Philosophische Fakultät III
Sprach- , Literatur- und Kulturwissenschaften
Institut für Information und Medien, Sprache und Kultur (I:IMSK)
Professur Digital Humanities

AI vs. Human Generated Fairy Tales
A Stylometric and Folkloristic Comparison across GPT-4o, GPT-5, and German Folk Tales

Name: Sophia-Magdalena Heigl
Matr.-Nr.: 1781725
Semesteranzahl/Studiengang: 8
E-Mail: Sophia-Magdalena.Heigl@stud.uni-regensburg.de
Private E-Mail: sophiaheigl@gmx.de
Erstgutachter*in: Jun.-Prof. Mareike Schumacher
Zweitgutachter*in: Prof. Dr. Udo Kruschwitz

Abgegeben am 30.09.2025

Abstract

Defining what makes a fairy tale a real fairy tale has been an enduring endeavor in the field of folklore studies. With an arsenal of computational methods at hand, the question still stands, if there is an empirical way to capture the essence of the folk fairy tale and if linguistic-stylistic measures can – can AI capture its essence as well? This study investigates to what extent large language models (GPT-4o, and the 2025 released GPT-5) emulate the stylistic and orally transmitted characteristics of the fairy-tale genre. The study applies stylometry, supported by POS profiles, sentiment analysis, n-grams, and readability measures. Results indicate overreliance on formulaic markers (e.g., "Once upon a time"), reduced dialogues, and more uniform syntax; GPT-5 is closer to folk tales than GPT-4o.

Zusammenfassung

Festzuhalten, was das Märchen märchenhaft macht, ist ein andauerndes Unterfangen der Volkskunde. Trotz einer Vielzahl von computergestützten Methoden bleibt die Frage offen, ob es einen empirischen Weg gibt, das Wesen des Volksmärchens zu erfassen, und ob sprachlich-stilistische Maßnahmen dazu geeignet sind und ob KI sein Wesen ebenfalls erfassen kann? Diese Studie untersucht, inwieweit große Sprachmodelle (GPT-4o und das 2025 veröffentlichte GPT-5) die stilistischen und mündlich überlieferten Merkmale des Märchen-Genres nachahmen. Die Studie wendet Stilometrie an, unterstützt durch POS-Profile, Sentiment Analyse, N-Gramme und Lesbarkeitsmaße. Die Ergebnisse zeigen eine übermäßige Abhängigkeit von formelhaften Markern (z. B. „Es war einmal"), eine reduzierte Dialogizität und eine einheitlichere Syntax; GPT-5 ist den Volksmärchen näher als GPT-4o.

Inhalt

1 Introduction

> "There was once upon a time, in a lonely valley between high mountains and dark woods, a poor woodcutter who lived with his only daughter. She was as bright and good as the morning sun, and though their hut was but a handful of planks and moss, it shone with her cheerfulness as if a king dwelt there."

While the passage above sounds and feels like the familiar opening of a folktale, it is, in fact, not. It is a story generated by ChatGPT-5 designed to imitate a classical fairy tale. With the release of ChatGPT-3.5 in late 2022, large-scale Natural Language generation (NLG) has been made much more available for the general public (Herbold et al., 2023: Related Work; Brown et al., 2020). This has brought about profound shifts in how content and cultural artifacts are created and experienced. NLG has already found application across diverse domains, including medical research and diagnostics (Javaid et al., 2023), language teaching and learning (Kohnke et al., 2023) and academic writing (Amirjalili et al., 2024).

At the same time, this technology has increasingly been applied to creative writing and fiction (Shidiq, 2023). Among the many genres explored, the fairy tale has emerged as a particularly compelling field of experimentation. Applications such as *FairyLandAI: Personalized Fairy Tales* utilizing ChatGPT and DALL·E-3 (Makridis et al., 2024) and *FairyTailor* (Bensaid et al., 2021) demonstrate how Large Language Models (LLMs) can be used to generate personalized fairy tales, often enriched with artificial intelligence (AI) generated illustrations and tailored to individual children's names, interests, or situations.

These developments highlight both the creative and educational potential, as well as innovative forms of storytelling and engagement, while on the other hand, they raise questions about authenticity, tradition, pedagogical functions and genre fidelity. This study ties in with research about genre identification and genre adherence in connection with texts generated by artificial intelligence. Such studies on genre classification highlight the importance of linguistic-stylistic and establish stylometric features, such as word frequency distributions, function-

word usage, and sentence complexity, as a stable indicator of genre (Melliti, 2024; Vanetik et al., 2024). Computational stylistic markers have further been found to aid in the differentiation and identification of AI-generated texts (Kumarage et al., 2023).

