton (DFA), as shown in Figure A.5. $\Sigma*$ includes all sequences that can be produced from Σ using these production rules, including the empty sequence ε with $|\varepsilon| = 0$. The automaton can be used to automatically check the validity of the production rules against the hyper-hyleme sequences (i.e. the words in the language L over $\Sigma*$). The DFA is *non-empty*, i.e. it accepts any language other than the empty language, *non-universal*, i.e. the input language of the DFA is different from $\Sigma*$ and therefore *interesting* (i.e. non-empty and non-universal).

Table 6.3: Production Rules of the Regular Grammar for the Orpheus and Eurydice myth

Start	Non-Terminals	Terminals
Seq → V	V → vT	H → h
Seq → T	T → tE	U → u
Seq → E	T → tR	D → d
Seq → R	E → eR	A → a
Seq → U	R → rU	F → f
Seq → ε	R → rZ	M → m
	R → rB	R → r
	R → rH	
	R → rP	
	H → hD	
	U → uH	
	U → uB	
	U → uP	
	U → uQ	
	B → bH	
	B → bD	
	P → zbH	
	Q → bzD	
	Z → zF	
	D → dA	
	D → dM	
	A → aF	
	F → fM	

Based on this grammar, it can be determined that the hyper-hylemes have different functions, depending on where in the sequence they appear. For instance, 13 out of 18 sequences start with a hyper-hyleme from the set $PREP = \{v, t, e, r\}$. Those preparatory functions include Aristeus chasing Eurydice/Agriope[1], Eurydice's death, Orpheus' decision for counter-action and

[1] Hermesianax, Fragment 7,1–14, alternative name for Eurydice, presented by Hermesianax and others e.g.

the subsequent descend to the netherworld. The sequences that do not start with one of those hyper-hylemes are the single-hyleme sequences, the empty sequences and sequence 186, derived from Pseudo-Vergil *Culex* 68-295a.

The core of the story is communicated by one or multiple of the hyper-hylemes $CORE = \{u, z, b, h, d\}$. The end sequences communicate the result of Orpheus failure to rescue Eurydice. We define them as $RES = \{a, f, m\}$. Out of all sequences, eleven end with one of these hyper-hylemes. However, it is worth taking a closer look at the sequences which end with one of the other hyper-hylemes.

The first sequence that diverges from the general pattern was derived from Hermesianax *Fragment* 7, 1-14 (Sequence 1). Its hylemes are listed in Table A.1. In this variant Orpheus manages to rescue Agriope (Eurydice) from Hades without mention of any failure. He successfully brings her back to life. Therefore, the sequence ends with hyper-hyleme h. Secondly, Sequence 140, derived from Diodorus (4,25,4)[2] contains two hyper-hylemes ru, realised in three hylemes. It ends with Persephone allowing Orpheus to take Eurydice back with him, but no mention that he actually does so or the manner in which it is achieved.[3]

Both sequences are from the pre-Vergilian tradition of the *Orpheus and Eurydice-Stoff*, they tell a variant that ends with the successful rescue of Eurydice. The *single-hyleme* sequences (SEQ118 and SEQ169), which contain the hyper-hylemes u and r also allude to the successful rescue.

The third sequence that differs from the common end pattern consists of nine hylemes, of which the final hyleme is *durativ-resultative* ("Euridyke bleibt zurück.", engl. "Eurydice remains behind."). It was derived from Pseudo-Vergil *Culex* 68-295a. The sequence consists of three hyper-hylemes, ubd. The sequence starts with Orpheus already in the netherworld, omitting the preparatory hyper-hylemes. It begins with hylemes describing Orpheus singing in the netherworld, his successful convincing of Dis' (Hades) wife, and her subsequent conditions for Eurydice's ascend. In wanting to kiss his wife, Orpheus neglects those conditions and looks back. The sequence does, however, not contain any mention of his descend to the netherworld, or Orpheus beginning any form of ascend with Eurydice. Therefore, the result of Orpheus' failure to bring his wife back to life is her staying in the netherworld (conveyed in a *durative* hyleme) instead of an action or event, e.g. her fading or sliding back.

One sequence is substantially different to the others, although it contains the hyper-hylemes from the $PREP$ set at the beginning and the RES set at the end. This sequence from Plato, *symposium* 179d contains no hyper-hylemes from the $CORE$ set, because Orpheus does not convince the

Athenaeus, Deipnosophists, 13.71 `https://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2013.01.0003%3Abook%3D13%3Achapter%3D71`

[2]`http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2008.01.0540%3Abook%3D4%3Achapter%3D25%3Asection%3D4`

[3]Although the source does not mention Orpheus's success beyond convincing Persephone, it might be argued that the successful rescue is implied, because there is no mention of conditions for the ascend, or other obstacles which could lead to Orpheus' failure. Following this argument, an implicit hyleme could be added which would complete the sequence with the positive result similar to the first diverging example.

inhabitants of the netherworld, who only show him an illusion of his wife, effectively setting him up for failure. The English translation of the source is given here:

> (1) "But Orpheus, son of Oeagrus, they sent back with failure from Hades, showing him only a wraith of the woman for whom he came; her real self they would not bestow, for he was accounted to have gone upon a coward's quest, too like the minstrel that he was, and to have lacked the spirit to die as Alcestis did for the sake of love, when he contrived the means of entering Hades alive. Wherefore they laid upon him the penalty he deserved, and caused him to meet his death."[4]

Marlow [182] argues that this includes the possibility that Orpheus could have saved his wife if he had followed Alcestis' example and died in his wife's place. This aspect is not present in the hyleme sequence that was derived for this source. Arguably, a possibility of this sort would be conveyed by a German Konjunktiv II construction, which is not in accordance to the hyleme extraction standards. Additionally, hyleme sequences are mainly plot-driven, they communicate events, which possible alternative outcomes like 'had he been as brave as Alcestis' is not.

From these divergences it becomes clear that the myth by no means always ends in failure and twice-lost love. Instead, the functions in the RES set indicate that both endings are valid elements of the *Stoff* from which we derive the myth of *Orpheus and Eurydice*.[5] Bowra [175] discusses versions of the myth where Orpheus is allowed to take Eurydice with him for a short time only. There is no version that contains that restriction in the hyleme data as of date. However, the condition can be subsumed under hyper-hyleme b, "The inhabitants of the netherworld set conditions for the ascent of Orpheus and Eurydice.".

Describing a New Variant with the Regular Grammar

The regular grammar has to be investigated on how well it can be applied to sequences that were not part of the orignal data set of 18 variants of the *Orpheus and Eurydice* myth. Firstly, we test it against a myth variant communicated in Plutarch, Amatorius 17. The variant is relatively short, containing only five hylemes, of which four are *durative-constant*. The last hyleme is classified as single-event and only alludes to the myth without re-telling the plot:

"Hades zeigt Alkestis, Protesilaos, Eurydike und Orpheus gegenüber Gnade."
(engl. "Hades shows mercy towards Alcestis, Protesilaos, Eurydice and Orpheus.")

[4] Plato Symp. 179d `http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0174:text=Sym.:section=179d&highlight=Orpheus`

[5] Additionally, the empty Sequence 54, derived from Isokrates *Busiris* 11, contains one *durative-constant* hyleme: "Orpheus führt die Toten zurück aus dem Hades." (engl. "Orpheus leads the dead back from Hades."). Despite describing an action, this hyleme is annotated as *durative-constant*, because the verb in the original Greek is in imperfect, indicating not a habit but the power and aptitude to do so. However, it also indicates success, i.e. if he is someone who has the ability to bring people back from Hades, he would have certainly demonstrated it by bringing back his wife. [175] follows the same argument. Therefore, sequence no. 54 is in accordance with the other sequences that communicate the success of bringing back Eurydice, although it does not appear in the aligned data.

Unfortunately, this hyleme is so abstract that the closest equivalent hyper-hyleme is u (Orpheus convinces the inhabitants of the netherworld.) $Seq \to U$, $U \to u$ are valid production rules. The nature of the variant, it only hinting at the myth of *Orpheus and Eurydice* may be the reason it was not included in the hyleme data set in the first place.

As a second variant, we apply the regular grammar on a variant of the story that was produced by the chat bot ChatGPT[6]. We select this tool because it can produce coherent, high-quality natural language text, which is constructed using a (essentially) statistical representation of the large training data using a pre-trained transformer (GPT) approach. Therefore, similarity between the produced sequence and the data used for constructing the grammar might indicate which version is most present in the training data. When prompted for the myth of *Orpheus and Eurydice*, the chat bot artificial intelligence agent produced the following text. We indicate hyleme-relevant information in square brackets.

> [Orpheus was a legendary musician and poet]$_1$, [gifted with a voice that could charm anyone who heard it]$_2$. [He fell deeply in love with a woman named Eurydice]$_3$, [a beautiful nymph]$_4$. [They were married and deeply in love]$_5$. However, their happiness was short-lived. One day, [while out walking in the woods]$_6$, [Eurydice was bitten by a snake]$_7$ and [died from the venom]$_8$. [Orpheus was inconsolable with grief]$_9$ and [decided to descend into the Underworld to try and bring her back]$_{10}$. [Using his music, he was able to charm the ferryman Charon and cross the River Styx]$_{11}$. [He then stood before Hades, [the god of the Underworld]$_{13}$, and [his queen Persephone]$_{14}$]$_{12}$, and [begged for Eurydice's release.]$_{15}$ [Moved by Orpheus's music and his love for Eurydice]$_{16}$, [they agreed to let her go with him]$_{17}$, [but with one condition:]$_{18}$ [Orpheus must not look back at Eurydice until they had both left the Underworld.]$_{19}$ [Orpheus agreed]$_{20}$ and [began the journey back to the world above,]$_{21}$ with [Eurydice following behind him.]$_{22}$ As they walked, [he could hear her footsteps behind him]$_{23}$, but [resisted the urge to look back]$_{24}$ until [they were just a few steps away from the exit]$_{25}$. [Unable to bear the suspense any longer]$_{26}$, [Orpheus turned around]$_{27}$, but unfortunately, [Eurydice had not yet crossed the threshold back into the world of the living]$_{28}$, and [[she was pulled back into the Underworld]$_{29}$ forever]$_{30}$. [Orpheus was devastated by his mistake]$_{31}$ and [[[spent the rest of his life in mourning]$_{32}$, wandering the earth]$_{33}$ and singing sad songs of lost love]$_{34}$. It is said that [he was eventually killed by the followers of the god Dionysus]$_{37}$, [who were enraged by [his rejection]$_{36}$ of their worship]$_{35}$.

The single-event hylemes that can be derived from the text, and their corresponding hyper-hylemes are shown in Table 6.4. Hyleme 35 refers to a different variant of the narrative material, and hence is not part of the vocabulary of the formal grammar. We report durative hylemes separately, in Table 6.5.

[6] `https://chat.openai.com/`

Table 6.4: Single-event hylemes and corresponding hyper-hylemes derived from chatGPT version of the Orpheus and Eurydice myth

	Hyleme	Hyper-Hyleme
7	A snake bites Eurydice.	t
8	Eurydice dies from the venom.	t
9	Orpheus is grieving.	e
10	Orpheus decides to descend into the netherworld.	e
11	Orpheus convinces Charon to let him cross the River Styx.	e
12	Orpheus stands before Hades and Persephone.	r
15	Orpheus begs for Eurydice's release.	u
16	Orpheus' moves Hades and Persephone with his love for his wife and his music.	u
17	Hades and Persephone agree to let Eurydice go.	u
18	Hades and Persephone set a condition for Eurydice's acent.	b
19	Orpheus is not allowed to look at Eurydice until they reach the surface.	b
20	Orpheus agrees to the condition.	b
21	Orpheus begins the ascent.	h
22	Eurydice follows Orpheus.	h
23	Orpheus hears Eurydices footsteps.	h
24	Orpheus resists the urge to look at Eurydice.	h
25	Orpheus and Eurydice are a few steps away from the surface.	h
26	Orpheus cannot bear the suspense any longer.	d
27	Orpheus turns around.	d
28	Eurydice has not reached the surface yet.	d
29	Eurydice returns back to the netherworld.	a
31	Orpheus is devastated by his mistake.	f
35	Orpheus rejects the worship of Dionysus.	-
37	The followers of Dionysus kill Orpheus.	m

From the hyper-hylemes, a sequence for the ChatGPT variant can be produced with the regular grammar using the following production rules. Seq $\rightarrow T$, $T \rightarrow tE$, $E \rightarrow eR$, $R \rightarrow rU$, $U \rightarrow uB$, $B \rightarrow bH$, $h \rightarrow hD$, $D \rightarrow dA$, $A \rightarrow tF$, $F \rightarrow fM$, $M \rightarrow m$. The resulting sequence is $terubhdafm$. It is very similar to the sequence for sequence 172, derived from Vergil *Georgics*, 4 ($terzbhdafm$). The text produced by ChatGPT can be interpreted as a combination of components from various sources. GPT-3 on which ChatGPT is based is trained on a large amount of data from the internet. The similarity to Vergil's version might be due to its popularity across the training data. Hence, it is not surprising that the regular grammar works well on the resulting sequence. It would most likely not contain new information, only a combination of existing components of the myth. However, this would technically be true for some of the antique versions as well.

Table 6.5: Durative hylemes derived from chatGPT version of the Orpheus and Eurydice myth

	Hyleme	Hyleme-Type
1	Orpheus is a legendary musician and poet.	durative-constant
2	Orpheus charms everyone with his voice.	durative-constant
3	Orpheus loves Eurydice.	durative-constant
4	Eurydice is a nymph.	durative-constant
5	Orpheus and Eurydice are married.	durative-constant
6	Eurydice is walking in the woods.	durative-initial
13	Hades is the god of the netherworld.	durative-constant
14	Persephone is Hades' queen.	durative-constant
30	Eurydice has to remain in the netherworld forever.	durative-resultative
32	Orpheus spends the rest of his life mourning.	durative-resultative
33	Orpheus spends the rest of his life wandering.	durative-resultative
34	Orpheus spends the rest of his life singing sad songs.	durative-resultative
36	The followers of Dionysus are enraged by Orpheus rejection of their worship.	durative-resultative

6.2 Sequence Similarity

Before this chapter moves on to sequence alignment, we want to investigate sequence similarity based on the sequences derived from the regular grammar. In order to find out which sequences are similar, different methods can be applied. All the methods work directly on the sequences, using the alphabet Σ for comparison. If two sequences contains a hyper-hyleme represented by a literal, the literal will be matched, even if the hyleme representation is different.

6.2.1 Longest Common Subsequence and Substring

We can apply two measures to compute the similarity of strings with regard to the how many elements in the sequences match, see Section 2.3. Those measures tell us which hyper-hylemes appear in both sequences.

The first similarity measure is *longest common subsequence*, which computes how many matching elements are in the respective sequences. LCS neglects ordering of elements and includes all matching elements, regardless of whether they are connected or not.

The second measure we can apply for the sequence similarity is *Longest Consecutive Substring* (LCStr). This measure calculates how many consecutive elements appear in both strings. For instance, the longest consecutive substring of the sequences $vtrbhda$ (Seq34) and $vterubzda$ (Seq57) is vt resp. da with a length of 2.

Both of these measures have the disadvantage that they favour longer sequences. Therefore, we

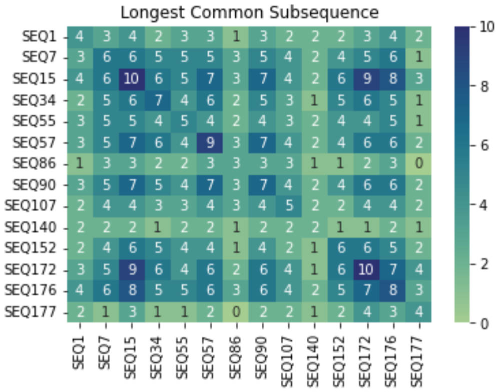


Figure 6.1: Longest Common Subsequence, $min = 0$, $max = 9$

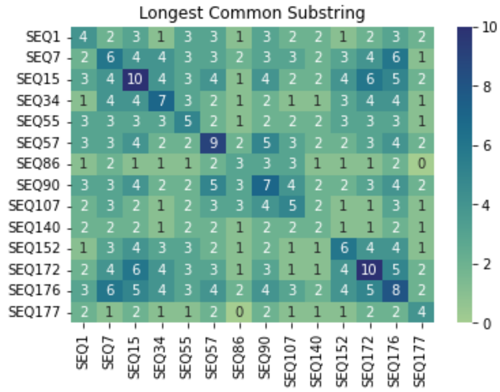


Figure 6.2: Longest Common Substring, $min = 1$, $max = 6$

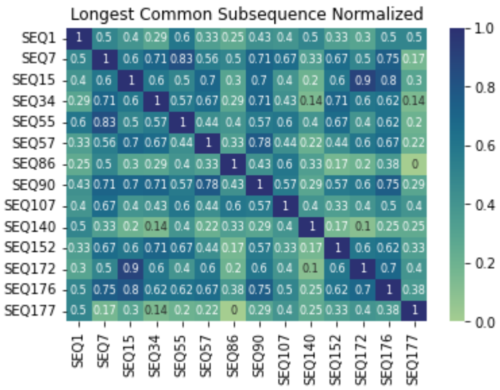


Figure 6.3: Longest Common Subsequence normalized, $min = 0$, $max = 0.9$

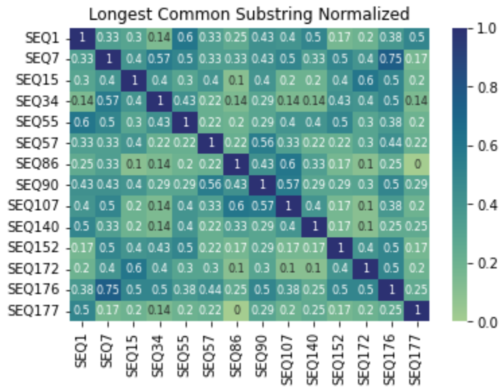


Figure 6.4: Longest Common Substring Normalized, $min = 0$, $max = 0.75$

normalize LCS and LCStr of two sequences L_1 and L_2.

$$LCS_{norm}(L_1, L_2) = \frac{LCS}{L}, \tag{6.1}$$

where $L = L_1$, if $|L_1| > |L_2|$, else $L = |L_2|$

LCS, LCStr and their normalized variants are reported in Figure **??**. We can see that the values on the diagonal of the similarity matrices show values of 1 for the normalized measures, whereas the diagonals on the matrices for the non-normalized measures show the length of the sequences. The highest similarity values occur for the pairs Seq7-176 (LCStr) and Seq15-172 (LCS), for the normalized and non-normalized measures.

6.2.2 Levensthein Distance

Levenshtein Distance, also known as Edit Distance, is a measure of the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string of characters into another. We report the Levensthein Distance for the sequences that can be produced using the regular grammar in Figure 6.5. The minimal Levensthein distance occurs between sequences 7 (SHF) and 55 (SHO), both by Seneca. Sequences 140 and 172, D_4254 and VG_4, have the highest Levensthein Distance in the data set.

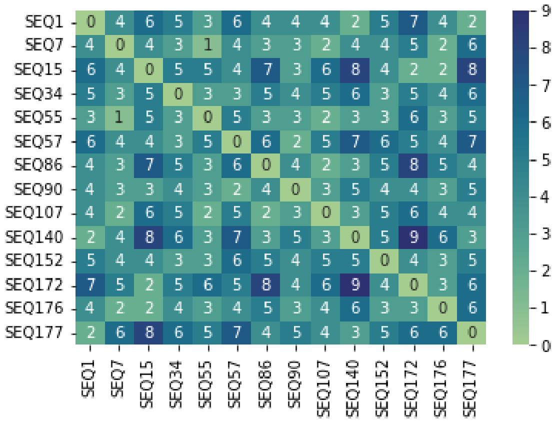


Figure 6.5: Levensthein Distance Heatmap, $min = 1$ (Seq7-Seq55), $max = 9$ (Seq140-Seq172)

For the sequences 7 and 55, which have the closest Levensthein distance, can see that both sequences start with r, realised in one hyleme. In both cases, the hyper-hyleme u is realised in four consecutive hylemes. Sequence 55 does not have an explicit mention of the conditions for Eurydice's rescue, hence b is missing in the sequence. We see that the attempted ascend d is realised in three hylemes in sequence 55, and in four hylemes in sequence 7. Both sequences end in the same hyleme "Orpheus verliert Eurydike" (engl. "Orpheus loses Eurydice").

We report the hylemes, the GermaNet word classes and the hyper-hylemes of sequences 7 and 55 in Table 6.6. The sequences that can be generated with the regular grammar are $rubhda$ and $ruhda$, resulting in a edit distance (Levensthein distance) of 1 (one insertion). This example demonstrates why automatic alignment on hylemes might be challenging. For instance, "betritt die Unterwelt durch die taenarische Pforte" (engl. "enters the netherworld through the taenaric gate") and "steigt in die Unterwelt hinab" (engl. "descends into the netherworld") communicate the same event, but are too semantically dissimilar to be matched automatically with reasonable confidence.

Table 6.6: Hylemes and Hyper-Hylemes for the sequence pair 7-55

	Sequence 55			Sequence 7		
	Hyleme	Word Category (Verb)	Hyper-Hyleme	Hyleme	Word Category (Verb)	Hyper-Hyleme
1	Orpheus betritt die Unterwelt durch die taenarische Pforte. *Orpheus enters the netherworld through the taenaric gate.*	Lokation	r	(Orpheus steigt in die Unterwelt hinab.) *(Orpheus descends into the netherworld.)*	Lokation	r
2	Orpheus schlägt die Lyra an. *Orpheus strikes the lyre.*	Gesellschaft	u	Orpheus singt. *Orpheus sings.*	Kommunikation	u
3	Orpheus gewinnt die Unterirdischen mit rührendem Gesang (für sich). *Orpheus wins over the inhabitants of the netherworld with touching song.*	Kommunikation	u	Orpheus (rührt und) besänftigt die Unterirdischen. *Orpheus touches and appeases the inhabitants of the netherworld.*	Gefuehl	u
4	(Schicksals-)Göttinnen ersetzen Eurydikes aufgebrauchte (Lebens-)Fäden. *(Fate) goddesses replace Eurydice's used up (life) threads.*	Veraenderung	u	Orpheus fordert seine (geliebte) Eurydike zurück. *Orpheus demands his (beloved) Eurydice back.*	Kommunikation	u
5	Orpheus erhält Eurydike als Belohnung für seinen Gesang wieder. *Orpheus receives Eurydice as a reward for his song.*	Besitz	u	Orpheus stimmt mit Gesang und flehentlicher Bitte die unerbittlichen Herrscher der Schatten um. *Orpheus convinces the implacable rulers of the shadows with songs and pleading requests.*	Kommunikation	u
6	Eurydike folgt Orpheus. *Eurydice follows Orpheus.*	Lokation	h	Der Gebieter über den Tod erlegt (dem Orpheus und der Eurydike) Bedingung(en) für den Aufstieg nach oben auf. *The lord of death imposes condition(s) on (Orpheus and Eurydice) for ascending.*	Gesellschaft	b
7	Orpheus glaubt nicht an Eurydike (hinter ihm). *Orpheus does not believe that Eurydice is behind him.*	Kognition	d	Eurydike geht hinter Orpheus her. *Eurydice walks behind Orpheus.*	Lokation	h
8	Orpheus denkt nicht (an die Bedingungen des Unterweltaufstiegs). *Orpheus does not remember the conditions of the ascend from the netherworld.*	Kognition	d	Orpheus erträgt den Aufschub (von Eurydikes Anblick) nicht. *Orpheus can not bear the delay of seeing Eurydice.*	Kognition	d
9	Orpheus blickt sich um. *Orpheus looks around.*	Perzeption	d	Orpheus blickt vor Erscheinen des Tageslichtes und (vor) Erreichen der Pforte des spartanischen Taenarus zu Eurydike. *Orpheus looks to Eurydice before the appearance of daylight and (before) reaching the gate of the Spartan Taenarus.*	Perzeption	d
10	Orpheus verliert Eurydike. *Orpheus loses Eurydice.*	Veraenderung	a	Orpheus verliert Eurydike. *Orpheus loses Eurydice.*	Veraenderung	a

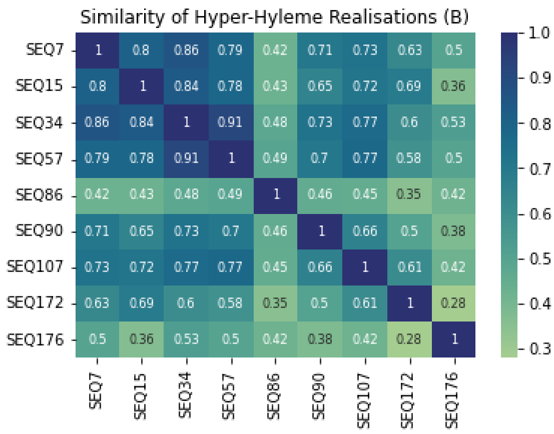


Figure 6.6: Cosine similarity of the hylemes belonging to the hyper-hyleme *B*

6.2.3 Comparing Plot Elements using Hyper-Hylemes

The hyper-hylemes presented at the beginning of the chapter can also be examined and used for comparison on the hyleme level. It is an interesting comparatistic research objective to inspect how exactly one element of a narrative, e.g. the communication of the conditions for Eurydice's successful rescue, is represented in different variants. For instance, we can compare the individual hylemes used to describe the same hyper-hyleme. In this work, this is achieved by grouping hylemes of a variant together if they belong to the same hyper-hyleme.

We can then employ sentence semantic similarity measures to find out, which variants are similar with regard to the representation of a certain function.

This is achieved by employing a sentence transformer approach using SBERT [183]. Sentence transformers can be used for sentences, but also on short texts that consist of multiple sentences. For that purpose, we group hylemes by their hyper-hyleme. For every variant, a SBERT embedding is created. Subsequently, for every pair of variants, the cosine similarity between the sentence embeddings is calculated. This allows us to compare hyper-hylemes that are represented by different numbers of hylemes. As an example, the cosine distance between different representations of the hyper-hyleme *b* (Conditions for Eurydice's rescue) are illustrated in Figure 6.6.

The maximum similarity (0.91), apart from the self-distance (1) on the matrix diagonal, is found in the pair SEQ34-SEQ57. There, the hyper-hyleme is represented as the following two hylemes:

- SEQ34: "Orpheus nimmt die Bedingung (für den Aufstieg und die Mitnahme Eurydikes) an." (engl. "Orpheus accepts the conditions (for the ascend and the taking of Eurydice).")
- SEQ57: "Orpheus nimmt die Bedingung (für den Aufstieg mit Eurydice) an." (engl. "Orpheus accepts the conditions (for the ascend with Eurydice).")

6.3 Event and Plot Alignment

6.3.1 Optimal Alignment

The sequences of hyper-hylemes that can be produced with the regular grammar can be aligned manually to create an optimal alignment. This optimal alignment is used to evaluate the performance of alignment algorithms. For that purpose, we exclude the empty sequences and the single-hyleme sequences. Figure 6.7 shows the manually created alignment. Gaps in the aligned sequence occur when at least one other sequence (target sequence) has a different hyper-hyleme inserted in the position of the hyper-hyleme in the source sequence. Sequence 15 contains the longest common subsequence (LCS $= 9$), and the least gaps ($|g| = 1$, position: 8).

	SEQ1	SEQ7	SEQ15	SEQ34	SEQ55	SEQ57	SEQ86	SEQ90	SEQ107	SEQ140	SEQ152	SEQ172	SEQ176	SEQ177
1				V		V								
2			T	T		T		T			T	T		
3	E		E			E		E				E	E	E
4	R	R	R	R	R	R		R	R	R	R	R	R	R
5	U	U	U		U	U	U	U	U	U			U	
6			Z									Z		Z
7		B	B	B		B	B	B	B			B	B	
8						Z								
9	H	H	H	H	H						H	H	H	
10		D	D	D	D	D	D	D	D		D	D	D	
11		A	A	A	A	A		A			A	A	A	
12			F								F	F	F	F
13									M			M		M

Figure 6.7: Optimal alignment of the non-empty sequences with sequence length > 1, no substitutions

The hyper-hyleme z, which can appear in two positions, either before or after the conditions for Eurydice's rescue are laid out to Orpheus, can be substituted in one case (Sequence 57). For the purpose of aligning narrative sequences, we allow a substitution in position i under the following condition:

A substitution of a hyper-hyleme h_t in a target sequence S_t may be performed, if and only if, there is at least one other sequence S_s, where

$$h_{s_i-1} = h_{t_i-1} \wedge h_{s_i+1} = h_{t_i+1} \wedge \forall S_t, h_{t_i} \in S_t : h_{t_i} \neq h_{s_i}, \tag{6.2}$$

i.e. the preceding and succeeding hyper-hylemes in source- and target-sequence match, and there is no matching hyper-hyleme in the position i in any target-sequence. If we allow substitutions, the optimal alignment would perform one substitution ($h \rightarrow z$). The resulting alignment is shown in Figure 6.8. When substitutions are applied, Sequence 15 remains intact (without gaps). The optimal alignment does not allow mismatches, with exception of $h \rightarrow z$ for the substitution, because each of the hyper-hylemes fulfills its own narrative role in the myth variants. Therefore, a mismatch would narratologically be a replacement of one narrative component with another. This would mean, for instance, Eurydice's death *t* (SEQ152) would be replaced by Orpheus descending into the netherworld *e* (SEQ1). Instead, the optimal alignment favours gaps, which can be interpreted as a certain narratological component being not present.

We define the cost function of the alignment as follows:

- $c(gap) = 0.5$, if gap occurs in *PREP* or *RES* positions.
- $c(gap) = 1$, if gap occurs in the *CORE* positions.
- $c(sub) = 0.25$, if substitution applies.

With this cost function, missing hyper-hylemes from the core set are punished more severely than missing hyper-hylemes from the *PREP* or *RES* set. The nature of the *PREP* and *RES* functions, e.g. m, is often to allude to other, related myths, e.g. the Thrakian women killing Orpheus, because he did not share the mysteries with them. The gap cost is modelled as a *gap open* penalty [3], in contrast to a *gap extension* penalty, i.e. each gap is punished the same, resulting in a linear growth of penalty over long gaps.[7] We aim to minimize the cost of possible alignments to find the optimal alignment.

Figure 6.8 reports the values of the sequences of hyper-hylemes according to the cost function in comparison with the maximum sequence S_{max}, i.e. all the hyper-hylemes of Σ in consecutive order.

	SEQ1	SEQ7	SEQ15	SEQ34	SEQ55	SEQ57	SEQ86	SEQ90	SEQ107	SEQ140	SEQ152	SEQ172	SEQ176	SEQ177
1	0.5	0.5	0.5	V	0.5	V	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	0.5	0.5	T	T	0.5	T	0.5	T	0.5	0.5	T	T	0.5	0.5
3	E	0.5	E	0.5	0.5	E	0.5	E	0.5	0.5	0.5	E	E	E
4	R	R	R	R	R	R	0.5	R	R	R	R	R	R	R
5	U	U	U	1	U	U	U	U	U	U	1	1	U	1
6	1	1	Z	1	1	0.25	1	1	1	1	1	Z	1	Z
7	1	B	B	B	1	B	B	B	B	1	1	B	B	1
8	H	H	H	H	H	Z	1	1	1	1	H	H	H	1
9	1	D	D	D	D	D	D	D	D	1	D	D	D	1
10	0.5	A	A	A	A	A	0.5	A	0.5	0.5	A	A	A	0.5
11	0.5	0.5	F	0.5	0.5	0.5	0.5	0.5	0.5	0.5	F	F	F	F
12	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	M	0.5	0.5	M	0.5	M
	5.5	3.5	**1**	3.5	4.5	1.25	5.5	3.5	4.5	7	4.5	**1.5**	2.5	5.5

Figure 6.8: Optimal alignment of the non-empty sequences with sequence length > 1, with substitution (SEQ57) and cost function, S_i against S_{max}

6.3.2 Alignment Algorithms

For the application of alignment algorithms, we can apply two different approaches. Global alignments take whole sequences into consideration. For the purpose of myth variant comparison, i.e. the comparison of hyper-hyleme sequences that are by default similar, global alignments are useful. In contrast, local alignments try to find matching subsequences in two or more sequences. For exploratory hyleme sequence alignment, local alignments are better suited, because they can identify patterns and sub-sequences in sequences from different domains. Therefore, the optimal

[7] In biological sequences, longer gaps might be expected and it would not be desirable to punish them as severely.

alignment of the sequences pertaining to the myth of *Orpheus and Eurydice* is used to test the performance of the alignment methods.

For global alignment, the Needleman-Wunsch algorithm [30][8], as introduced in Section 2.3.1, can be used. In order to see how well the algorithm performs with regard to the gold standard, we check if the alignment in Figure 6.8 can be re-produced for the sequences against the maximum sequence S_{max}, i.e. the consecutive sequence of all hyper-hylemes in Σ. For that purpose, we set the cost function for the Needleman-Wunsch algorithm to:

- $c(match)$ = 1
- $c(missmatch)$ = -1
- $c(gap)$ = -0.5

Additionally, we check if the alignment between two sequences S_i and S_j can be reproduced. For that purpose, the alignment of S_i and S_j is the alignment of S_i against S_{max} resp. S_j against S_{max}, with the reduction of a gap in position k, if $S_{i_k} = S_{j_k} = gap$, i.e. a gap that occurs in the same position in both sequences is removed.

The agreement between the alignment obtained from the Needleman-Wunsch algorithm using the aforementioned cost function and the reference alignment described earlier is depicted in Figure 6.9. The algorithm accurately predicts all alignments between S_i and S_{max}. Moreover, the algorithm performs well in aligning the longest sequence, SEQ15, with the remaining sequences, except for SEQ107, where discrepancies are observed. These discrepancies primarily arise from mismatches, which the Needleman-Wunsch algorithm permits, but the gold standard does not allow due to the reasons mentioned above. The overall accuracy of the algorithm on the hyper-hyleme sequences is 0.61.

The local alignment algorithm of Smith-Waterman [32] algorithm can identify sub-sequences in the sequences, as introduced in Section 2.3.1. In contrast to the LCStr method, the Smith-Waterman patterns can include gaps. This way, patterns can be identified between a pair of sequences, even if one sequence has an inserted hyper-hyleme. Figure 6.10 displays the alignment of two sequences (SEQ1 and SEQ90) as achieved by applying the Smith-Waterman algorithm.[9] One pair of sequences, SEQ86 and SEQ177, does not share a common pattern. All other pairings of sequences (nodes in Figure 6.10) have at least one common hyper-hyleme, which can serve as a minimal pattern. The frequency of shared hyper-hylemes in the Smith-Waterman subpatterns are shown in Figure 6.11. The hyper-hylemes *r,u,d,a,b* occur frequently in patterns. *u,b,d* are hyper-hylemes from the *CORE*-set, *a* is the result of Orpheus' failure, i.e. the *second loss*. Based on the frequency of these hyper-hylemes in the data, the hypothetical sequence *RUBDA*, which can be constructed from the

[8]The Needleman-Wunsch and Smith-Waterman algorithms are implemented in the python package minineedle, which has been used here. `https://pypi.org/project/minineedle/`

[9]An interactive version of the plot can be accessed under: `https://teaching.gcdh.de/hyleme/Smith-Waterman/`

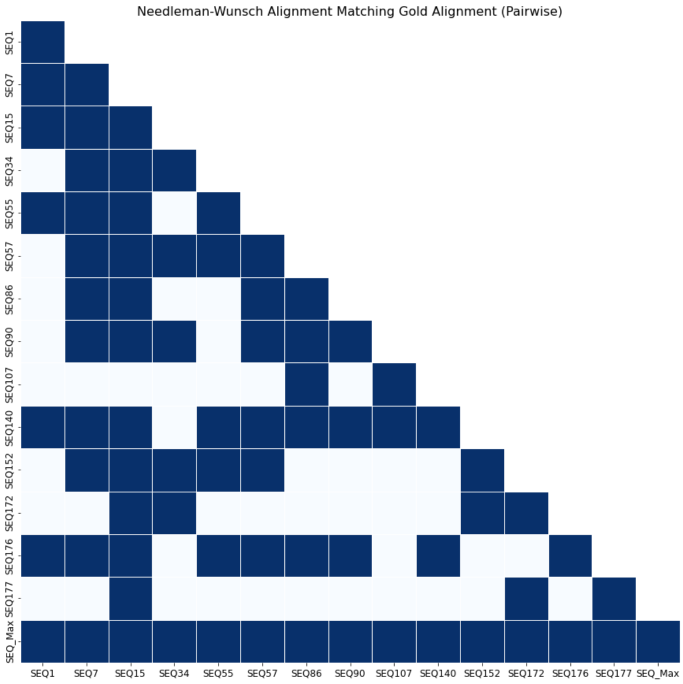


Figure 6.9: Needleman-Wunsch alignment corresponding to gold alignment, blue=gold alignment reproduced

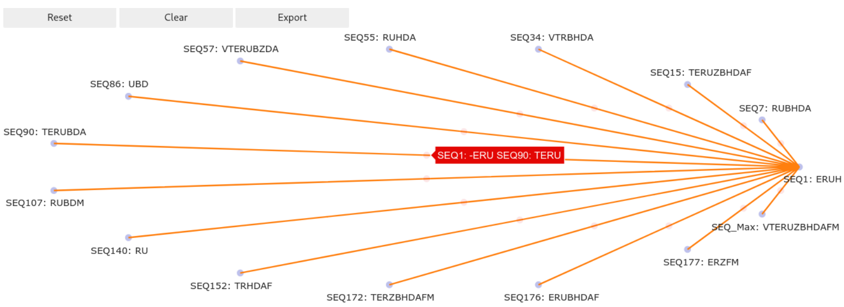


Figure 6.10: Smith-Waterman alignment patterns corresponding, highlighted example: SEQ1-SEQ90

regular grammar, can be seen as one possible minimal example of the Orpheus' *Stoff*.

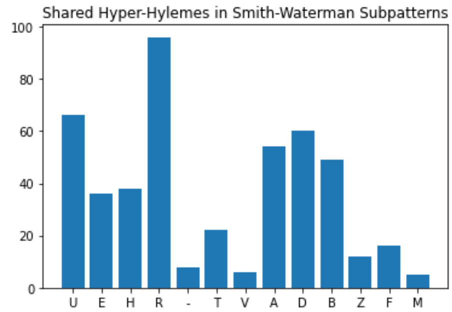


Figure 6.11: Hyper-Hyleme Frequency in Smith-Waterman alignment patterns

6.4 Summary

In this section, we modelled the event and plot structure of variants of the popular myth *Orpheus and Eurydice* using hyleme sequences. For that purpose, we introduced hyper-hylemes which make the comparison and alignment of core plot elements possible. On the basis of these hyper-hylemes, a regular grammar was constructed. Using the production rules of the grammar, sequences of hyper-hylemes for all variants of the *Orpheus and Eurydice* myth can be constructed.

The regular grammar can help verify if new versions follow the same structure as the versions present in the hyleme data set. In turn, new sequences help verify the validity of grammar, if they can be generated by the production rules. We demonstrated this on two new variants, a variant based on a version in Plutarch, Amatorius 17, and a version generated by the artificial intelligence agent ChatGPT.

Sequence similarity algorithms, such as Longest Common Substring (LCStr) or Levensthein distance, can be applied to the sequences of hyper-hylemes. We introduce an alignment approach including a specific cost function for the alignment of hyper-hyleme sequences. Additionally, the hyleme representation of the hyper-hylemes can be examined. This way, we can inspect closely how variants differ in the communication of a single event, e.g. the communication of the conditions of Orpheus' and Eurydice's ascend. This can be achieved by employing semantic similarity measures, such as cosine similarity of sentence embeddings.

In this chapter, an optimal alignment based on narratological considerations, is proposed and tested against global and local alignment approaches. The Needleman-Wunsch algorithm achieves 0.61 accuracy against the gold alignment. The Smith-Waterman algorithm identifies common patterns in the sequences. Most frequently, these patterns include the hyper-hylemes *r,u,b,d,a*, a hypothetical sequence which can be produced using the regular grammar.

The regular grammar will be published under the title *To Love and to Lose: A regular grammar for the hylistic comparison of the Orpheus and Eurydice-Stoff* as a chapter in the upcoming anthology *Mächte und Bereiche der Unterwelt in Mesopotamischen und Griechischen Quellen* (working title) in the series *Mythological Studies (MythoS)*, Vol. 6, 2024, de Gruyter, Berlin/Boston.