The proposed approach contributes to this discourse by developing several genre-defining characteristics of human and AI-generated folk fairy tales, derived from theories of folklore studies and well-established writing style indicators. Chosen for this examination are a corpus of classical German folk fairy tales and two corpora tales generated by OpenAI GPT models, in this case ChatGPT-4o and ChatGPT-5-main. To assess the ability of LLMs to emulate this particular genre, this analysis draws on the methodology of quantitative text analysis used in the Digital Humanities and Computational Linguistics and examines a range of features in the texts generated by the two models, including properties on the word, syntax and document level. Building on Max Lüthi's concepts of form and function in fairy tales (1986), the study applies stylometric analysis in combination with other computational methods developed within the Digital Humanities, such as topic modelling and sentiment analysis to further highlight the properties of the generated texts. The findings support the notion that, while LLMs reliably reproduce surface genre cues like, canonical openings, restricted colour palette and contrastive adjectives, they underperform on oral-pragmatic features (dialogicity, variable pacing, moral evaluation). GPT-5 converges more closely on the human reference than GPT-4o but remains detectably distinct.

Although LLMs have been widely studied in terms of their capabilities on the imitation of human language, in academic writing as well as in creative text production (Amirjalili et al., 2024; Herbold et al., 2023; Bahndarkar et al., 2024), their ability to imitate folktales has not been extensively studied, to the researcher's best knowledge. Classical German fairy tales represent a highly formalized genre, characterized by recurring motifs, stylized expressions, and recognizable structures, as proven by research from the field of folklore studies (Propp, 1968; Uther, 2004), which makes them an excellent test case for examining how successful GPT models can reproduce narrative structures and stylistic markers. By combining

folklore theories with computational methods such as opinion mining, lexical diversity methods and frequency based stylometry this study contributes to an interdisciplinary dialogue between folklore research, literary studies, and the Digital Humanities.

Central to this development is the role of LLMs, and particularly OpenAI's ChatGPT, a chatbot that relies on generative pre-trained transformer (GPT) architecture (Berryman/ Ziegler, 2023: 9). The capabilities of ChatGPT-3.5 and GPT-4o have been under extensive investigation (Mikros, 2025; Cai et al., 2023; Georgiou, 2024; Herbold et al., 2025). In August 2025, OpenAI released its new ChatGPT-5-main model to the public and advertised higher performance on benchmarks for its latest model, a reduced answering time, reduction in hallucinations and a higher performance in health, coding and writing (OpenAI, 2025). While LLMs are getting better at emulating human text, their results still show discernible differences compared to human-produced writing (Beguš, 2024; Sandler et al., 2024, Fedoriv, 2023). Since the more recent version of ChatGPT was released in August 2025, this study will serve as one of the first assessments of performance of the new ChatGPT-5 model against its predecessor, ChatGPT-4.

1.1 Folklore Key Terminology

The term fairy tale has been defined in various ways within folklore scholarship. Max Lüthi describes it as "a prose narrative of oral origin, marked by motifs of the supernatural and wonder, a lack of concrete historical or geographical setting, and a stylized, simple mode of expression," (Lüthi, 1986: 1) and as "a world-encompassing adventure story told in a swift, sublimating style" (Lüthi, 1986: 82). In the context of this study, the term *fairy tale* will be used in this narrower folkloristic sense. In this effort, it is necessary to distinguish between fairy tales and the broader category of the folktale. In folklore studies, the folktale functions as an umbrella term for orally transmitted prose narratives of anonymous origin (Carrassí, 2016: 99). This category includes a wide variety of subtypes, such as animal tales, trickster anecdotes, tall tales, jokes, and legends. Within this framework, the fairy tale represents a particular subtype, often referred to in German scholarship as *Zaubermärchen*, which is defined by the presence of magical

elements, supernatural beings, miraculous transformations, and motifs of wonder (cf. Uther, 2004).

A further distinction in terminology must be made for the purposes of this paper. The term classical fairy tale will be used here specifically to denote *Volksmärchen*, which by definition, are orally transmitted folk fairy tales characterised by their anonymous origin and their transmission across generations within a collective tradition (Geldern-Egmond, 2000: 7). In contrast, *Kunstmärchen* (literary fairy tales) are consciously composed by an identifiable author, recorded in writing from the outset, and often bear a distinctive stylistic and thematic signature that reflects the author's individual voice (Geldern-Egmond, 2000: 7). According to this definition, an AI-generated story cannot be considered a folk fairy tale. Nevertheless, this study investigates whether the characteristic features of the genre can be sufficiently imitated by the models.

1.2 Research Questions

This master's thesis examines if AI generated fairy tales can be statistically recognised as classical folk fairy tales and investigates the stylistic differences and similarities of human-written and AI-generated fairy tales. The main motivation is ascertained if there are measures that can be extracted by computational analysis that capture the essence of the fairy tale genre and secondly, if tales created by LLMs not only imitate surface-level characteristics but also reproduce deeper stylistic and folkloristic markers associated with abstraction, oral tradition, and narrative simplicity. Based on this main objective, the following central research question (RQ) is formulated.

RQ: To what extent do fairy tales generated by ChatGPT-4o and ChatGPT-5 reproduce the stylistic markers and oral-tradition features characteristic of the classical German folk fairy tale, and how do they compare to human-written tales on a deeper stylistic level?

This central RQ integrates three distinctive lines of inquiry. The first two derive from theories of folk tales on the narrative form of this genre and the third captures statistical differences between the human-written and the AI generated

tales. The following three subordinate research questions (SRQs) address these dimensions in more detail:

SRQ1: Which genre-specific stylistic markers of abstraction, as identified in folkloristic theory by Max Lüthi (1986), are preserved or altered in AI-generated fairy tales compared to German folk fairy tales?

SRQ2: To what extent do ChatGPT-4o and ChatGPT-4 generated fairy tales reflect features of oral tradition characteristic of the genre?

SRQ3: To what extent do ChatGPT-4o and ChatGPT-4 generated fairy tales statistically diverge from or resemble classical German fairy tales when compared through word frequency analysis?

By structuring the inquiry in this way, the study first establishes whether significant metrics can be derived from folkloristic theory, for one, on the abstract style of fairy tales according to Max Lüthi (1986) and secondly, on features described as being characteristic for orally transmitted narratives, before moving on to examine the degree of stylometric comparability between the corpora.

1.3 Goals

The primary goal of this thesis is to determine whether central concepts folklore theory can be operationalised into measurable metrics that allow for the identification and differentiation of human-written and AI-generated texts (see SRQ1 and SRQ2). Building on this foundation, the study applies stylometric analysis as a complementary approach (see SRQ3) to capture implicit stylistic features that unconsciously shape the genre of classical folk tales. The goal is to develop an analytical pipeline capable of systematically extracting metrics from the texts using quantitative text analysis methods used in the Digital Humanities and computational linguistics, first on the basis of folklore theory and secondly based on stylometry.

Based on preliminary observations and existing research, the following hypotheses are proposed:

1. Since large language models excel at recognizing and extending patterns (Brown et al., 2020: 34; Berryman/ Ziegler, 2024: 4), it is expected that the generated tales will reproduce formulaic features of fairy-tale style, such

opening phrases like "There once was…" ("*Es war einmal*…"). This will be measured by extracting n-grams from the text.

2. It is reasonable to assume that ChatGPT-5-main will perform better on the emulator task than ChatGPT-4o according to OpenAI's system information on the more recent model (2025). This will be determined by looking at the *Cosine Delta (*Smith/Aldridge, 2011) between the corpora, a well-established stylometric measure for capturing the statistical differences in the use of the most frequently used words.
3. Following studies on the performance of LLMs to imitate human-like language (Sandler et al., 2024) the AI-generated fairy tales are expected to show less narrative and stylistic variability than the corpus of classical tales. This will be measured by looking at the number of unique words of the texts and examining lexical diversity.
4. The simplicity of sentence structure, which Rölleke (1975) classifies as being characteristic of orally transmitted narratives, not only reflects on the syntax level but also on the word and document level of the tales. This can be verified by a low lexical diversity, shorter words and high readability.

To test these hypotheses and answer the RQ, the study employs a mixed-methods computational approach. First, features will be investigated with methods such as the extraction of n-grams and corpus-based pattern mining, drawing on Max Lüthi's (1986) concepts on the form and function of fairy tales to evaluate whether AI-generated tales preserve narrative and stylistic conventions. To complement these surface-level measures, stylometric analysis provides a further quantitative insight into linguistic features such as word frequency, sentence length, vocabulary richness, and readability. Together, these methods allow for a systematic assessment of the degree to which AI-generated fairy tales conform to the genre tradition.

1.4 Thesis Outline

What follows is an outline of the approaches and structure of this work. Chapters two and three provides the theoretical foundation by giving a brief overview of the German fairy tale tradition and discussing central concepts of fairy-tale research. Subsequently, the current state of OpenAI's GPT models is highlighted and limitations of ChatGPT's abilities of producing human-like language is discussed. Additionally, recent studies in the use of stylometry in conjunction with AI detection are presented. Chapter four focuses on the operationalization of the study, presenting the applied methodology and feature selection and introduces the tools used for analysis. Chapter five describes the construction of the corpus, including the prompting process for generating AI-based texts, and provides a preliminary description of the data. It also presents the methodology process in detail which is applied in this study as well as its limitations. Chapter six documents the analysis and presents the results of the selected features. Chapter seven further interprets these findings and puts them in relation to existing scholarship. Finally, Chapter eight closes the thesis with a reflection on the findings and a discussion of possible directions for future research.