Chapter 7

Hyleme Matching Approaches and Semantic Proximity

Alignment algorithms are used to find structural similarities between sequences. In bioinformatics, alignments algorithms find patterns from DNA (and other types of) sequences. In natural language processing (NLP), alignment often refers to finding mentions of the same events in one or multiple texts, or detecting paraphrases or translations of the same statement across documents.

In order to achieve an alignment between two or more sequences of sentences (e.g. texts), a similarity measure between the individual items to be compared (e.g. clauses, sentences, predicate arguments or hylemes in our case) is needed. This similarity measure determines when two items are similar enough to be considered for alignment. What exactly this means depends largely on the intention of the user. For this thesis, two applications are relevant: 1. If a user is interested in comparing variants of the same myth with many lexically and semantically similar hylemes, the matching approach needs to be relatively strict. 2. If a user wants to perform exploratory alignment, to uncover similarities in hyleme sequences, the matching approach needs to find similarities that are less plain than just lexical agreement. It would need to consider semantic variations, such as slight shifts in meaning, or instances where only the actions as implied in the hyleme predicate is the same, but the arguments (such as characters, locations, and objects) are different.

In the simplest case, e.g. when investigating base pairs in DNA sequences, this match would include a simple string comparison operation on a vocabulary of four different bases (ATGC[1]). Each base in one sequence is compared to the base in the other sequence at the same position. For instance, a match is found if both sequences contain the same literals at the same positions[2]. If sequence S_1 is represented as $ACGT$ and sequence S_2 is $ACCT$, the alignment predicted would be

[1](Adenin, Thymin, Guanin and Cytosin)

[2]For illustration purpose, I ignore substitutions in this example.

AC-T, with one gap.

However, for hyleme sequences, the matching is not as straightforward, because a full string match, e.g. two hylemes are identical in every aspect, rarely occurs in two hyleme sequences. Important points of alignment, i.e. semantically similar hylemes, might be missed. Hyleme sequences do not use a controlled vocabulary, are sensitive to different domains, and in parts also to the language of the source text (which can influence the translation). Furthermore, different domain experts might use different phrases when extracting hylemes for similar events or circumstances, depending on their research background or personal preference.

As demonstrated in the Section 6.1, an ideal alignment would be performed on higher-level hylemes, i.e. hyper-hylemes. However, the data sets do not include hyper-hyleme information. For the German data set, the hyleme management and annotation software provides options to include hyper-hylemes. However, to date this functionality has not yet been used by the domain experts. The English data set includes Callaway's headings (see Chapter 4), which can be interpreted as hyper-hylemes in some cases. However, their form does not strictly follow hylistic theory (SPO-structure, no passives, present tense, one finite verb), so that an alignment on the sub-headings would be technically feasible, but would not yield interpretable results. Hence, in this chapter, the identification of alignment candidates is performed on the hylemes directly.

In this chapter, I first aim to establish a gold standard for ranking hyleme pairs in terms of semantic *proximity*. Therefore in Section 7.1, the annotation of a gold standard for the semantic proximity of the German hyleme data set is presented. I introduce a survey undertaken with experts in mythological studies and non-experts, in which annotators were asked to rate 4-tuples of hyleme pairs according to the perceived semantic promximity. Subsequently, in Section 7.2, various matching approaches are presented and evaluated according to different criteria. One or multiple measures can be applied to determine the semantic *proximity* of hylemes in two sequences. For that purpose, a match is returned, either if the semantic *distance* is below a certain threshold (i.e. hylemes are close in meaning), or if the matching method returns the boolean value *True*.

7.1 Semantic Similarity Annotation

Best-worst scaling approaches have proven to be robust measures for annotation tasks in the field of natural language processing, e.g. for semantic relatedness [149] or sentiment analysis [148] tasks. In a best-worst scaling setting, each annotator has to select the best (closest) and the worst (furthest) fit from n tuples according to the task in question. In a semantic similarity or relatedness task, this means that the closest fit is the tuple which is the most similar or the closest related in comparison to all other pairs. Most commonly, the number of pairs is $n = 4$. One of the advantages of best-worst scaling approaches is that it provides a partial ranking with only two ratings, p_i (best fit) and p_j (worst fit). This yields different ratings of fits, where $p_i > p_j$ indicates a higher similarity of the pair p_i compared to p_j For instance, the following ratings can be inferred if p_1 is the best fit

and p_4 is the worst fit:
$p_1 > p_4$, $p_1 > p_2$, $p_1 > p_3$, as well as $p_4 < p_2$, and $p_4 < p_3$, where p_1 is the worst fit (furthest semantic distance), and p_4 is the best (closest) fit. However, full ranking cannot be achieved with a best-worst approach, because the values of p_2 and p_3 can not be determined.

7.1.1 Annotation Design

In order to achieve a gold standard for hyleme similarity, a small subset of hylemes was assessed. For that purpose, 178 unique hylemes where selected from the German data set, e.g. *Der Mann sieht jugendlich aus* (engl. "The man looks youthful"). Each hyleme was then used as a basis to create one or more variants of that hyleme, e.g. *Der Mann sieht stark und jugendlich aus* (engl. "The man looks strong and youthful."). The base hylemes were selected from each of the disciplines present in the German hyleme data set. In order to model certain grammatical or semantic phenomena of two hyleme alignment candidates, the base hyleme was transformed into a new hyleme. Therefore, each pair p consists of two hylemes h_1 and h_2, where $h_2 = t(h1)$, and t is one of the possible transformations. Transformations t are selected from a number of effects that are present in the hyleme data set. These transformations can apply to different hyleme components, e.g. concerning hyleme subject or hyleme predicate.

The transformed hylemes need to fulfill the following requirements:

- The hyleme needs to be a grammatically correct German sentence.
- The hyleme needs to be semantically correct, and meaningful.
- The hyleme needs to be plausible in the broader mythological context.

Each annotated item consists of a 4-tuple (p_1-p_4) of pairs of hylemes h_1 and h_2. The transformations can be grouped into nine coarse classes, and 24 sub-types (see Appendix Section A.3.3). The coarse classes are:

- *(1) Additions/Deletions*: Additional component present in one hyleme, that is not present in the other hyleme. Example: *Der Mann sieht (stark und) jugendlich aus.* (engl. "The man looks (strong and) youthful.")
- *(2) Taxonomical Relationship*: Hyponym/hypernym relationship of one hyleme component, e.g. hyleme predicate.[3] Example: *Der Pharao/König sieht die Bewohner der Unterwelt.* (engl. "The pharaoh/king sees the inhabitants of the netherworld.")
- *(3) Changes concerning* durative-constant *hylemes*: Changes concerning the subject in *durative-constant* hylemes, e.g. subject complements. Example: *Sie ist eine Amme (und eine Mutter).*

[3]Candidates for taxonomical relationships were identified using GermaNet [46]. The component of the base hyleme was queried in GermaNet. Then the candidate synsets were manually reviewed, the correct synset was chosen, and one of the candidate replacements was selected from the hyper-/hyponyms by the author according to how well it fit the base hyleme with respect to the requirements mentioned above.

(engl. “She is a wet nurse (and a mother).”)

- *(4) Changes of hyleme components*: Replacement of one hyleme element with a variation, which is not taxonomically related, but fulfills the above requirements. Example: *Er bringt eine Person/ein Lämmchen herein.* (engl. “He brings in a person/a little lamb”.)
- *(5) Changes of grammatical number*: One hyleme component in plural, with the corresponding component in singular. Example: *Das/Die Feuer sind gleichmäßig.* (engl. “The fire is even./The fires are even.”)
- *(6) Change or specification of quantity*: Quantities inserted or changed for one hyleme component, where applicable, e.g. *one priest* vs. *seven priests*, or change in expression of quantity (*some/many*). Example: *Die Größe der Riesen ist dreihundert/dreitausend Ellen.* (engl. “The size of the giants is three hundred/thousand cubits.”)
- *(7) Hyleme rephrasing*: Rephrasing of one hyleme into the other while keeping original meaning as closely as possible. Example: *Sie ist keine Göttin/nicht göttlich.* (engl. “She is not a goddess/not divine.”)
- *(8) Negations*: Negations present in one hyleme, but not in the other, including indefinite pronoun (*kein*) or negation particle (*nicht*). Example: *Elischas Mantel fällt (nicht) herunter.* (engl. “Elisha’s cloak (falls/does not fall) down.”)
- *(9) Antonymy*: Antonymy relationship of one hyleme component. Example: *Ba’l soll sich vom Berg Knkny fernhalten/soll sich zum Berg Knkny begeben.* (engl. “Ba’l has to (stay away from/go to) Mount Knkny.”)

The annotation of hyleme similarity was performed as a best-worst-ranking experiment using 4-tuples of hyleme pairs. It was presented as an online survey, which was distributed to experts of the STRATA research group, and to students and graduate students of computer science, data science, and the humanities.

On the landing page of the annotation survey, the participants were given an instruction text. The text contained a description of the research objective. However, the text did not include hylistic terms, since the survey was designed to address participants who did not need to have prior engagement with hylistic analysis, see Section 7.1.2. Since the matching of hylemes is a matter of semantic similarity as well as relatedness, the participants were asked to take both into consideration when judging the “semantic proximity” (Ger: “semantische Nähe”). Participants were given examples for semantic similarity and semantic relatedness. The introductory text can be found in Appendix A.3.1.

Each page of the survey contained four 4-tuples, and a short introductory text. In the introductory text, participants were reminded of the objective and instructed to provide an answer, even if they found multiple pairs to be equally close. The text also guided participants to use individual assessment criteria, which they were later asked to provide.

In a short pre-study with fewer participants, who were members of the author's peer group of fellow graduate students with a background in computational linguistics (CL)/natural language processing (NLP), it was determined that the annotation task requires a high degree of concentration. Therefore, the final study was divided into two parts, with a short interlude after 12 4-tuples. This included an assessment question, in which the participants where asked to rate the difficulty of the task and provide some feedback in form of free text. Additionally, participants were asked if they recognized any mythological contexts from the set of sentence pairs they just evaluated. This question aimed to prevent participants from feeling fatigued by the annotation process and to help them refocus their attention.

The survey was created with the intended purpose of a gold standard annotation. With 27 4-tuples, the sample size is too small to gain meaningful statistical insights on the influence of the transformations t on the similarity between h_1 and h_2. A larger study, which would allow a statistical analysis, would require a larger amount of items per transformation, resulting in an overall larger number of 4-tuples.

Annotator Information

In total, 20 participants annotated the 4-tuples of German hylemes according to the perceived semantic similarity and relatedness. The data was collected in form of an anonymous survey. Two out of the 20 annotators only partially answered the questionnaire. The annotators were presented with the 4-tuples and asked to select the best (closest) match and the worst (furthest) match by selecting them from two multiple choice options. The results were inspected for malicious annotations, e.g. patterns (ABAB etc.). No malicious annotations were found.

While preserving anonymity, the annotators were asked to provide general information on their academic background. At the time of the survey, nine of 20 annotators held a Master's degree, six held a PhD, four participants had a Bachelor's degree and one was a habilitated professor.

The majority of annotators (14) has a background in the humanities, but a number of annotators from scientific fields were included. The distribution of annotators by field can be found in Figure 7.2, multiple selections were allowed.

An important piece of information for evaluating the annotation in the context of hyleme theory is the question of previous experience with hyleme extraction or analysis. Eleven of the 20 annotators had no previous experience with hylemes, while nine had worked with hylemes and hyleme sequences before.

The survey also investigated to what extend annotators engaged with mythological content in their academic work or studies. The options ranged from constant engagement with mythological content as the participants main area of study/research (4) to no engagement with mythological

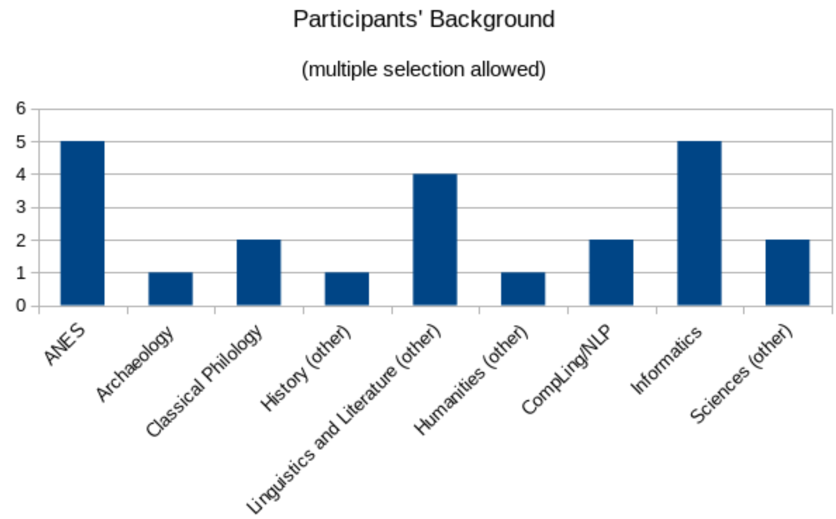


Figure 7.1: Annotators' academic background

content (8). A detailed breakdown can be found in Figure 7.2. Most of the annotators are German native speakers (18), two annotators have an advanced German proficiency level (C1-2 CEFR[4]).

After four pages of annotations (12 4-tuples), the participants where asked to give feedback regarding the perceived difficulty of the task. The answer options were *not difficult* (0), *a little bit difficult* (6), *challenging* (11), *very difficult* (3). Furthermore, participants were asked if they identified any myths among the sentence pairs they had just annotated. Out of the 20 participants, 14 indicated that they recognized specific contexts. In a free text field, they were requested to provide further details on the myths they identified. Several participants correctly identified all or most of the contexts, while others only recognized certain individual contexts, such as the Bible.

The inter-rater agreement (*Fleiss Kappa*) between the 18 annotators who provided all 54 answers (min/max for each of the 27 4-tuples), is $\kappa = 0.3162$.[5] It has to be noted that the first set of questions (12 4-tuples) has a slightly higher agreement $\kappa = 0.3358$ than the second set of questions (15[6] 4-tuples) $\kappa = 0.2981$. Overall the annotation results indicate *fair* agreement [9]. This does not necessarily mean that annotations were of poor quality. It rather indicates a high difficulty of the annotation task, which is confirmed by the feedback given on the perceived difficulty of the task. Furthermore, the inter-rater agreements may indicate a considerable degree of subjectivity involved in the human ratings.

After the second set of questions, the participants were asked to provide details on how they

[4]Common European Framework of Reference for Languages

[5]Python package statsmodels `https://www.statsmodels.org/stable/generated/statsmodels.stats.inter_rater.fleiss_kappa.html#statsmodels-stats-inter-rater-fleiss-kappa`

[6]The last question containing four 4-tuples had to be excluded from the evaluation of the study, because the answer was accidentally not provided as a set of multiple choices, but as a free text option.

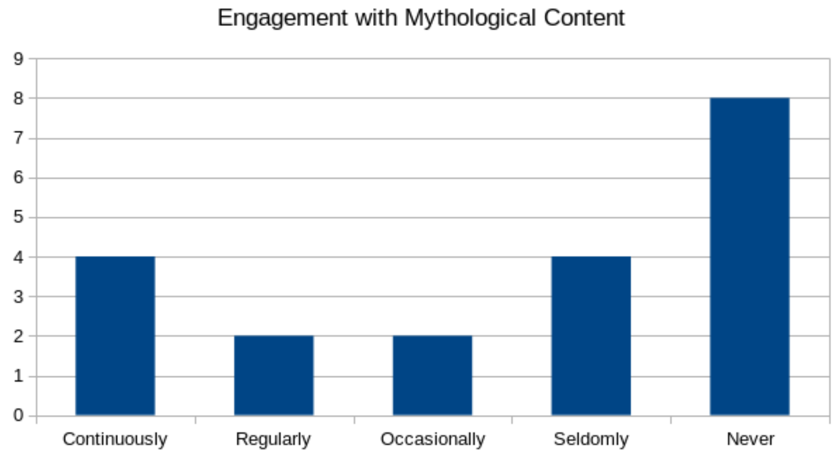


Figure 7.2: Annotators' academic engagement with mythological content

approached the task. They were presented with multiple options, such as *linguistic*, *by content*, *by instict*, or *other* and given the opportunity to provide free text for additional explanations.

The participants provided different levels of detail concerning their approach to solving the task. Some participants gave very detailed answers, indicating a high level of introspection and structured methodology. For example, participant #8 provided a list of linguistic considerations that he or she used to determine the semantic proximity of the sentence pairs:

> "Entailment rated higher and Contradiction less high. Weighting of differences by obliqueness: Subject/agent differences weighted higher (i.e. further away) than Patiens/dir. Obj., Instrument/indir. Obj. etc., synonyms considered closer than paraphrases/antonyms considered closer than explicit negation with "not" etc." (transl. by author)

One participant specified the *instinct* option further, indicating that he or she imagined the scenes described in the sentence pairs and judged the difference in the mental pictures that they invoked, e.g. in judging the different between quantities, e.g. "The size of the giants is three (hundred/thousand) cubits."

Table 7.1: Annotation study statistics

#Hyleme Pairs	#4-tuples	#Annotations per tuple	#Annotations
108	27	min. 18	518

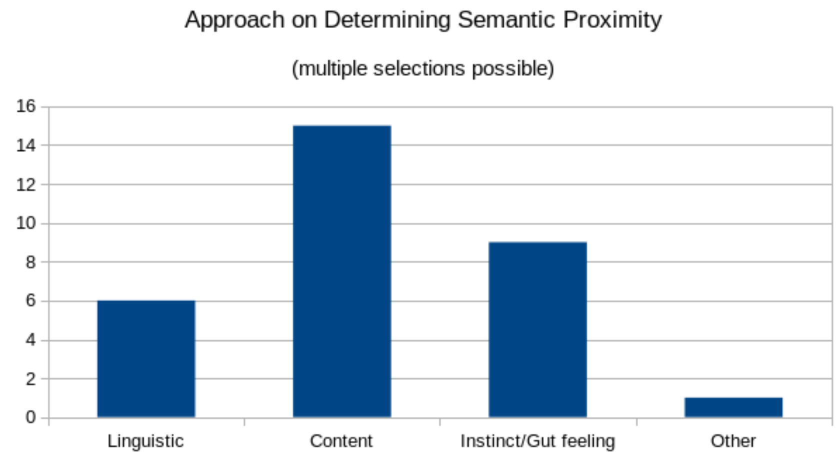


Figure 7.3: Annotators' approaches to semantic proximity annotation

7.1.2 Annotation Results

In this section, the results for the gold standard annotation for the hyleme similarity ranking task are reported. As mentioned above, the annotation data contains a set of 27 4-tuples. The data combined with the high number of transformation effects does not provide enough data to gain meaningful statistical insights. Therefore, we report the results as absolute quantities.

Paar A	'Noah nimmt von dem reinem Vieh je sieben mit.'
	'Noah nimmt von dem reinen Vieh je siebzehn mit.'
transl.	'Noah takes (seven/seventeen) of each of the pure animals with him.'
Paar B	'Noah nimmt von dem reinem Vieh je sieben mit.'
	'Noah betrachtet von dem reinem Vieh je sieben.'
transl.	'Noah takes (takes/beholds) of each of the pure animals (with him).'
Paar C	'Die sieben Ritualkundigen befestigen den unendlich langen Faden für Dumuzi. '
	'Die sieben Ritualkundigen spinnen den unendlich langen Faden für Dumuzi.'
transl.	'The seven ritual experts (fasten/spin) the endless thread for Dumuzi.'
Paar D	*'Die sieben Ritualkundigen befestigen den unendlich langen Faden für Dumuzi.'*
	'Die Ritualkundigen befestigen den unendlich langen Faden für Dumuzi.'
transl.	'The (seven) ritual experts fasten the endless thread for Dumuzi.'

Table 7.2: 4-tuple no. 14, emphasis=best/closest match (Agreement: 18 annotators)

Two items in the set have a particularly high agreement over all annotators. Tuple no. 14 (see Table 7.2) was annotated by 19 participants, of which 18 agreed that hyleme *pair D* had is the closest in similarity/proximity. Similarly, tuple no. 17 (see Table 7.3) had 18 of 19 participants agreeing that hyleme pair D was the furthest/worst match.

Two items did not have a majority vote for one hyleme pair. 4-tuple no. 13 has no conclusive best match (Pair A, B, and C: five votes each, Pair D: four votes), while tuple no. 26 has no majority

Pair A	'Saturus und Perpetua sehen ein großes Licht.'
	'Saturus und Perpetua sehen große Lichter.'
transl.	'Saturus and Perpetua see (a big light/big lights).
Pair B	'Baumeister des Grabes darf das Grab nicht finden.'
	'Baumeister der Gräber darf das Grab nicht finden.'
transl.	'The master builder of the (grave/graves) is not allowed to find the grave.
Pair C	'Der Mann sieht jugendlich aus.'
	'Der Mann sieht stark und jugendlich aus.'
transl.	'The man looks (strong and) youthful.'
Pair D	*'Der Sohn Lamechs verbirgt sich.'*
	'Der Sohn und die Tochter Lamechs verbergen sich.'
transl.	'Lamech's son (and daughter) (hides/hide).'

Table 7.3: 4-tuple no. 17, emphasis=worst/furthest match (Agreement: 18 annotators)

vote for the worst match (Pair A and B, seven votes each). Table 7.4 shows the results of the gold standard annotation.

In Table 7.5 the number of ratings as best and worst match per hyleme component affected in the transformation, and the total occurrence of the component transformation across all 4-tuples are reported. The subject transformation occurs more often than predicate or object transformations. Not all base hylemes have a hyleme object, hence the number is lower than predicate and object transformations. Addition and deletion effects can only be applied to predicate determinations, since a hyleme can only have one predicate. In contrast, addition effects on hyleme subject and object can be applied on both the subject itself ('Saturus (and Perpetua)'[7]) and on the subject determination ('Eurydike's (much beloved) husband'). Therefore, effects on the hyleme subject occur more often than on the predicate or object. As mentioned above one 4-tuple did not have a majority vote for best fit, and one 4-tuple did not have a majority vote for worst fit. Hence, the sum of the best/worst columns in Tables 7.5 and 7.6 equals 26.

From Table 7.5 no distinct preference for one of the hyleme components in terms of best or worst match can be observed.

Table 7.6 shows how often certain effects were selected as best/worst fit, the number of total occurrences of the effect in the data, and the number of 4-tuples in which the effect is present at least once. Effects 1,2, and 4, i.e. *Addition/Deletion*, *Taxonomical Relationship*, and *Changes of hyleme components* occur more often than the other effects, because they manifest themselves in different ways (e.g. *Addition/Deletion* of one word or entire phrases), resulting in many different possible pairings with the hyleme components. While *Addition/Deletion* effects occur as best (8) and worst (10) fit, *Changes*, which occur without taxonomical relationship are most often selected as the worst fit, unless paired with effects which have a strong tendency towards dissimilarity, i.e. *Antonymy*, or *Negation*. *Hyper/Hyponymy* effects are rated as best match (10) more often than worst match (4).

[7]The reader may be reminded that the hyleme subject is different from the grammatical subject. A conjunction involving *and* is typically counted as two distinct hyleme subjects.

Table 7.4: Annotation results for the best-worst annotation task, Best/Worst Fit aggregated annotation results, i.e. number of raters per sentence pair, Coarse Effects (s. Section 7.1.1), Hyleme components: s = subject, p = predicate, o = object, incl. component determinations

Item	Best Fit	Worst Fit	Coarse Effects	Hyleme component
1	[0,1,3,**16**]	[**10**,9,1,0]	[1,1,2,2]	[s,o,o,o]
2	[**17**,3,0,0]	[0,0,**15**,5]	[1,3,4,4]	[o,s,p,o]
3	[7,0,**11**,2]	[3,5,2,**10**]	[1,1,1,1]	[p,p,s,s]
4	[**13**,2,5,0]	[2,6,2,**10**]	[1,3,2,2]	[p,s,s,s]
5	[6,**12**,0,2]	[0,1,**18**,1]	[1,2,3,2]	[o,p,s,s]
6	[4,3,1,**12**]	[1,7,**11**,1]	[1,2,1,2]	[o,s,p,p]
7	[2,2,1,**15**]	[**8**,6,5,1]	[1,2,1,2]	[o,o,p,o]
8	[0,6,**13**,1]	[**13**,4,0,3]	[1,2,2,1]	[p,p,s,p]
9	[4,**14**,2,0]	[0,0,1,**19**]	[1,2,4,1]	[s,o,s,p]
10	[6,**8**,0,6]	[0,1,**19**,0]	[1,2,4,4]	[p,p,s,p]
11	[2,**9**,3,6]	[6,0,**13**,1]	[1,2,2,2]	[s,o,o,p]
12	[**13**,6,0,1]	[3,2,4,**11**]	[1,2,5,6]	[s,o,s,s]
13	[5,5,5,4]	[4,**7**,2,6]	[1,2,1,4]	[o,s,o,o]
14	[1,0,0,**18**]	[5,**10**,4,0]	[6,4,4,6]	[o,p,p,o]
15	[**10**,5,0,4]	[0,4,**10**,5]	[1,6,4,4]	[s,s,s,p]
16	[3,0,**12**,4]	[4,**12**,0,3]	[6,6,6,5]	[s,o,s,s]
17	[2,**12**,5,0]	[0,0,1,**18**]	[5,5,3,1]	[o,s,s,s]
18	[0,**12**,6,1]	[4,0,0,**15**]	[4,5,5,4]	[s,o,s,p]
19	[1,**16**,2,0]	[**10**,0,3,6]	[1,1,1,1]	[o,p,o,s]
20	[**7**,3,5,4]	[6,**7**,4,2]	[1,1,4,3]	[p,o,p,s]
21	[7,**11**,0,0]	[0,0,6,**12**]	[3,2,4,4]	[s,o,p,o]
22	[0,2,**16**,0]	[**12**,0,1,5]	[4,2,2,1]	[s,s,p,s]
23	[**12**,1,1,4]	[1,5,**11**,1]	[2,2,1,1]	[p,p,p,p]
24	[**12**,1,0,5]	[3,4,**9**,2]	[3,2,1,2]	[s,s,s,o]
25	[8,**10**,0,0]	[0,0,7,**11**]	[7,7,8,9]	[s,s,p,p]
26	[1,0,1,**16**]	[7,7,3,1]	[8,9,8,4]	[p,p,s,p]
27	[3,**9**,4,2]	[7,0,3,**8**]	[1,3,1,5]	[p,o,o,o]

Table 7.5: Number of best-worst ratings per component effect,Total=Number of hyleme pairs with any effect on component present, sum of best/worst columns = 26

Component	Best	Worst	Total
S	8	11	42
P	9	9	35
O	9	6	31

Table 7.6: Number of best-worst ratings per coarse effect, Total=Number of hyleme pairs with effect present, #4-tuples=Number of 4-tuples with effect present at least once

Effect	Best	Worst	Total	#4-tuples
1	8	10	36	21
2	10	4	26	16
3	2	1	8	8
4	1	7	17	11
5	2	1	7	5
6	2	2	7	4
7	1	0	2	1
8	0	0	3	2
9	0	1	2	2

Conclusion

The gold standard annotation of sentences (hylemes), especially with mythological content where multiple layers of interpretation are possible, is not an easy task as the relatively low inter-rater agreement (Fleiss κ = 0.3162) indicates. This is also confirmed by the participants difficulty estimation as mostly *challenging*. A set of multiple studies with less transformation effects per study, but more examples for each effect and each hyleme component would reveal if any of the effects has a statistical significance with regard to the hyleme similarity. This study can be seen as a pre-study towards this research question, but since effects and components have too little representation in the 27 annotation items (4-tuples), no meaningful statistical insights can be gained from it. With relatively few different examples per effect, the gold standard annotations would most likely reveal a lexical relation rather than the influence of a certain transformation effect as introduced in Section 7.1.1.

As indicated above, the aim of this annotation was to create a gold standard for the hyleme alignment task. Therefore, a larger study which would deliver interesting insights, is left for future work.

7.2 Hyleme Similarity Modelling

As explained in the introduction to this section, a similarity measure is needed in order to align hylemes in two or more sequences. What exactly makes two hylemes similar depends on the context and the research question. That the ranking of hyleme pairs with regard to semantic similarity or proximity is not a trivial matter was shown through the annotation task introduced in the previous section. Additionally, changes in hylemes that are candidates for alignment can occur through many different transformation effects, which will have to be considered when semantic similarity methods are applied.

For the purpose of automatically achieving a hyleme matching, i.e. the automatic identification of candidates for points of alignment, a number of different approaches was implemented. The following section presents the different measures, roughly ordered by their strictness and computational complexity. The objective of all the approaches is to yield a boolean value $true$ or $false$ if two hylemes h_1 and h_2 are the same with regard to a matching method, or semantically similar/close enough based on a certain threshold. For demonstration purposes, the threshold for the methods that return a *distance* value presented here will be set to 0.1. Where applicable, the method will return a match if the distance $d(h_1, h_2) < 0.1$ is smaller than the threshold. If h_1 and h_2 are the same, $d(h_1, h_2) = 0$, i.e. the distance is minimal. For the application in an analysis tool, this value should be made adjustable by the user.

The methods were implemented for both the German and English hyleme data sets.

7.2.1 Full string match

The simplest matching method is the full string match. In this variant, only hylemes that are identical will be matched. It is the strictest method, because it does not allow any deviations between h_1 and h_2. A full string match is only a true match in every sense, if two variants of the same *Stoff* are aligned. Especially across cultures, two hylemes that are the same in their textual representation might not actually imply the same event, because common sense and real-life considerations only play a secondary role. Consider the following example:

1. NN goes into the netherworld. (Cultural background: Greek mythology)

2. NN goes into the netherworld. (Cultural background: ANES)

The two hylemes are identical in their lexical representation. However, based on their narrative contexts they are different, because they refer to two fundamentally different ideas on what exactly the netherworld is.

7.2.2 Full lemmata match

The lemmata match is slightly more lenient than the full string match. In this approach, the hyleme text is tokenized and lemmatized, using the spaCy-internal tokenizer and lemmatizer. h_1 and h_2 are then compared based on the lists of lemmata. If the lists of lemmata, including repetitions (e.g. of articles), are the same, the method returns a match. No stopword removal is performed. For example, the following two hylemes would be matched:

1. The child returns with the bull.
2. The children return with the bulls.

However, the hylem *The child returns the bull* or *The bull child returns*[8] would not be matched, because lemmata are missing.

7.2.3 Lemmata intersection

A variant of the lemmata match is the lemmata intersection. In this method, the sets of lemmatas of h_1 and h_2 are compared, i.e. no repetitions are preserved. The method returns a match if the intersection between the two set of lemmata is at least as long as the smaller of the both sets. This method can identify matches where one hyleme is an extended version of a second hyleme, for example:

1. The man looks strong and youthful.
2. The man looks youthful.

Hence, the lemmata intersection will consider *Addition/Deletion* effects to be similar. Additionally, it will return a match for all combinations of hylemes from the previous section including *The bull child returns.*

7.2.4 Jaccard Distance

The Jaccard distance, as introduced in Section 2.3.1, was implemented in three variants: matching all tokens in a hyleme (J_{full}), matching all tokens without stopwords (J_{stop}), and matching all lemmata of a hyleme (J_{lemma}). The Jaccard distance was implemented using the NLTK metrics module.[9] It is a fast and simple way to calculate the distance between two sets.

However, Jaccard distance has disadvantages with regards to hylemes. Mainly, it cannot distinguish between different elements of a hyleme, which according to the hylistic theory might have different influence on the hyleme similarity. Additionally, in cases where one hyleme has a predicate determination that consists of a number of tokens, the similarity can be skewed. Consider the following example,

[8] In this context, a *bull child* could refer to a child who assumed the form of a bull or the child of a bull, a calf.
[9] https://www.nltk.org/api/nltk.metrics.distance.html#module-nltk.metrics.distance

1. The child rides the chief's ox.
2. The child rides the chief's ox in the mountains.

The Jaccard distance of the two hylemes is $J_{full} = 0.22$, $J_{stop} = 0.14$, and $J_{lemma} = 0.14$. If we assume a distance threshold of 0.1, none of the three methods returns a match between the two hylemes. Additionally, the Jaccard Distance is influenced by the length of the hyleme. That means that the same deviation is punished more severely in shorter hylemes than in longer hylemes.

Furthermore, the Jaccard Distance is applied to sets of tokens in a hyleme, which means that two hylemes where hyleme subject and object are switched will be assumed to be the same.

1. Usikulumi hides Unthlatu.
2. Unthlatu hides Usikulumi.

In this case, the Jaccard Distance would be 0 in all three variants, i.e. the method assumes the two hylemes to be exactly the same.

7.2.5 Predicate match

Another matching method compares the hyleme predicates[10] of h_1 and h_2. It does not include other hyleme components. Using this method, we can align actions or events rather than entire hylemes. This is useful if the user is interested in finding occurrences of the same action in sequences that describe variants of different myths, e.g. finding instances where a character performs a ritual or uses a specific mode of transport.

1. Kain slays Abel.
2. Lamaštu slays the worm.

7.2.6 Predicate synonym match

The synonym match is the most lenient hyleme match. In this case, we match h_1 and h_2 if their predicate share a synonym. This way actions, and events that are expressed in different ways but essentially mean the same can be matched. The synonym match was implemented for English hylemes using NLTK's WordNet implementation[11], and for German using GermaNet through the Python library *germanetpy*[12]. For example, the German verbs *aufgehen* (engl. 'to open', something opens by itself, e.g. a blossom) and *aufmachen* (engl. 'to open sth'), share a synonym *öffnen* ('to open sth. up').

This allows to match hylemes that use predicates which have different word senses, like in this case *aufmachen* can also mean 'to start a journey' and *aufgehen* can also refer to celestial bodies ('Die

[10] Lemma matching has previously been used as a baseline for predicate-argument alignment, by [97] and [98]
[11] `https://www.nltk.org/howto/wordnet.html`
[12] `https://pypi.org/project/germanetpy/`

Sonne geht auf.', engl. 'The sun rises.')

7.2.7 Strict predicate synonym match

The predicate synonym match delivers a relatively high number matches, because it will align every occurrence of a match across word senses. A slightly stricter variant of this method yields a match if the predicate lemma of h_1 is part of the synset of the predicate of h_2, or vice versa.

In this case, only the word sense 'to start a journey' would be matched, when *aufbrechen* and *aufmachen* are compared (not 'to open/break open something'), because they are both lexical units of the synset *s58082* in GermaNet.

1. Orpheus bricht zur Unterwelt auf. (engl. 'Orpheus starts his journey to the netherworld.')
2. Orpheus macht sich zur Unterwelt auf. (engl. 'Orpheus starts his journey to the netherworld.')

7.2.8 Hyleme-SP/O match

Another matching method is sensitive to the hyleme components, namely hyleme subject, hyleme predicate, and hyleme object. This approach is called Hyleme-SPO[13]. Similarly, we can match hylemes by subject and predicate, without the hyleme object. The first method returns closer matches, while the second variant is useful for exploratory alignments, such as:

1. The child approaches the kraal.
2. The child approaches his mother.

We match the three components of h_1 and h_2 against each other. Deviations of hyleme component determinations will be neglected. If the SP resp. SPO-structure of h_1 and h_2 match, the method returns a match.

7.2.9 Resolved Entities

In this approach, named entities in hylemes are resolved using information provided about them in the sequence. For that purpose, we exploit the hyleme type information. All *durative-constant* hylemes in the sequences S_1 and S_2, where $h_1 \in S_1$ and $h_2 \in S_2$ are parsed. If any *durative-constant* hyleme contains a subject complement and the predicate is a form of *to be*, all occurrences of the named entity in the sequence are resolved with the subject complement. After the resolution of the named entities, h_1 and h_2 are compared using the *full string match*. If no resolution is found, the method defaults to the regular *full string match*.

[13]If the two hylemes do not have objects, this method returns a match. If one hyleme has an object but the other one does not have an object, it does not return a match.

With this approach, we can match characters with different named that fulfill the same roles. For instance,

1. Usikulumi is the king. (durative-constant) Usikulumi kills the cannibal.

2. The king kills the cannibal.

The *durative-constant* hyleme would invoke the entitiy resolution of *Usikulumi* with *the king*. Afterwards, the two hylemes would be matched accordingly.

7.2.10 Domain-Specific Pretrained Embeddings

Two embeddings models, Word2Vec and FastText (see Section 2.3), have been trained on domain relevant data for the German and English data sets. The German data consists of freely available texts related to the mythological domain, e.g. from Wikipedia articles on various types of deities, but also plot discription of movies with mythological content. Additionally, some translations of the source texts from which the hylemes in the data set were derived, such as the Homeric Hymns, were used. The data has been pre-processed by removing special characters and line breaks. Afterwards, it was tokenized and lemmatized using the spaCy pipeline.

For the English data set, the OCR text of Callaway's Zulu folktales was used. For that purpose, each page was processed using the OCR module tesseract[14]. Then the text was parsed and the Python language detection module langdetect[15] detected the English parts of the text. This text was then pre-processed, by removing line breaks and special characters, before sentence tokenization and lemmatization using spaCy.

Word2Vec models work well if a word is present in the vocabulary. However, since the hyleme data has a lot of specific vocabulary that might not even be present in domain-specific texts, the FastText model is used as a back-up model. Whenever a term is not present in the Word2Vec model vocabulary, the FastText model is used. FastText is able to query for subword information. Therefore, the similarity between two FastText word vectors is inherently more based on the morphological structure of a word, and less on semantics (although morphology influences semantics).

Since both embedding methods query for words, and not sentences, we selected the hyleme predicate as the part of the hyleme on which the match will be performed. The embeddings are queried for the predicates of h_1 and h_2. The cosine distance between the word vectors corresponding to the predicates of h_1 and h_2 is calculated. A match is returned if the cosine distance is below the threshold.

[14]https://pypi.org/project/pytesseract/
[15]https://pypi.org/project/langdetect/

7.2.11 TF-IDF cosine distance

For a TF-IDF-based matching method, the hylemes h_1 and h_2 are vectorized in a shared TF-IDF vector space. This way, it remains ensured that $h_1 = h_2$ yields a value of 0, in contrast to vectorizing them separately among the hylemes of S_1 and S_2. A match is returned if the cosine distance between the two vectors is smaller than the threshold. It would be possible to create a shared TF-IDF vector space for both sequences, and then calculate the cosine distance between h_1 and h_2. This would ensure that $h_1 = h_2$ yields a value of 0. However, since some sequences are relatively long, the processing time of this approach would make up for the slight improvement in accuracy.

7.2.12 Sentence Embeddings

We can also match hylemes using a sentence transformer embeddings approach. For this purpose, the the sentence transformer model *distiluse-base-multilingual-cased-v2* [183] is used. Embeddings are encoded for each hyleme sequence. Then, cosine distance scores are calculated for each pairing h_1 and h_2 of hylemes. The method returns all pairings $h_1 \in S_1$ and $h_2 \in S_2$ where the cosine similarity (i.e. $1 - cosine\_score(h_1, h_2)$) is below the threshold.

7.2.13 Combining methods

The presented methods can be combined to return a set of different alignment candidates. Figure 7.4 shows an example from two sequences in the English data set. For demonstration purposes, we include a subset of methods: the lemmata intersection match, the predicate match and the Hyleme-SP match. The predicate match returns four different candidates for alignment. However, since the method is relatively lenient, the pairings might not always be meaningful, e.g. ‘The mother asks NN to come and listen to the unborn child.’ and ‘The father asks for proof.’ The Hyleme-SP match returns a subset of the candidates of the predicate match. Hence, it is slightly more expressive. The lemmata intersection method is the only method which returns the candidate pair ‘The child demands to be born at once.’ and ‘The child repeats his demand to be born at once’.

7.3 Results

The different matching approaches for $h_1 \in S_1$ and $h_2 \in S_2$ are presented in Table 7.7. Three important factors play a role in determining the validity of matching methods. Firstly, some methods are sensitive to the SPO structure of the hyleme, i.e. the method does not use a *bag-of-X* (e.g. words, or lemmata) approach. Therefore, *Usikulumi hides Unthlatu* and *Unthlatu hides Usikulumi* will be treated differently. Secondly, a method may or may not take (hyleme subject, predicate or object) determinations into consideration. The importance of determinations might differ from use case to use case. For instance, a user might want to align all hylemes that use a variant of *to sail*, without considering different determinations, such as *alone*, *on the shore*, or *through*

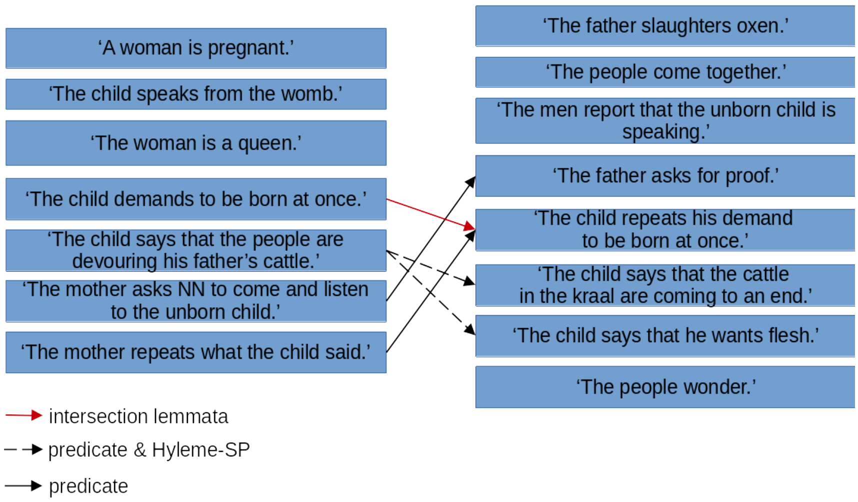


Figure 7.4: Example alignment candidates identified by three different methods, dashed line indicates match returned by both methods

a storm. Thirdly, methods that are computationally expensive have a disadvantage if sequences are particularly long, or if they are applied alongside other methods. Namely the embeddings methods suffer from a high processing time, because embeddings need to be encoded (sentence embeddings) or word vectors need to be accessed (pre-trained embeddings), and cosine scores need to be calculated for each possible pairing of $h_1 \in S_1$ and $h_2 \in S_2$. To measure computational time, a subset of one hundred sequences from the English data set was processed. The processing time in seconds in presented in Table 7.7. The results vary across methods. The fastest approaches are the predicate match, and the Hyleme-SP/O matching method. The pretrained embeddings approach is naturally the slowest, for the aforementioned reasons.

Additionally, we can investigate the threshold methods which are applied to pairs of hylemes, namely Jaccard, and TF-IDF, against the rating of manual annotations. For this purpose, all items in the gold standard annotation (four pairs of hylemes each), where processed using the threshold methods, and ranked according to their similarity. These similarity rankings can be matched against the manual annotations. The sentence embeddings method cannot be directly applied to the gold standard annotation, because h_1 and h_2 have no corresponding hyleme sequences. The method creates embeddings for all the hylemes in a sequence. Hence, embedding and comparing two individual hylemes separately would not produce the same results as intended by the approach. From Tables 7.8 and 7.9, we see that the full Jaccard methods produces the gold standard alignment once, the stopwords variants twice, and the Jaccard lemma variant four times. The TF-IDF cosine distance produces the gold standard ranking twice out of 27 4-tuples. The performance of the ranking-producing methods (Jaccard and TF-IDF) is measured as precision by position, i.e. how often the method predicts the correct ranking position (*best*, *worst* or *rank 2/3*. The precision values for the different methods is given in Table 7.10. The Jaccard lemma method works best in reproducing the ranking of best and worst fits of hyleme pairs. The results show only a marginal improvement over random chance (which would be 0.25 for best and worst, 0.50 for rank 2/3). However, given the complexity of the annotation task, demonstrated by the relatively low inter-rater agreement ($\kappa = 0.3162$), it becomes clear that hyleme similarity is highly subjective. Therefore, it has to be approached by different similarity measures as introduced in Section 7.2, depending on the research question and preference of the user.

Table 7.7: Comparison of different hyleme matching methods

Method	SPO-sensitive	incl. determinations	Processing time (s, n=100)
Full string	yes	yes	.0059
Full lemmata	no	yes	.0866
Lemmata intersection	no	yes	.1022
Jaccard Distance (full)	no	yes	.1628
Jaccard Distance (lemma)	no	yes	.5382
Jaccard Distance (stopwords)	no	yes	.5542
Hyleme-SPO	yes	no	.0052
Hyleme-SP	yes	no	.0045
Resolve entities	yes	yes	.0070
Predicate match	no	no	.0053
Predicate synonym match	no	no	.4082
Predicate synonym match (strict)	no	no	.3554
Domain-specific Pretrained Embeddings	no	no	90.4864
TF-IDF cosine	no	yes	28.6789
Sentence Embeddings cosine	no*	yes	34.3915

(*) Strictly speaking, sentence embeddings do not capture the SPO-structure of a hyleme. However, since embeddings are created from entire sentences, the position of a word in a sentence and grammatical function will influence the embedding. Therefore, *Usikulumi hides Unthlatu* and *Unthlatu hides Usikulumi* will not be treated as the same hyleme.

7.4 Summary

This chapter presents approaches towards identifying possible candidates for hyleme alignment. To that end, gold standard annotations were performed on pairs h_1 and h_2 of hylemes, where h_2 is $t(h_1)$, and t is one of various possible transformations. The gold standard annotation consists of a best-worst-ranking between 4 pairs of hylemes (4-tuples). The best fit is the pair h_1 and h_2 with the highest perceived semantic proximity. The gold standard data set consists of 27 4-tuples, combining different types of transformation effects. 20 annotators provided annotations for the task. Fleiss kappa indicates a fair agreement of $\kappa = 0.3162$.

The gold standard annotation was performed on hylemes derived from the German data set. To our knowledge, it is the first, albeit small, data set on semantic proximity annotations for German in the mythological domain.

For hyleme alignment, whether or not two hylemes h_1 and h_2 can be considered a match may differ depending on the user's research question. Researchers from different backgrounds might consider a match based on the taxonomic relation, e.g. hypernomy, between hyleme components, or only if a full string match, i.e. the hylemes are exactly the same in every aspect, occurs. Lenient methods, such as predicate matches, are well suited for exploratory alignments, where two or more sequences are not directly related. Stricter methods, such as full string matches, or SPO matches, are better suited for the alignment of sequences that pertain to the same myth variant. Therefore, a variety of matching methods to identify possible candidates is presented in this chapter. The resulting candidates are fundamental for the alignment of hyleme sequences.

The matching approaches are evaluated based on three consideration: Firstly, is the matching method sensitive to the structure of the hyleme? Secondly, does it include information from hyleme component determinations? Lastly, how performant is the matching method in terms of computational time? In terms of computational time, the match of hyleme subject and predicate is the fastest (0.0045s for 100 sequences), while using domain-specific embeddings is the slowest (90.4864s for 100 sequences).

Table 7.8: Rankings produced by distance measures and gold standard, ordered closest to furthest/best to worst, 1 = “Pair A”, 2 = “Pair B”..., bold = ranking matches the gold standard, a/b = ranking of *a* equals ranking of *b*

4-Tuple	TF-IDF Cosine Distance	Jaccard (full, stop-words, lemma)	Gold Standard Rating
1	[1,2,4,3]	[1,2,3/4] [1/2,4,3] [1/2,4,3]	[4,2/3,1]
2	**[1,2,4,3]**	**[1,2/4,3]** **[1,2/4,3]** **[1,2/4,3]**	[1,2/4,3]
3	[3,4,2,1]	[3/4,2,1] [4,2,1/3] [4,1/3,2]	[3,1/2,4]
4	[1,3,4,2]	[1,4,2/3] **[1,2,3,4]** **[1,2,3,4]**	[1,2/3,4]
5	[1,4,3,2]	[1,3,4,2] [1,2/4,3] [1,2/4,3]	[2,1/4,3]
6	[1,2,4,3]	[1,2,4,3] [1/3,2/4] [1/3,2/4]	[4,1/2,3]
7	[3,2,1,4]	[2/3,1,4] [1,3,2/4] [1,3,2/4]	[4,2/3,1]
8	[1,4,2,3]	[1,2/3/4] [1,4,2/3] [1,4,2/3]	[3,2/4,1]
9	[1,3,2,4]	[1,3,2,4] [1,3,2,4] [1,3,2,4]	[2,1/3,4]
10	[1,4,2,3]	[1,4,2,3] [1,2/4,3] [1,2/4,3]	[2,1/4,3]
11	[3,1,2,4]	[3,1,2/4] [3,1,4,2] [3,1/2,4]	[2,1/4,3]
12	[4,1,2,3]	[4,1,2,3] [1,4,2,3] [3,1,4,2]	[1,2/3,4]
13	[3,2,4,1]	[3,2,4,1] [3,1/2,4] [3,2/4,1]	[-]
14	[4,3,2,1]	[4,3,2,1] [4,3,2,1] **[4,3,1,2]**	[4,1/3,2]
15	[3,2,1,4]	[2,3,1,4] [1,2/3/4] [1,2/3/4]	[1,2/4,3]
16	[3,1,2,4]	[3,1,2,4] [1,3,2,4] [4,2,1,3]	[3,1/4,2]

Table 7.9: Rankings produced by distance measures and gold standard, ordered closest to furthest/best to worst, 1 = "Pair A", 2 = "Pair B"..., bold = ranking matches the gold standard, a/b = ranking of *a* equals ranking of *b*, Cont'd

4-Tuple	TF-IDF Cosine Distance	Jaccard (full, stop-words, lemma)	Gold Standard Rating
17	[3,2,4,1]	[3,2,1/4] [3,2,1,4] [1,2,3/4]	[2,1/3,4]
18	[4,1,2,3]	[4,1/2,3] [3,2,4,1] [2/3,4,1]	[2,1/3,4]
19	[3,2,1,4]	[3,2,4,1] [2,3,1,4] **[2,4/3,1]**	[2,3/4,1]
20	**[1,3,4,2]**	[3/4,1,2] [1,2,3,4] [1,2,3,4]	[1,3/4,2]
21	[1,2,4,3]	[1,2,4,3] [1,2,3,4] [1,2,4,3]	[2,1/3,4]
22	[1,2,3,4]	[1,2,3,4] [1/4,2,3] [1/4,2,3]	[3,2/4,1]
23	[4,3,1,2]	[4,1,3,2] [4,3,1/2] [4,3,1/2]	[1,2/4,3]
24	[3,1,4,2]	[1,3,2,4] [1,3,4,2] [1,3,4,2]	[1,2/4,3]
25	[3,2,4,1]	[3,2/4,1] [3,2/4,1] [2/3,4,1]	[2,1/3,4]
26	[1,3,2,4]	[1,3,2,4] [1/3,2,4] [1/3,2,4]	[-]
27	[3,1,2,4]	[3,1/2,4] [3,1,2,4] [4,3,1,2]	[2,1/3,4]

Table 7.10: Performance of the ranking-producing methods again gold standard hyleme similarity ranking

	TF-IDF	Jaccard (full)	Jaccard (stopwords)	Jaccard (lemma)	Support
best	0.24	0.21	0.32	0.33	25
worst	0.46	0.30	0.26	0.28	50
rank 2/3	0.24	0.40	0.51	0.55	25
weighted average	0.35	0.37	0.40	0.43	

Chapter 8

Conclusion

> "[...] the society that cherishes and keeps its myths alive will be nourished from the soundest, richest strata of the human spirit."
>
> J. Campbell, *Myths to live by* [20, p.24]

In this work, I present approaches towards a digital study of mythological and folkloristic content derived by means of the *Hylistic* approach. It is a first step towards *Digital Hylistics*.

In that, the focus of this work lies in the comparison of narrative variants (*Stoffe*). These variants are presented according to the *Hylistic approach* as so called *hyleme sequences*. The individual statements in those sequences are the *hylemes*, the basic plot units, e.g. "Harry Potter boards the Hogwarts express."

The first step in comparing those sequences of *hylemes* is to compare the individual *hylemes*. In which case two *hylemes* can be considered similar enough for alignment largely depends on the research question. If fine deviations are already considered meaningful, a stricter method for comparison such as the full string match, is needed. If *hylemes* can have slight deviations, e.g. different objects or determinations, a more lenient method can be employed. A set of these methods are presented in Section 7.2. When these methods are employed, a set of alignment candidates is derived. From those matching *hylemes*, an alignment of the sequences can be performed, e.g. by using the *Needleman-Wunsch* algorithm (global alignment). For long sequences, or when only a part of the narrative variants (e.g. only the ending) is to be used for the sequence comparison, local alignment can be performed by using the Smith-Waterman approach, see Section 6.3.

Additionally, this work presents the first large English hyleme data set (see Section 4.1). This data set is derived from 30 folktales from Henry Callaway's collection of isiZulu folktales [7]. In that, it is also the first *Hylistic* data set for the domain of Folkloristics. Callaway's collection is particularly interesting for folkloristic research, because it contains English and vernacular isiZulu text. Furthermore, the text is close to the oral narration, which allows for interesting further research (see below).

In this work, two annotation studies are presented: The annotation of *hyleme* types, and the annotation of *semantic proximity* between pairs of *hylemes*. While the first annotation study is relatively straightforward and resulted in a high overall inter-annotator agreement, the semantic proximity estimation is a particularly hard task. This result is interesting and useful for the interpretation of results delivered by automatic approaches.

In this thesis, I suggest that *hylemes* are treated differently according to their *hyleme*-type. While the plot-driving *single-event hylemes*, which contain mainly actions, e.g. "Harry Potter boards the Hogwarts express" are compared and subsequently aligned, *durative-constant hylemes*, which contain background information, states and habituals, are processed with knowledge engineering methods. In two case studies, we demonstrate the ideal shallow (i.e. flat hierarchy) ontologies for the *Orpheus and Eurydice* myth and on myth variants concerning the Mesopotamian deity *Dumuzi*. These hand-crafted shallow or minimal ontologies are a knowledge base for the representation of characters and other important concepts, their relations and attributes. Since relationships and other types of information can be contradictory in different myth variants, each *hyleme sequence* is represented by a dedicated ontology. These individual ontologies can be subsequently used for comparing the background information of the myth variants. For that purpose, class overlap and individual overlap is calculated (essentially using Jaccard similarity).

This work aimed to select all methods and approaches in a way to pay justice to the *Hylistic* theory and keep potential application by scholars working with narratives in mind.

8.1 Future Work

This work opens the field of studying *Hylistic* data from a number of different perspectives, through different methods. Firstly, the *Hylistic* approach does not only apply to myths and folklore, it can also be applied to a variety of other narrative domains and genres. Any source material that contains a form narrative can be modelled as a *hyleme sequence*. For instance, *hyleme sequences* can be applied to fiction and fanfiction, with subsequent comparison of variants of both through the methods proposed in this work. *Hylistic* analysis can also be applied to various narrative forms of media, not only textual sources, for instance *political caricatures*.

The *hyleme* alignments that are derived using the methods proposed in Chapter 7.2 can be interpreted as the common narrative of two or more sequences. In turn, those common narratives can be used as input to downstream tasks, such as Semantic Overlap Summarization (SOS), for instance following the approach by Bansal et al. [184].

Another interesting future direction of *Digital Hylistics* is to (automatically or semi-automatically) identify the causal links between *hylemes*, especially *single-event* and *durative-resultative* hylemes, e.g. "Eurydice dies" (single-event) and "Orpheus is heartbroken" (durative-resultative). This task is challenging because *hyleme* sequences do not use any discourse markers that would indicate the causal link (such as *because*). Therefore, such a study would have to be modelled completely through knowledge-based and semantic methods.

Hyleme sequences would also be an ideal basis for the application in language education. For instance, in studies of reading comprehension, language learners can be asked to identify where a *hyleme* is entailed in a text, or to order hyleme sequences according to the order in the text and the order in the narrative. Additionally, by leaving out one *hyleme* from the sequence, logical reasoning can be assessed when learners are asked to predict the missing *hyleme* (*narrative cloze test*).

In order to create *hyleme*-annotated resources, it would be an interesting task to (semi-automatically) map *hylemes* back to their textual representation in the source material. This would include the subtask of identifying the right context-windows in the source. The *hyleme sequence*-to-text mappings would be a very useful basis for digital editions of ancient sources, including the *Stoff*-layer. A working tool for this purpose would also be interesting for the data set derived from the Zulu folktales, because it can serve as a parallel edition for a low-resource language (English-isiZulu). Particularly for scholars of historical isiZulu, this resource would be of great value.

For the purpose of creating these annotated resources, an XML-TEI schema for *Hylistic* in-text annotations has been developed[1], but not yet employed. This annotation would have to rely on domain experts, espcially for the annotation of *implicit hylemes*.

The task to automatically identify *hylemes* from a text will remain challenging for certain domains, mainly in *mythological* and *folkloristic* materials. However, for domains where aspects of the narrative is less often implied and series of events are told in a more straightforward manner, e.g. news texts, this could be more easily achievable.

Previous works on identifying predicate argument structures (PAS) from text, which include filtering events and non-events, could serve as a starting point. In hylistic terms, all *single-event* hylemes are events of some sort. The task to derive a hyleme sequence from a source can be assisted if suitable language models are available for the source language, such as Ancient Sumerian. However, for the mythological domain the task will always have to rely on informed scholars, as well as scholarly discussion. Large Language Models (LLMs) have shown great capacity for inferring knowledge and common sense reasoning in texts, but for the research objective at hand, expert knowledge is crucial to determine how to transform a source statement into an appropriate hyleme. Hylemes are more than text summaries, or mere descriptions of vase paintings. They include interpretations of the source that are beyond common sense and everyday knowledge. Moreover, background information and plots of mythological and folkloristic material may be

[1] I thank Uwe Sikora and Florian Barth for their help in creating the XML-TEI schema.

even contradictory to common sense reasoning of the real world, i.e. humans cannot give birth to children who are not human, rendering selectional preferences of arguments, e.g. in frame semantic approaches, largely moot for the application in the mythological domain.

A completely new task of measuring similarity in *hylemes* would be to consider instances where *hylemes* can be matched semantically within the context of the *narrative* variant but do not follow basic semantic similarity. In these cases, a match between two *hylemes* is only valid within the context of the *Stoff*, e.g. "Eurydice receives a tender breath of life" and "The gods replace Eurydice's threads of life".

Lastly, the *Hylistic* theory can be combined with other narrative theories. For instance, if Afanasyev's Russian Magic Tales [61] were used as a basis for *Hylistic* analysis, the Proppian Morphology as introduced in 3.1 and its functions of the Folk tale could be interpreted as *hyper-hylemes*. Other efforts could include bridging *Hylistics* and folktale motif indices, such as the Thomson Motif Index (TMI), by investigating which motifs follow a *hyleme* structure, and by investigating if an automatic assignment of motifs to *hyleme sequences* can be solved as a classification task.

Chapter 9

Acknowledgements

سُبْحَانَكَ اللَّهُمَّ رَبَّنَا وَبِحَمْدِكَ اللَّهُمَّ اغْفِرْ لِي

I would like to express my gratitude to my supervisor, Prof. Caroline Sporleder, who always encourages me in my academic endeavours without question or hesitation. For her years of continuous professional and personal support she has my deepest gratitude. I also thank Prof. Ramin Yahyapour for agreeing to serve as co-referent to this thesis, and for his helpful comments during the thesis committee meetings. I would also like to thank Prof. Christian Zgoll, for his support, all his insights, and especially for his open-mindedness to let me do all sorts of things with his theory. Furthermore, I am grateful to Prof. Carsten Damm, Prof. Wolfgang May, and Prof. Bela Gipp for agreeing to be part of my thesis examination board.

Prof. Annette Zgoll, whose passion for everything mythical is contagious, Dr. Josephine Fechner, Dr. Brit Kärger, Dr. Balbina Bäbler, and all other members of the STRATA family have my gratitude. I would also like to thank Hanna Varachkina, who always had an open ear and open mind for me. I also thank Stefan Ziehe, Tillmann Dönicke, Theresa Blaschke, Lucy Ritter and my other colleagues for their support. Ms Petra Kemmling, Ms Heike Jachinke, Ms Tina Bockler and all the other members of the administration have my special gratitude for 'keeping the ship running'.

I am grateful to my family for believing in me, namely my parents and my grandparents, to whom this work is dedicated, since it was their stories which laid the foundation to this thesis. Lastly, and most importantly, want to thank my husband, Yasar Mahomed Abbas, who loves me in my best and my worst moments, whose patience never wavers, and without whose ongoing encouragement and support this work would not have been possible.

Bibliography

[1] C. Zgoll, *Tractatus mythologicus*. Berlin, Boston: De Gruyter, 2019. [Online]. Available: https://doi.org/10.1515/9783110541588

[2] M. Moens and M. Steedman, "Temporal ontology and temporal reference," *Computational Linguistics*, vol. 14, no. 2, pp. 15–28, 1988. [Online]. Available: https://aclanthology.org/J88-2003

[3] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. New York, NY, USA: Cambridge University Press, 1998.

[4] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation," in *Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2017, pp. 165–174. [Online]. Available: https://doi.org/10.1109/DSAA.2017.61

[5] A. Scheidel, *An Augmented Annotation Scheme for Fairy Tales*. Bachelor Thesis, Saarland University, 2010.

[6] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of the Association for Computational Linguistics-08: Human-Language-Technology (HLT)*. Columbus, Ohio: Association for Computational Linguistics, Jun. 2008, pp. 789–797. [Online]. Available: https://aclanthology.org/P08-1090

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

[8] A. Zgoll, B. Cuperly, and A. Cöster-Gilbert, "In search of Dumuzi: An introduction to Hylistic narratology," in *The Shape of Stories*. Leiden: Brill, 2023, pp. 285–350. [Online]. Available: https://doi.org/10.1163/9789004539761_013

[9] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[10] W. Burkert, "Literarische Texte und Funktionaler Mythos: Zu Ištar und Atrahasis," in *Funktionen und Leistungen des Mythos: Drei altorientalische Beispiele*, J. Assmann, W. Burkert, and F. Stolz, Eds., 1982, p. 63.

[11] S. Albert, J. Anderssen, R. Bader, S. Becker, T. Bracht, S. Brants, T. Brants, V. Demberg, S. Dipper, P. Eisenberg *et al.*, *TIGER Annotationsschema*. Universität des Saarlandes and Universität Stuttgart and Universität Potsdam, 2003.

[12] R. A. Wilcox and E. M. Whitham, "The symbol of modern medicine: why one snake is more than two," *Annals of Internal Medicine*, vol. 138, no. 8, pp. 673–677, 2003.

[13] G. A. Caduff, *Antike Sintflutsagen*, ser. HYPOMNEMATA: Untersuchungen zur Antike und zu ihrem Nachleben, A. Dihle, H. Erbse, C. Habicht, H. Lloyd-Jones, G. Patzig, and B. Snell, Eds. Göttingen: Vandenhoeck und Ruprecht, 1986, vol. 82.

[14] C. Lévi-Strauss, "The structural study of myth," *The Journal of American folklore*, vol. 68, no. 270, pp. 428–444, 1955.

[15] E. Götting, "Lamaštu: Ikonographie einer altorientalischen Dämonin," Magisterarbeit, Freie Universität Berlin, 2009.

[16] W. Farber, "Lamaštu—agent of a specific disease or a generic destroyer of health?" in *Disease in Babylonia*. Leiden: Brill, 2007, pp. 137–145.

[17] C. Lévi-Strauss, *Mythos und Bedeutung: Vorträge*. Berlin: Suhrkamp, 1995.

[18] B. B. Serekoane, *Tricks, Trickster Characters and Trickster Tales in Tswana Folktales*. Master's Thesis University of Johannesburg, South Africa, 2012.

[19] J. Pieterse, "Trickster tropes: Female storytelling and the re-imagination of social orders in four nineteenth-century Southern African communities," *Historia*, vol. 55, no. 1, pp. 55–77, 2010.

[20] J. Campbell, *Myths to Live by: The Collected Works of Joseph Campbell*. Los Angeles, CA, USA: Joseph Campbell Foundation, 2011.

[21] C. Zgoll, "Myths as Polymorphous and Polystratic Erzählstoffe," in *Mythische Sphärenwechsel*, A. Zgoll and C. Zgoll, Eds. Berlin, Boston: De Gruyter, 2020, pp. 9–82. [Online]. Available: https://doi.org/10.1515/9783110652543-002

[22] D. Herman, *Story logic: Problems and possibilities of narrative*. Lincoln, Nebraska, USA: University of Nebraska Press, 2004.

[23] A. Zgoll and C. Zgoll, Eds., *Mythische Sphärenwechsel*. Berlin, Boston: De Gruyter, 2020. [Online]. Available: https://doi.org/10.1515/9783110652543

[24] G. I. Gabriel, B. Kärger, A. Zgoll, and C. Zgoll, Eds., *Was vom Himmel kommt: Stoffanalytische Zugänge zu antiken Mythen aus Mesopotamien, Ägypten, Griechenland und Rom*. Berlin, Boston: De Gruyter, 2021. [Online]. Available: https://doi.org/10.1515/9783110743005

[25] T. McEnery and A. Hardie, *Corpus Linguistics: Method, Theory and Practice*, ser. Cambridge Textbooks in Linguistics. New York, NY, USA: Cambridge University Press, 2011.

[26] B. Kortmann, "The triad "Tense-Aspect-Aktionsart": Problems and possible solutions," *Belgian Journal of Linguistics*, vol. 6, no. 1, pp. 9–30, 1991.

[27] C. Baker, "FrameNet, present and future," in *Proceedings of the 1st International Conference on Global Interoperability for Language Resources (ICGL-2008)*, J. Webster, N. Ide, and A. C. Fang, Eds., City University. Hong Kong: City University, 2008.

[28] A. Barr and E. A. Feigenbaum, *The Handbook of Artificial Intelligence, vol. 1*. Los Altos, California: William Kaufmann, 1981.

[29] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks." *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.89.22.10915

[30] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283670900574

[31] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, no. 3, pp. 705–708, 1982.

[32] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283681900875

[33] J. B. Kruskal, "An overview of sequence comparison: Time warps, string edits, and macromolecules," *Society for Industrial and Applied Mathematics (SIAM) Review*, vol. 25, no. 2, pp. 201–237, 1983.

[34] P. Jaccard, "Lois de distribution florale dans la zone alpine," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 38, pp. 69–130, 1902.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations*, 2014, pp. 55–60.

[36] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, 2013, artificial Intelligence, Wikipedia and Semi-Structured Resources. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370212000276

[37] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the Association for Computing Machinery (ACM)*, vol. 8, no. 10, p. 627–633, oct 1965. [Online]. Available: https://doi.org/10.1145/365628.365657

[38] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[40] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” *Advances in Neural Information Processing Systems*, vol. 29, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the International Conference on Learning Representations (ICLR’13)*, 2013.

[42] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[43] D. Jurafsky and J. H. Martin, *Speech and Language processing*, 3rd ed., 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[44] G. A. Miller, “WordNet: a lexical database for English,” *Communications of the Association for Computing Machinery (ACM)*, vol. 38, no. 11, pp. 39–41, 1995.

[45] C. Fellbaum, *WordNet: An electronic lexical database*. Cambridge, MA, USA: MIT press, 1998.

[46] B. Hamp and H. Feldweg, “Germanet-a lexical-semantic net for German,” in *Automatic information extraction and building of lexical semantic resources for NLP applications*, 1997.

[47] V. Henrich and E. W. Hinrichs, “GernEdiT-the GermaNet Editing Tool.” in *Proceedings of 48th annual meeting of the Association for Computational Linguistics: system demonstrations*, 2010, pp. 19–24. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf

[48] D. Sánchez, M. Batet, D. Isern, and A. Valls, “Ontology-based semantic similarity: A new feature-based approach,” *Expert Systems with Applications*, vol. 39, no. 9, pp. 7718–7728, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417412000954

[49] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and application of a metric on semantic nets,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[50] M. Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” in *Proceedings of the 5th Annual International Conference on Systems Documentation*, 1986, pp. 24–26.

[51] P. Resnik, “Using information content to evaluate semantic similarity,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montreal: IEEE Press, 1995, pp. 448–453.

[52] R. Mihalcea, C. Corley, C. Strapparava *et al.*, “Corpus-based and knowledge-based measures of text semantic similarity,” in *American Association for Artificial Intelligence (AAAI)*, vol. 6, 2006, pp. 775–780.

[53] D. M. Blei, “Probabilistic topic models,” *Communications of the Association for Computing Machinery (ACM)*, vol. 55, no. 4, pp. 77–84, 2012. [Online]. Available: https://doi.org/10.1145/2133806.2133826

[54] J. Grimm and W. Grimm, Eds., *Kinder- und Hausmärchen*, 7th ed. Göttingen: Verlag der Dieterichschen Buchhandlung, 1857.

[55] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[56] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the 8th Association for Computing Machinery (ACM) International Conference on Web Search and Data Mining (WSDM'15)*, 2015, pp. 399–408. [Online]. Available: https://doi.org/10.1145/2684822.2685324

[57] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[58] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, p. 378, 1971.

[59] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[60] V. Propp, *Morphology of the Folktale*. Austin, Texas, USA: University of Texas Press, 1968, vol. 10.

[61] A. N. Afanasyev, *Russian Folk-Tales*. New York: E. P. Dutton & Company, 1916.

[62] I. Okodo, "Obaraedo: Conformity to Proppian morphology," *AFRREV IJAH: An International Journal of Arts and Humanities*, vol. 1, no. 2, pp. 100–111, 2012.

[63] C. Azuonye, "Morphology of the Igbo folktale: Ethnographic, historiographic and aesthetic implications," *Folklore*, vol. 101, no. 1, pp. 36–46, 1990.

[64] F. Pannach, C. Sporleder, W. May, A. Krishnan, and A. Sewchurran, "Of Lions and Yakshis," *Semantic Web*, vol. 12, no. 2, pp. 219–239, 2021.

[65] D. L. Jones, *The television teen drama as folktale*. Master's Thesis, Nelson Mandela Metropolitan University, 2011.

[66] M. Kim, D. Srinivasan, and X. Zhou, "The morphology of dignity: Service storytelling and prototypes for a service design tool," *The Design Journal*, vol. 22, no. 6, pp. 793–812, 2019. [Online]. Available: https://doi.org/10.1080/14606925.2019.1662633

[67] A. Brusentsev, M. Hitchens, and D. Richards, "An investigation of Vladimir Propp's 31 functions and 8 broad character types and how they apply to the analysis of video games," in *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System*, ser. IE '12. New York, NY, USA: Association for Computing Machinery, 2012. [Online]. Available: https://doi.org/10.1145/2336727.2336729

[68] C. Lévi-Strauss, *Structural Anthropology, Volume 1*. New York: Basic Books, 1968.

[69] W. Labov and J. Waletzky, "Narrative analysis: oral versions of personal experience," in *Essays on the Verbal and Visual Arts, Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society, ed. June Helm (Seattle, 1967)*. University of Washington Press, 1967.

[70] W. G. Lehnert, "Plot units and narrative summarization," *Cognitive Science*, vol. 5, no. 4, pp. 293–331, 1981. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0504_1

[71] A. Goyal, E. Riloff, H. Daumé III, and N. Gilbert, "Toward plot units: Automatic affect state analysis," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 17–25. [Online]. Available: https://dl.acm.org/doi/pdf/10.5555/1860631.1860634

[72] A. Goyal, E. Riloff, and H. D. Iii, "A computational model for plot units," *Computational Intelligence*, vol. 29, no. 3, pp. 466–488, 2013. [Online]. Available: https://doi.org/10.1111/j.1467-8640.2012.00455.x

[73] R. Abelson and R. C. Schank, "Scripts, plans, goals and understanding: An inquiry into human knowledge structures," 1977.

[74] M. Regneri, A. Koller, and M. Pinkal, "Learning script knowledge with web experiments," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 979–988. [Online]. Available: https://aclanthology.org/P10-1100

[75] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 25–36, 2013. [Online]. Available: https://aclanthology.org/Q13-1003

[76] B. Jans, S. Bethard, I. Vulić, and M. F. Moens, "Skip N-grams and Ranking Functions for Predicting Script Events," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2012)*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 336–344. [Online]. Available: https://aclanthology.org/E12-1034

[77] M. Manshadi, R. Swanson, and A. S. Gordon, "Learning a probabilistic model of event sequences from internet weblog stories," in *Proceedings of the Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 2008, pp. 159–164. [Online]. Available: https://aaai.org/papers/flairs-2008-044/

[78] S. Ostermann, M. Roth, S. Thater, and M. Pinkal, "Aligning script events with narrative texts," in *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 128–134. [Online]. Available: https://aclanthology.org/S17-1016

[79] L. Wanzare, A. Zarcone, S. Thater, and M. Pinkal, "Inducing script structure from crowdsourced event descriptions via semi-supervised clustering," in *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–11. [Online]. Available: https://aclanthology.org/W17-0901

[80] A. Belyy and B. Van Durme, "Script induction as association rule mining," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, 2020, pp. 55–62. [Online]. Available: https://aclanthology.org/2020.nuse-1.7.pdf

[81] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen, "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 839–849. [Online]. Available: https://aclanthology.org/N16-1098

[82] Z. Li, X. Ding, and T. Liu, "Constructing narrative event evolutionary graph for script event prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 4201–4207. [Online]. Available: https://doi.org/10.24963/ijcai.2018/584

[83] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. Voss, "Connecting the Dots: Event graph schema induction with path language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 684–695. [Online]. Available: https://aclanthology.org/2020.emnlp-main.50

[84] Y. Zhou, X. Geng, T. Shen, J. Pei, W. Zhang, and D. Jiang, "Modeling event-pair relations in external knowledge graphs for script reasoning," in *Findings of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021, pp. 4586–4596. [Online]. Available: https://aclanthology.org/2021.findings-acl.403

[85] N. Weber, R. Rudinger, and B. Van Durme, "Causal inference of script knowledge," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7583–7596. [Online]. Available: https://aclanthology.org/2020.emnlp-main.612

[86] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative schemas and their participants," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (AFNLP)*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 602–610. [Online]. Available: https://aclanthology.org/P09-1068

[87] M. Granroth-Wilding and S. Clark, "What happens next? Event prediction using a compositional neural network model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016. [Online]. Available: https://doi.org/10.1609/aaai.v30i1.10344

[88] Z. Wang, Y. Zhang, and C.-Y. Chang, "Integrating order information and event relation for script event prediction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 57–67. [Online]. Available: https://aclanthology.org/D17-1006

[89] H. Kiyomaru, K. Omura, Y. Murawaki, D. Kawahara, and S. Kurohashi, "Diversity-aware event prediction based on a conditional variational autoencoder with reconstruction," in *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 113–122. [Online]. Available: https://aclanthology.org/D19-6014

[90] S. Yang, F. Wang, D. Zha, C. Xue, and Z. Tang, "NarGNN: Narrative graph neural networks for new script event prediction problem," in *2020 IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2020, pp. 481–488. [Online]. Available: https://ieeexplore.ieee.org/document/9443758

[91] A. Modi, "Event embeddings for semantic script modeling," in *Proceedings of The 20th Special Interest Group on Natural Language Learning (SIGNLL) Conference on Computational Natural Language Learning*, 2016, pp. 75–83. [Online]. Available: https://aclanthology.org/K16-1008.pdf

[92] I.-T. Lee and D. Goldwasser, "Feel: Featured event embedding learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. [Online]. Available: https://doi.org/10.1609/aaai.v32i1.11936

[93] N. Weber, R. Rudinger, and B. Van Durme, "Causal inference of script knowledge," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7583–7596. [Online]. Available: https://aclanthology.org/2020.emnlp-main.612

[94] S. Lv, F. Zhu, and S. Hu, "Integrating external event knowledge for script learning," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 306–315. [Online]. Available: https://aclanthology.org/2020.coling-main.27

[95] M. Aldawsari and M. Finlayson, "Detecting subevents using discourse and narrative features," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4780–4790. [Online]. Available: https://aclanthology.org/P19-1471

[96] G. Glavaš, J. Šnajder, M.-F. Moens, and P. Kordjamshidi, "HiEve: A corpus for extracting event hierarchies from news stories," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3678–3683. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1023_Paper.pdf

[97] M. Roth and A. Frank, "Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task," in **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Montréal, Canada: Association for Computational Linguistics, 7-8 Jun. 2012, pp. 218–227. [Online]. Available: https://aclanthology.org/S12-1030

[98] T. Wolfe, B. Van Durme, M. Dredze, N. Andrews, C. Beller, C. Callison-Burch, J. DeYoung, J. Snyder, J. Weese, T. Xu, and X. Yao, "PARMA: A predicate argument aligner," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 63–68. [Online]. Available: https://aclanthology.org/P13-2012

[99] T. Wolfe, M. Dredze, and B. Van Durme, "Predicate argument alignment using a global coherence model," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 11–20. [Online]. Available: https://aclanthology.org/N15-1002

[100] E. Hovy, T. Mitamura, F. Verdejo, J. Araki, and A. Philpot, "Events are not simple: Identity, non-identity, and quasi-identity," in *Workshop on Events: Definition, Detection, Coreference, and Representation*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 21–28. [Online]. Available: https://aclanthology.org/W13-1203

[101] J. Eisenberg and M. Sheriff, "Automatic extraction of personal events from dialogue," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Online: Association for Computational Linguistics, Jul. 2020, pp. 63–71. [Online]. Available: https://aclanthology.org/2020.nuse-1.8

[102] V. D. Lai, T. H. Nguyen, and F. Dernoncourt, "Extensively matching for few-shot learning event detection," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Online: Association for Computational Linguistics, Jul. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.nuse-1.5

[103] A. Judea and M. Strube, "Event extraction as frame-semantic parsing," in *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 159–164. [Online]. Available: https://aclanthology.org/S15-1018

[104] M. Vauth and E. Gius, "Richtlinien für die Annotation narratologischer Ereigniskonzepte," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5078175

[105] M. Vauth, H. O. Hatzel, E. Gius, and C. Biemann, "Automated event annotation in literary texts." in *Proceedings of the 2021 Computational Humanities Research Conference (CHR'2021)*, 2021, pp. 333–345. [Online]. Available: https://ceur-ws.org/Vol-2989/short_paper18.pdf

[106] E. Gius and M. Vauth, "Towards an event based plot model. A computational narratology approach," *Journal of Computational Literary Studies*, vol. 1, no. 1, 2022. [Online]. Available: https://doi.org/10.48694/jcls.110

[107] A. Baikadi and R. E. Cardona-Rivera, "Towards finding the fundamental unit of narrative: A proposal for the narreme," in *Proceedings of the The Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27, 2012*, 2012, pp. 44–46. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[108] F. Pannach, ""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)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10–18. [Online]. Available: https://aclanthology.org/2023.law-1.2

[109] J. Pustejovsky, R. Ingria, R. Saurí, J. M. Castaño, J. Littman, R. J. Gaizauskas, A. Setzer, G. Katz, and I. Mani, "The specification language timeML," in *The Language of Time - A Reader*, I. Mani, J. Pustejovsky, and R. Gaizauskas, Eds., 2005.

[110] J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí, "Temporal and event information in natural language text," *Language resources and evaluation*, vol. 39, pp. 123–164, 2005. [Online]. Available: https://doi.org/10.1007/s10579-005-7882-7

[111] M. Ocal, A. Perez, A. Radas, and M. Finlayson, "Holistic evaluation of automatic timeML annotators," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, 2022, pp. 1444–1453. [Online]. Available: https://aclanthology.org/2022.lrec-1.155.pdf

[112] E. Kearns, "Annotating and quantifying narrative time disruptions in modernist and hypertext fiction," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Online: Association for Computational Linguistics, Jul. 2020, pp. 72–77. [Online]. Available: https://aclanthology.org/2020.nuse-1.9

[113] H. Peng, S. Chaturvedi, and D. Roth, "A joint model for semantic sequences: Frames, entities, sentiments," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 173–183. [Online]. Available: https://aclanthology.org/K17-1019

[114] A. Alvarez and J. Font, "TropeTwist: Trope-based narrative structure generation," in *Proceedings of the 17th International Conference on the Foundations of Digital Games*, ser. FDG '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3555858.3563271

[115] P. Papalampidi, F. Keller, L. Frermann, and M. Lapata, "Screenplay summarization using latent narrative structure," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1920–1933. [Online]. Available: https://aclanthology.org/2020.acl-main.174

[116] N. Reiter, O. Hellwig, A. Frank, I. Gossmann, B. M. Larios, J. Rodrigues, and B. Zeller, "Adapting NLP tools and frame-semantic resources for the semantic analysis of ritual descriptions," in *Language Technology for Cultural Heritage*, C. Sporleder, A. van den Bosch, and K. Zervanou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 171–193. [Online]. Available: https://doi.org/10.1007/978-3-642-20227-8_10

[117] N. Reiter, "Discovering structural similarities in narrative texts using event alignment algorithms," Ph.D. dissertation, University of Heidelberg, 2014.

[118] M. Aldawsari, E. Asgari, and M. A. Finlayson, "Story fragment stitching: The case of the story of Moses," in *1st Workshop on Artificial Intelligence for Narratives (AI4N 2020)*, 2021. [Online]. Available: https://ceur-ws.org/Vol-2794/paper9.pdf

[119] O. O. Kwong, "Annotating the structure and semantics of fables," in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp. 275–282. [Online]. Available: https://aclanthology.org/Y11-1029.pdf

[120] P. M. Broadwell and T. R. Tangherlini, "Trollfinder: Geo-semantic exploration of a very large corpus of danish folklore," in *Proceedings of the The Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27, 2012*, 2012, pp. 50–57. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[121] O. Ninan and O. Odejobi, "Towards a digital resource for african folktales," in *Proceedings of the Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27, 2012*, 2012, pp. 77–81. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[122] M. M. A. Finlayson, "Learning narrative structure from annotated folktales," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[123] M. A. Finlayson, "Inferring Propp's functions from semantically annotated text," *Journal of American Folklore*, vol. 129, no. 511, pp. 55–77, 2016.

[124] F. Karsdorp and A. Van den Bosch, "The structure and evolution of story networks," *Royal Society Open Science*, vol. 3, no. 6, p. 160071, 2016. [Online]. Available: https://doi.org/10.1098/rsos.160071

[125] W. V. H. Yarlott and M. A. Finlayson, "Learning a better motif index: Toward automated motif extraction," in *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2016/6708/pdf/OASIcs-CMN-2016-7.pdf

[126] F. Karsdorp, M. van der Meulen, T. Meder, and A. van den Bosch, "MOMFER: A search engine of thompson's motif-index of folk literature," *Folklore*, vol. 126, no. 1, pp. 37–52, 2015. [Online]. Available: https://doi.org/10.1080/0015587X.2015.1006954

[127] S. A. Malec, "Proppian structural analysis and XML modeling," *Proceedings of Computers, Literature and Philology (CLiP 2001)*, 2001.

[128] T. Declerck, A. Scheidel, and P. Lendvai, "Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales," in *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*. Springer, 2011, pp. 155–170.

[129] P. Lendvai, T. Declerck, S. Darányi, and S. Malec, "Propp revisited: Integration of linguistic markup into structured content descriptors of tales," in *Digital Humanities 2010, London, United Kingdom*, 2010. [Online]. Available: https://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-753.html

[130] P. Lendvai, T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado, "Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/654_Paper.pdf

[131] A. Scheidel and T. Declerck, "APftML-augmented Proppian fairy tale markup language," in *Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, 2010, p. 95.

[132] W. V. H. Yarlott and M. A. Finlayson, "ProppML: A complete annotation scheme for Proppian morphologies," in *7th Workshop on Computational Models of Narrative (CMN 2016)*, 2016. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2016/6709/pdf/OASIcs-CMN-2016-8.pdf

[133] S. Malec, "AutoPropp: Toward the automatic markup, classification, and annotation of Russian magic tales," in *Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, 2010, p. 68.

[134] J. Valls-Vargas, J. Zhu, and S. Ontanón, "Toward automatic role identification in unannotated folk tales," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 10, no. 1, 2014, pp. 188–194. [Online]. Available: https://doi.org/10.1609/aiide.v10i1.12732

[135] L. Jahan, R. Mittal, W. V. Yarlott, and M. Finlayson, "A straightforward approach to narratologically grounded character identification," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6089–6100. [Online]. Available: https://aclanthology.org/2020.coling-main.536

[136] M. A. Finlayson, "ProppLearner: Deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory," *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 284–300, 12 2015. [Online]. Available: https://doi.org/10.1093/llc/fqv067

[137] L. Jahan, G. Chauhan, and M. Finlayson, "A new approach to Animacy detection," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1–12. [Online]. Available: https://aclanthology.org/C18-1001

[138] E. Gianitsos, T. Bolt, P. Chaudhuri, and J. P. Dexter, "Stylometric classification of ancient greek literary texts by genre," in *Proceedings of the Third Joint Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities (SIGHUM) Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, 2019, pp. 52–60. [Online]. Available: https://aclanthology.org/W19-2507

[139] A. Sahala, "Contributions to computational assyriology," Ph.D. dissertation, Helsingin yliopisto, 2021.

[140] P. Gervás, "Computational approaches to storytelling and creativity," *AI Magazine*, vol. 30, no. 3, pp. 49–49, 2009.

[141] A. Alabdulkarim, S. Li, and X. Peng, "Automatic story generation: Challenges and attempts," in *Proceedings of the Third Workshop on Narrative Understanding*. Online: Association for Computational Linguistics, Jun. 2021, pp. 72–83. [Online]. Available: https://aclanthology.org/2021.nuse-1.8

[142] F. Peinado, P. Gervás, and B. Díaz-Agudo, "A description logic ontology for fairy tale generation," in *Proceedings of the Workshop on Language Resources for Linguistic Creativity, Conference on Language Resources and Evaluation (LREC)*, vol. 4, 2004, pp. 56–61.

[143] P. Gervás, "Propp's morphology of the folk tale as a grammar for generation," in *2013 Workshop on Computational Models of Narrative (CMN 2013)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2013/4156/pdf/p106-gervas.pdf

[144] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 1–8. [Online]. Available: https://aclanthology.org/S14-2001

[145] I. Gurevych and H. Niederlich, "Computing semantic relatedness in German with revised information content metrics," in *Proceedings of OntoLex 2005-Ontologies and Lexical Resources*, 2005.

[146] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," *Association for Computing Machinery (ACM) Computing Surveys*, vol. 54, no. 2, feb 2021. [Online]. Available: https://doi.org/10.1145/3440755

[147] I. Vulić, S. Baker, E. M. Ponti, U. Petti, I. Leviant, K. Wing, O. Majewska, E. Bar, M. Malone, T. Poibeau, R. Reichart, and A. Korhonen, "Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity," *Computational Linguistics*, vol. 46, no. 4, pp. 847–897, 02 2020. [Online]. Available: https://doi.org/10.1162/coli_a_00391

[148] S. Kiritchenko and S. Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 465–470. [Online]. Available: https://aclanthology.org/P17-2074

[149] S. Asaadi, S. Mohammad, and S. Kiritchenko, "Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 505–516. [Online]. Available: https://aclanthology.org/N19-1050

[150] M. A. Hadj Taieb, T. Zesch, and M. Ben Aouicha, "A survey of semantic relatedness evaluation datasets and procedures," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4407–4448, 2020. [Online]. Available: https://doi.org/10.1007/s10462-019-09796-3

[151] J. J. Lastra-Díaz, J. Goikoetxea, M. A. H. Taieb, A. García-Serrano, M. B. Aouicha, and E. Agirre, "A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art," *Engineering Applications of Artificial Intelligence*, vol. 85, pp. 645–665, 2019. [Online]. Available: https://doi.org/10.1016/j.engappai.2019.07.010

[152] R. Navigli and F. Martelli, "An overview of word and sense similarity," *Natural Language Engineering*, vol. 25, no. 6, p. 693–714, 2019. [Online]. Available: https://doi.org/10.1017/S1351324919000305

[153] S. Zad, M. Heidari, P. Hajibabaee, and M. Malekzadeh, "A survey of deep learning methods on semantic similarity and sentence modeling," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2021, pp. 0466–0472. [Online]. Available: https://doi.org/10.1109/IEMCON53756.2021.9623078

[154] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 1–8. [Online]. Available: https://aclanthology.org/S14-2001

[155] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14. [Online]. Available: https://aclanthology.org/S17-2001

[156] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 task 10: Multilingual semantic textual similarity," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 81–91. [Online]. Available: https://aclanthology.org/S14-2010

[157] S. Chaturvedi, H. Peng, and D. Roth, "Story comprehension for predicting what happens next," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1603–1614. [Online]. Available: https://aclanthology.org/D17-1168

[158] D. K. Elson, "Detecting story analogies from annotations of time, action and agency," in *Proceedings of The Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27*, 2012, pp. 91–99. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[159] B. Saldias and D. Roy, "Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*. Online: Association for Computational Linguistics, Jul. 2020, pp. 78–86. [Online]. Available: https://aclanthology.org/2020.nuse-1.10

[160] B. Fisseni and B. Löwe, "Which dimensions of narratives are relevant for human judgments of story equivalence?" in *Proceedings of the The Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27, 2012*, 2012, pp. 114–118. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[161] D. Nguyen, D. Trieschnigg, and M. Theune, "Using crowdsourcing to investigate perception of narrative similarity," in *Proceedings of the 23rd Association for Computing Machinery (ACM) Conference on Information and Knowledge Management*, 2014, pp. 321–330. [Online]. Available: https://doi.org/10.1145/2661829.2661918

[162] M. P. Fay, "Story comparison via simultaneous matching and alignment," in *Proceedings of the Third Workshop on Computational Models of Narrative (CMN'12), Istanbul, Turkey, May 26-27, 2012*, 2012, pp. 98–102. [Online]. Available: http://narrative.csail.mit.edu/cmn12/proceedings.pdf

[163] R. Barzilay and L. Lee, "Learning to paraphrase: An unsupervised approach using multiple-sequence alignment," in *Proceedings of the 2003 Human Language Technology Conference of the North American*

Chapter of the Association for Computational Linguistics (NAACL-HLT), 2003, pp. 16–23. [Online]. Available: https://aclanthology.org/N03-1003

[164] ——, "Bootstrapping lexical choice via multiple-sequence alignment," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, Jul. 2002, pp. 164–171. [Online]. Available: https://aclanthology.org/W02-1022

[165] A. Dunlop, M. Mitchell, and B. Roark, "Prenominal modifier ordering via multiple sequence alignment," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 600–608. [Online]. Available: https://aclanthology.org/N10-1085

[166] V. F. Lacatusu, S. J. Maiorano, and S. M. Harabagiu, "Multi-document summarization using multiple-sequence alignment," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/408.pdf

[167] J. Prokić, M. Wieling, and J. Nerbonne, "Multiple sequence alignments in linguistics," in *Proceedings of the European Chapter of the Association for Computational Linguistics EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*. Athens, Greece: Association for Computational Linguistics, Mar. 2009, pp. 18–25. [Online]. Available: https://aclanthology.org/W09-0303

[168] L. Alonso, I. Castellón, J. Escribano, X. Messeguer, and L. Padró, "Multiple sequence alignment for characterizing the lineal structure of revision," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/332.pdf

[169] J. Grimm and W. Grimm, *Household Tales*. London: George Bell, 1884.

[170] T. Strauss, M. J. von Maltitz, and D. E. Blasi, "Word Frequency Distributions of African Languages," in *Book of Abstracts of the International Conference of the Digital Humanities Association of Southern Africa (DHASA) 2019*, vol. 2, 2019, pp. 11–12. [Online]. Available: https://doi.org/10.55492/dhasa.v2i01

[171] F. Pannach and A. Krishnan, "Spielräume zwischen Yakshis und Dibias: Vladimir Propps Morphologie des Märchens im ontologiegestützten interkulturellen Vergleich," in *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2020), Paderborn*. Zenodo, Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4621968

[172] H. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," *Advances in Neural Information Processing Systems*, vol. 22, 2009. [Online]. Available: https://papers.nips.cc/paper_files/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf

[173] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 63–70. [Online]. Available: https://aclanthology.org/W14-3110

[174] A. Zgoll, "Innana and En-hedu-ana: Mutual empowerment and the myth Innana conquers Ur," in *Powerful women in the ancient world: perception and (self) presentation: Proceedings of the 8th Melammu Workshop, Kassel, 30 January–1 February 2019*, no. 4. Münster: Zaphon, 2021.

[175] C. M. Bowra, "Orpheus and Eurydice," *The Classical Quarterly*, vol. 2, no. 3-4, p. 113–126, 1952. [Online]. Available: https://doi.org/10.1017/S0009838800007722

[176] S. N. Kramer, "The death of Dumuzi: A new Sumerian version," *Anatolian Studies*, vol. 30, pp. 5–13, 1980.

[177] M. E. Cohen, "A new piece of an Inanna/Dumuzi lamentation," in *He Has Opened Nisaba's House of Learning*, L. Sassmannshausen and G. Neumann, Eds. Leiden: Brill, 2014, pp. 37–49.

[178] B. Alster, *A Dumuzi Lament in Late Copies*. University of Hiroshima, Faculty of Letters, Department of Linguistics, 1985.

[179] M. E. Cohen, "Sumerian hymnology: the Ersemma," *Sumerian Hymnology*, pp. 1–217, 1981.

[180] M. Fritz, *„... und weinten um Tammuz": Die Götter Dumuzi-Ama'ušumgal'anna und Damu (= AOAT 307)*. Ugarit-Verlag, Münster, 2003.

[181] H. A. Ormerod, R. E. Wycherley, W. H. S. Jones *et al.*, *Pausanias description of Greece*. London: W. Heinemann, 1918, vol. 1.

[182] A. N. Marlow, "Orpheus in Ancient Literature," *Music and Letters*, vol. XXXV, no. 4, pp. 361–369, 10 1954. [Online]. Available: https://doi.org/10.1093/ml/XXXV.4.361

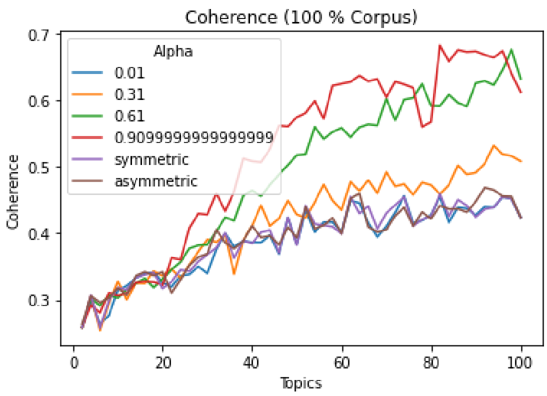
[183] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[184] N. Bansal, M. Akter, and S. K. Karmaker Santu, "Semantic overlap summarization among multiple alternative narratives: An exploratory study," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6195–6207. [Online]. Available: https://aclanthology.org/2022.coling-1.541

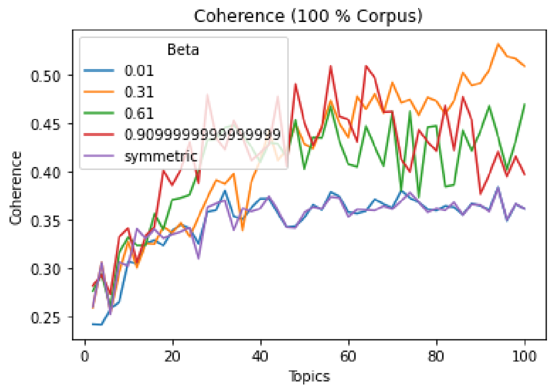
Appendix A

A.1 LDA Hyper-Parameter Tuning

LDA hyper-parameter tuning for the English hyleme data sets, coherence and complexity, in Figures A.1a-A.3d, refers to Section 4.4.1.

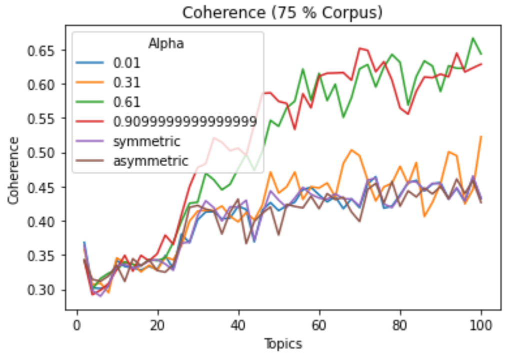


(a) Coherence for $0.01 <= \alpha < 1$, *S/AS*, $\beta = 0.31$, 100 % corpus

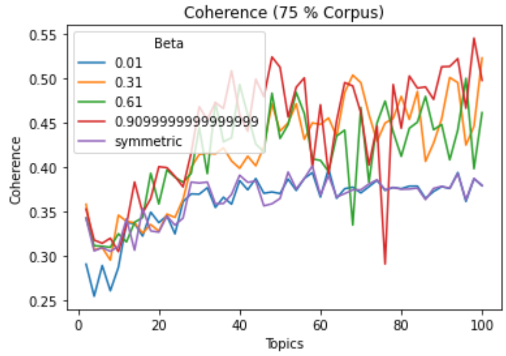


(b) Coherence for $0.01 <= \beta < 1$, *S*, $\alpha = 0.31$, 100 % corpus

Figure A.1: Coherence with different hyper-parameter settings, $2 < k < 100$ (English hyleme data)

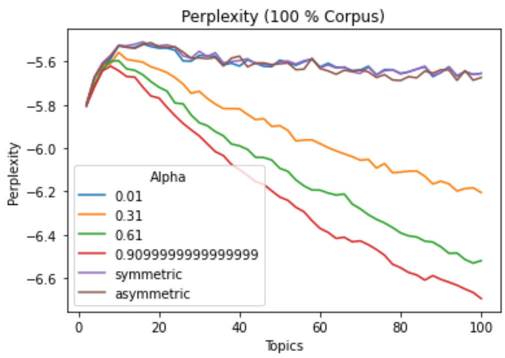


(a) Coherence for $0.01 <= \alpha < 1$, *S/AS*, $\beta = 0.31$, $\alpha = 0.3$, 75 % corpus

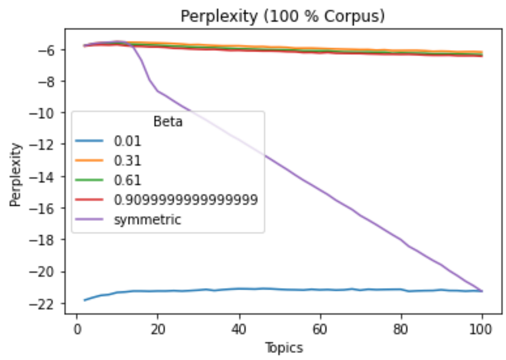


(b) Coherence for $0.01 <= \beta < 1$, *S*, $\alpha = 0.31$, 75 % corpus

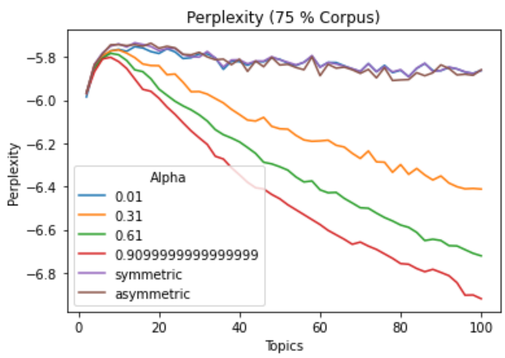
Figure A.2: Coherence with different hyper-parameter settings, $2 < k < 100$ (English hyleme data)



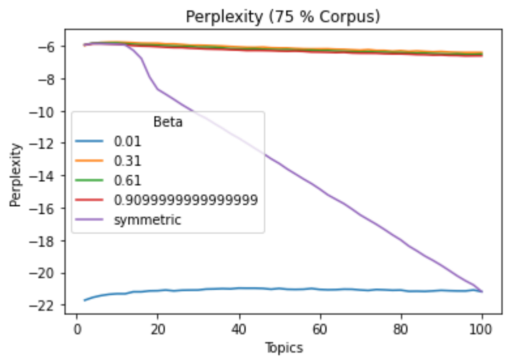
(a) Log-perplexity for $0.01 <= \alpha < 1$, *S/AS*, $\beta = 0.31$, 100 % corpus



(b) Log-perplexity for $0.01 <= \beta < 1$, *S*, $\alpha = 0.31$, 100 % corpus



(c) Log-perplexity for $0.01 <= \alpha < 1$, *S/AS*, $\beta = 0.31$, $\alpha = 0.3$, 75 % corpus



(d) Log-perplexity for $0.01 <= \beta < 1$, *S*, $\alpha = 0.31$, 75 % corpus

Figure A.3: Log-perplexity with different hyper-parameter settings, $2 < k < 100$ (English hyleme data set)

A.2 Modelling Event Structure

DFA corresponding to the right regular grammar introduced in Section 6.1.1 in Figure A.5. Word Cloud visualisations of hyper-hylemes in Figure A.4.

Figure A.4: WordCloud visualisation of the Orpheus' hyper-hylemes

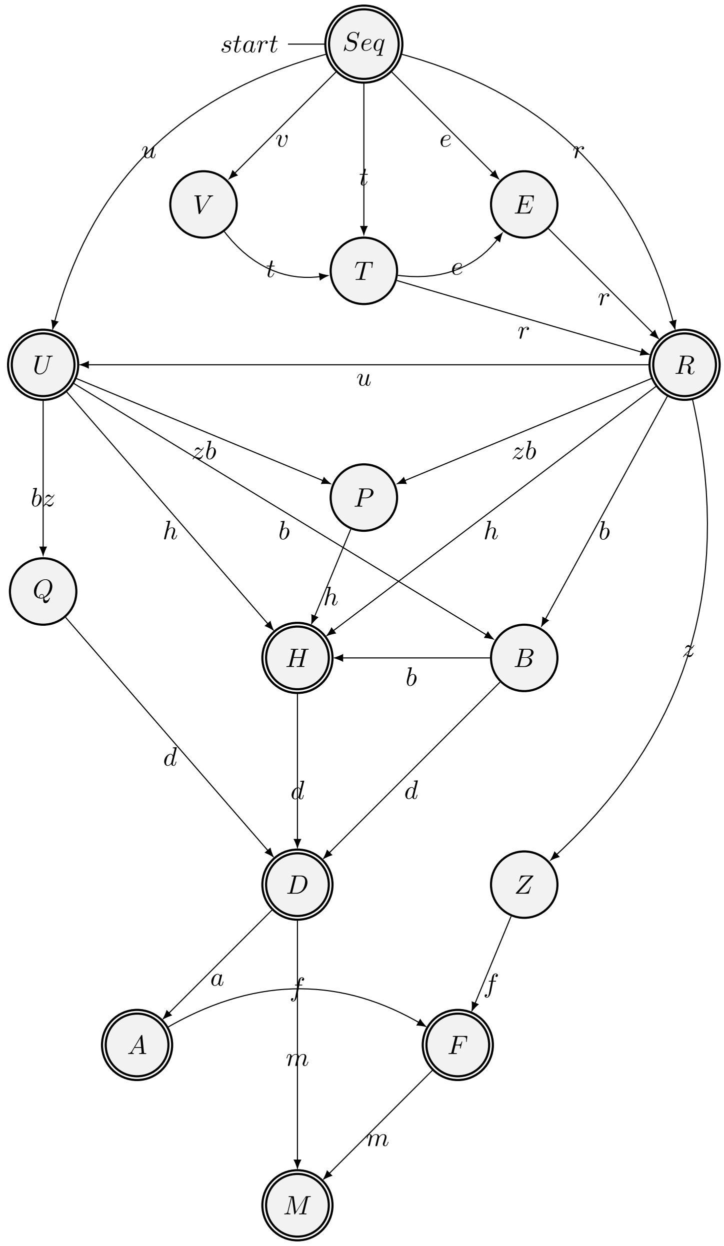


Figure A.5: Deterministic finite automaton of the Orpheus' rules

Table A.1: Hyleme and Corresponding GermaNet Synsets for the hyleme predicate lemma for Sequence 1 (Hermesianax Frag. 7)

Hyleme	GermaNet Synset ID	Word Class	No. of Synsets/Lemma No. of Lexunits (Correct Lemma)	Frame Subcategorization
Orpheus wappnet sich mit seiner Kithara.	s54812	Kognition	1,2 (2)	NN.AR.Pp
Orpheus segelt zu Charon.	s57989	Lokation	2,1	NN
Orpheus reist allein am Ufer entlang.	s57900	Lokation	1,1	NN.Bd
Orpheus spielt auf seiner Kithara.	s53565	Gesellschaft	15,1	NN.AN
Orpheus gewinnt verschiedene Götter für sich.	s56160	Kommunikation	6,2 (1)	NN.AN.PP
Orpheus widersteht dem Blick des furchtbaren Kerberos.	s55485	Kognition	3,1	NN.DN
Orpheus stimmt die großen Herrscher (der Unterwelt) mit Gesang um.	s56155	Kommunikation	2,1	NN.AN
Argiope bekommt zarten Lebensatem.	s52558	Allgemein	6,3 (1)	NN.AN
Orpheus bringt Argiope aus dem Hades herauf.	s57637	Lokation	11,1	NN.AN.BD

Table A.2: Realisation of Hyper-Hylemes in Orpheus' Descend to the netherworld, GermaNet word class corresponding to the hyleme predicate lemma, * Item not found in GermaNet, synonym or semantically similar lemma used (Cont'd)

<table>
<tr><th>Vocabulary</th><th>Word class</th><th>Lemmas</th></tr>
<tr><td>V</td><td>[Lokation, Lokation]</td><td>[verfolgen, fliehen]</td></tr>
<tr><td>T</td><td>[Lokation, Koerperfunktion, Koerperfunktion]</td><td>[gehen, beißen, sterben]</td></tr>
<tr><td></td><td>[Kontakt, Koerperfunktion, Veraenderung]</td><td>[treten, sterben, übersiedeln]</td></tr>
<tr><td></td><td>[Lokation, Lokation]</td><td>[entgehen, verursachen]</td></tr>
<tr><td></td><td>[Koerperfunktion, Koerperfunktion, Lokation]</td><td>[beißen, sterben, hinabsteigen]</td></tr>
<tr><td></td><td>[Koerperfunktion]</td><td>[sterben]</td></tr>
<tr><td>E</td><td>[Kognition]</td><td>[sich wappnen | wollen]</td></tr>
<tr><td></td><td>[Gefuehl]</td><td>[trauern]</td></tr>
<tr><td></td><td>[Veraenderung, Kognition]</td><td>[sehen, wollen]</td></tr>
<tr><td></td><td>[Kognition, Kommunikation]</td><td>[sehen, besingen]</td></tr>
<tr><td></td><td>[Kommunikation, Gefuehl, Kommunikation]</td><td>[Klagegesang führen*, trösten, bezichtigen]</td></tr>
<tr><td></td><td>[Kognition, Kognition]</td><td>[sehnen, ersinnen]</td></tr>
<tr><td>R</td><td>[Lokation, Lokation, Gesellschaft]</td><td>[segeln, reisen, spielen]</td></tr>
<tr><td></td><td>[Lokation]</td><td>[hinabsteigen | betreten | hinuntergehen | reisen | gehen]</td></tr>
<tr><td></td><td>[Lokation, Lokation]</td><td>[steigen, erreichen]</td></tr>
<tr><td></td><td></td><td>[folgen, herabsteigen]</td></tr>
<tr><td></td><td></td><td>[kommen, hinabsteigen]</td></tr>
<tr><td></td><td>[Lokation, Gesellschaft, Lokation]</td><td>[betreten, musizieren, gelangen]</td></tr>
<tr><td></td><td>[Gesellschaft]</td><td>[aufsuchen]</td></tr>
<tr><td>U</td><td>[Kommunikation, Kognition, Kommunikation, Allgemein]</td><td>[gewinnen, widerstehen, umstimmen, bekommen]</td></tr>
<tr><td></td><td>[Gefuehl, Kommunikation, Kommunikation]</td><td>[besänftigen, fordern, umstimmen]</td></tr>
<tr><td></td><td>[Kommunikation, Gefuehl, Kommunikation, Kommunikation]</td><td>[singen, ziehen, bannen, verweigern]</td></tr>
<tr><td></td><td>[Gesellschaft, Kommunikation, Veraenderung, Besitz]</td><td>[anschlagen, gewinnen, ersetzen, erhalten]</td></tr>
<tr><td></td><td>[Kommunikation, Gefuehl]</td><td>[erbitten, erweichen]</td></tr>
<tr><td></td><td>[Kommunikation, Kommunikation, Gesellschaft]</td><td>[singen, überzeugen, gestatten]</td></tr>
<tr><td></td><td>[Kommunikation, Gesellschaft]</td><td>[überzeugen, versprechen]</td></tr>
<tr><td></td><td>[Gefuehl, Besitz]</td><td>[betören, erhalten]</td></tr>
<tr><td></td><td>[Kognition]</td><td>[weichen]</td></tr>
<tr><td></td><td>[Kommunikation, Gesellschaft]</td><td>[überzeugen, erlauben]</td></tr>
<tr><td></td><td>[Gesellschaft, Gefuehl, Kommunikation, Gefuehl*, Besitz]</td><td>[spielen, berühren, bitten, Mitleid haben*, schenken]</td></tr>
<tr><td>Z</td><td>[Kommunikation, Lokation]</td><td>[herbeirufen, kommen]</td></tr>
<tr><td></td><td>[Allgemein]</td><td>[erhalten]</td></tr>
<tr><td></td><td>[Besitz]</td><td>[wiedergeben]</td></tr>
</table>

Table A.3: Realisation of Hyper-Hylemes in Orpheus' Descend to the netherworld, GermaNet word class corresponding to the hyleme predicate lemma, * Item not found in GermaNet, synonym or semantically similar lemma used (Cont'd)

<table>
<tr><th>Vocabulary</th><th>word class</th><th>lemmas</th></tr>
<tr><td>B</td><td>[Gesellschaft]</td><td>[auferlegen]</td></tr>
<tr><td></td><td>[Besitz, Allgemein]</td><td>[nehmen, dürfen]</td></tr>
<tr><td></td><td>[Kognition]</td><td>[annehmen]</td></tr>
<tr><td></td><td>[Kognition, Kognition]</td><td>[Vorschriften machen*, einhalten]</td></tr>
<tr><td></td><td>[Allgemein]</td><td>[dürfen]</td></tr>
<tr><td></td><td>[Besitz]</td><td>[bekommen]</td></tr>
<tr><td></td><td>[Kommunikation, Kommunikation]</td><td>[gebieten, verbieten]</td></tr>
<tr><td></td><td>[Veraenderung]</td><td>[einschränken]</td></tr>
<tr><td>H</td><td>[Lokation]</td><td>[heraufbringen | gehen | zurückführen | folgen | emporführen]</td></tr>
<tr><td></td><td>[Lokation, Lokation]</td><td>[antreten, nahe sein*]</td></tr>
<tr><td></td><td>[Lokation, Lokation, Lokation]</td><td>[aufmachen, gehen, fast am Licht sein*]</td></tr>
<tr><td>D</td><td>[Kognition, Perzeption]</td><td>[ertragen, blicken]</td></tr>
<tr><td></td><td>[Gefuehl, Gefuehl, Kognition, Perzeption]</td><td>[fürchten, fürchten, sehnen, blicken]</td></tr>
<tr><td></td><td>[Lokation, Perzeption]</td><td>[drehen, blicken]</td></tr>
<tr><td></td><td>[Kognition, Kognition, Perzeption]</td><td>[glauben, denken, blicken]</td></tr>
<tr><td></td><td>[Gefuehl, Perzeption]</td><td>[fürchten, blicken]</td></tr>
<tr><td></td><td>[Kognition, Gesellschaft, Perzeption]</td><td>[wollen, befolgen, blicken]</td></tr>
<tr><td></td><td>[Gesellschaft, Lokation, Perzeption]</td><td>[gehorchen, umdrehen, schauen]</td></tr>
<tr><td></td><td>[Kognition]</td><td>[vergessen]</td></tr>
<tr><td></td><td>[Kognition, Lokation]</td><td>[glauben, umdrehen]</td></tr>
<tr><td></td><td>[Kognition, Allgemein, Perzeption]</td><td>[missachten, Lokation]</td></tr>
<tr><td></td><td>[Perzeption]</td><td>[anblicken]</td></tr>
<tr><td>A</td><td>[Veraenderung]</td><td>[verlieren | umkehren]</td></tr>
<tr><td></td><td>[Lokation, Kommunikation, Kommunikation]</td><td>[gleiten, erheben, sagen]</td></tr>
<tr><td></td><td>[Veraenderung, Lokation]</td><td>[zunichte machen, zurückkehren]</td></tr>
<tr><td></td><td>[Perzeption, Lokation]</td><td>[ertönen, verschwinden]</td></tr>
<tr><td></td><td>[Lokation]</td><td>[umkehren]</td></tr>
<tr><td>F</td><td>[Gefuehl, Kommunikation, Kommunikation]</td><td>[starr sein*, erbitten, verwehren]</td></tr>
<tr><td></td><td>[Veraenderung]</td><td>[begehen]</td></tr>
<tr><td></td><td>[Kommunikation, Gefuehl]</td><td>[verbieten, weinen]</td></tr>
<tr><td></td><td>[Koerperfunktion]</td><td>[zugrunde gehen*]</td></tr>
<tr><td></td><td>[Besitz]</td><td>[zurückschicken]</td></tr>
<tr><td>M</td><td>[Gesellschaft, Veraenderung]</td><td>[lassen, reißen]</td></tr>
<tr><td></td><td>[Veraenderung]</td><td>[töten]</td></tr>
<tr><td></td><td>[Schoepfung]</td><td>[bewirken]</td></tr>
</table>

A.3 Semantic Similarity

A.3.1 Annotation Study Introductory Text (German)

Herzlich Willkommen zur Umfrage „Semantische Ähnlichkeit von Satzpaaren“!

Bitte lesen Sie diese Anmerkungen sorgfältig, bevor Sie mit der Bearbeitung beginnen.

In dieser Studie werden Ihnen je vier Satzpaare präsentiert, die sie nach ihrer semantischen Nähe (in der Computerlinguistik auch „Semantic Relatedness/Semantic Similarity“ genannt) bewerten sollen. Mit semantischer Nähe meinen wir sowohl die Ähnlichkeit zwischen zwei Begriffen oder Aussagen, z.B. „Auto“ und „Bus“, als auch weiter reichende Verwandschaft von bedeutungsfeld-relevante Konzepten, wie „Auto“ (Fahrzeug) und „Straße“ (Bauwerk/Infrastruktur) Ein Beispiel für ein Satzpaar mit großer semantischer Nähe ist z.B.„Der Mann betritt das Haus.“ - „Der Mann tritt in das Haus ein.“ Dabei müssen die Aussagen nicht zwangsläufig in Subjekt, Prädikat und Objekt übereinstimmen, um semantisch nah zu sein, z.B. „Peter kauft das Fahrrad von Max.“ - „Max verkauft das Fahrrad an Peter.“

Die beiden Sätze unterscheiden sich oft nur in einzelnen Bestandteilen, manchmal nur in einzelnen Wörtern. Bei Wörtern, die verschiedene Lesarten haben, gehen Sie bitte von der Lesart aus, die am engsten mit der Verwendung im Originalsatz verwandt ist, z.B. „Er betrat das Boot.“ – „Er betrat das Schiff.“ (im Sinne von Wasserfahrzeug, nicht Kirchenschiff.)

Die Originalsätze sind aus Erzählkontexten entnommen, die Ihnen zum Teil bekannt sind. Bitte ignorieren Sie etwaige Widersprüche zu den Originalkontexten. (z.B. „Odysseus besiegt den Zyklopen nicht.“) Konzentrieren Sie sich stattdessen nur auf die semantische Nähe der verschiedenen Satzpaare. Bitte gehen Sie bei der Bewertung der Semantischen Nähe nach Ihrer Intuition vor. Die Bearbeitung der Umfrage wird ca. 20min in Anspruch nehmen.

Vielen Dank für Ihre Unterstützung! Kontakt: [Contact information]

A.3.2 Annotation Study Introductory Text (English)

Welcome to the survey "Semantic similarity of sentence pairs"!

Please read these instructions carefully before you start.

In this survey, you will be presented with four pairs of sentences each, which you are asked to rate according to their semantic proximity (also called "semantic relatedness/semantic similarity" in computational linguistics). By semantic proximity we mean both the similarity between two terms or statements, e.g. "car" and "bus", as well as more far-reaching relatedness of concepts relevant to the field of meaning, such as "car" (vehicle) and "road" (building/infrastructure) An example of a sentence pair with high semantic proximity is e.g. "The man enters the house". - "The man enters the house." The statements do not necessarily have to agree in subject, predicate and object to be semantically close, e.g. "Peter buys the bicycle from Max." - "Max sells the bicycle to Peter."

The two sentences often differ only in individual components, sometimes only in individual words. For words that have different readings, please go by the reading that is most closely related to the use in the original sentence, e.g. "He entered the boat." - "He entered the ship." (in the sense of watercraft, not nave[1].)

The original sentences are taken from narrative contexts, some of which you are familiar with. Please ignore any contradictions with the original contexts. (e.g. "Odysseus does not defeat the Cyclops.") Instead, focus only on the semantic proximity of the different pairs of sentences. Please use your intuition when assessing the semantic proximity. The survey will take about 20min to complete.

Thank you very much for your support! Contact: [Contact information]

A.3.3 Transformation Effects

Transformation effects t for the hyleme pairs in the 4-tuples for the hyleme proximity annotation, see Section 7.1.

[1]The German term *Schiff* can have multiple meanings, including *ship* and *transept*

Table A.4: Hyleme transformations for gold standard annotation, transformation in ()

Effect	Subtype	Example	Translation
Addition/Deletion			
	one of multiple subjects	'Engel (und Menschenseelen) preisen Gottes Werk und Schöpfung.'	'Angels (and mortal souls) praise God's work and creation'
	one of multiple objects	'Er sieht die Erde (und den Himmel).'	'He sees the earth and the heaven.'
	subject determination	'Der (erste) Mensch nennt seine Frau Eva.'	'The (first) human names his wife Eva.'
	object determination	'Der Mann sieht (stark und) jugendlich aus.'	'The man looks (strong and) youthful. '
	predicate determination	'Charon treibt den Kahn (mit der Ruderstange) aufwärts.'	'Charon propels the boat up (with the oar).'
Taxonomic relationship (Hyper-/Hyponym relationship)			
	subject	'Der(Pharao/König) sieht die Bewohner der Unterwelt.'	The (pharaoh/king) sees the ' inhabitants of the netherworld.
	object	'Sie bringt (eine Person/einen Sterblichen) hinein.'	'She brings a (person/mortal) in.'
	predicate	'Dumuzi (liebt/mag) Geštinanna.'	'Dumuzi (loves/likes) Geštinanna.'
	object attribute/determination	'Er nennt Riesen (bitter)böse Geister.'	'He calls giants (evil/malicious) spirits.'
	predicate determination	'Der Mann sieht (jung/jugendlich) aus.'	'The man looks (young/youthful).'
Additional specifications for durative-constant hylemes			
	additional subject complement	'Sie ist eine Amme und eine Mutter.'	'She is a wet nurse (and a mother).'
	predicate adjective hyper-/hyponym relationship	'Odysseus und seine Gefährten sind (traurig/todunglücklich).'	'Odysseus and his companions are (sad/heartbroken).'
	change of predicate adjective (no semantic relationship)	'Das Feuer ist (gleichmäßig/heiß).'	'The fire is (smooth/hot).'

Table A.5: Hyleme transformations for gold standard annotation, transformation in () cont'd.

Effect	Subtype	Example	Translation
Change of hyleme component (without semantic relationship)			
	subject	'Die (Frauen/Kinder) der Söhne gehen in die Arche.'	'The (wifes/children) of the sons walk into the ark.'
	object	'Er bringt (eine Person /ein Lämmchen) hinein.'	'He brings in a person/ a little lamb.'
	predicate	'Enki (beschließt/unternimmt) eine Reise zur Unterwelt.'	'Enki (decides to go on/ undertakes) a journey to the netherworld.'
	object det.	'Die Annunaki sind (große/alte) Götter.'	The Annunaki are (great/old) deities.
Grammatical number			
	subject	'(Die/Das) Feuer sind gleichmäßig.'	'The fire(s) are even.'
	subject det.	'Baumeister (der Gräber/des Grabes) darf das Grab nicht finden.'	'The master builder of the (grave/graves) is not allowed to find the grave.
	object	'Saturus und Perpetua sehen (große Lichter/ein großes Licht).'	'Saturus and Perpetua see (big lights/a big light).'
Specification of quantity			
	Subject	'Eine (Anzahl/Vielzahl) von Toten geht hin und her um die Grube."	'A (number/large number) of dead persons walk back and forth around the pit.
	Object	'Die Größe der Riesen ist (dreitausend/dreihundert) Ellen.'	'The size of the giants is three(thousand/hundred) cubits.
	object det.	'Noah nimmt von dem reinem Vieh je sieben/siebzehn mit.'	'Noah takes (seven/seventeen) of each of the pure animals with him.'
Hyleme rephrasing		'Sie ist keine Göttin.'	'She is not a goddess.'
		'Sie ist nicht göttlich.'	'She is not divine.'
Negation		'Elischas Mantel fällt (nicht) herunter.'	'Elisha's cloak (does not) fall down.
Antonym		'Ba'l soll sich zum Berg Knkny begeben.'	'Ba'l is to go to Mount Knkny.'
		'Ba'l soll sich vom Berg Knkny fernhalten.'	'Ba'l is to stay away from Mount Knkny.'