2 German Fairy Tale Tradition

It is only since the early nineteenth century that the designation fairy tale has established itself as an umbrella term for fantastical and diverse oral narrative materials that include elements of animal tales and tall tales, fables, comic anecdotes (*Schwänke*), and legends. Literary scholarship draws sharp boundaries between the fairy tale and adjacent narrative types. In contrast to the fairy tale's artful and typically optimistic cast, the **saga** is characterized as simple and realistic, and often pessimistic and, unlike the fairy tale, it is meant to be believed (Baumgärtner, 1968: 8; see Geldern-Egmond, 2000: 5). The legend pursues religious aims and is anchored in a fixed religious system (Zitzelsperger, 1984: 19, see Geldern-Egmond, 2000: 5). A *Schwank* is a brief, sometimes grotesque, anecdotal narrative that lacks the poetry and lightness associated with the fairy tale. Fables incorporate moral instruction and often veil social critique through animal personae. Although there are **animal fairy tales**, they differ from the prescriptive logic of the fable (Geldern-Egmond, 2000: 5).

Under the term *fairy tale* fall both folk tales (*Volksmärchen*) and literary fairy tales (*Kunstmärchen*). Folk tales are based primarily on popular, anonymous narratives, whereas art tales can be attributed to a specific author. Important figures that collected and reworked folk tales include the brothers Jakob (1785-1863) and Wilhelm (1786-1859) Grimm, Ludwig Bechstein (1801-1860), and Johann Karl August Musäus (1735-1787). The transmitted materials were often rewritten, with new content added and other elements removed, as in the Grimms' collection.

Originally, the motivation behind the Grimm Brothers' collecting of fairy tales was to serve as philological documentation: evidence for the existence of an ancient, orally transmitted folk poetry. Soon, however, a literary ambition also emerged (Bleek, Die Brüder Grimm und die deutsche Politik, 2023). Wilhelm, in particular, began to shape the texts linguistically, making them suitable not only for scholarship, but also for a broader readership. The first publication of the *Children's and Household Tales* took place in December 1812. The second volume of the first edition appeared in 1815. The Grimm Brothers' motives for publishing the

Children's and Household Tales were rooted in Romantic theories of folk poetry, national-cultural programs, and scholarly interest (Rölleke, 2024: 82–86).

The second edition of the *Children's and Household Tales* appeared in 1819, with Wilhelm Grimm extensively revising the texts to suit a bourgeois readership (Rölleke2024, :87–88). In 1822, an annotation volume underscored the scholarly dimension of the project (Rölleke 2024, :92). Subsequent editions, especially from 1837 onward, introduced further changes such as linguistic simplification, moralization, and religious adjustments, highlighting the tales' didactic purpose and serving as reading material for education (Rölleke 2024, :94). The seventh and final edition of 1857 comprised 200 fairy tales and 10 children's legends (Grimm, 1857). In Folklore studies it is an established convention to use the *Kinder- und Hausmärchen* (KHM) number, when referring to the Brothers Grimm fairy tales (Lüthi, 1986: 13; Uther, 2024: 144). This index draws on the number of each fairy tale in the table of contents. "Cinderella" for example has the designation KHM 21.

Another major collector of fairy tales was Ludwig Bechstein, whose tales achieved the widest circulation, after the Grimm Brothers tales. His first literary publication was the collection *Thüringische Volksmärchen* (Thuringian Folk Tales) in 1823. He then published *Die Volkssagen, Märchen und Legenden des Kaiserstaates Österreich* (Folk Legends, Tales and Myths of the Austrian Empire) in 1840, which, however, did not contain any fairy tales. In 1845, he published his most famous work, the *Deutsches Märchenbuch* (German Fairy-Tale Book), which included 90 tales, 38 of these were based on Early New High German written sources, while 51 came from oral traditions. He also adopted and retold several tales from the Grimm Brothers' *Children's and Household Tales*. The second edition, containing 80 tales and several substantive changes, appeared in 1853. In 1856, he published the *Neues Deutsches Märchenbuch* (New German Fairy-Tale Book), which contained 50 tales. Bechstein distanced himself from the original Grimm tales by exercising greater narrative freedom. Guided by national concerns, he drew on orally transmitted and nearly lost folk poetry and traditions, which he sought to make accessible to the German people (Jurčáková, 2006: 99). Unlike the Brothers Grimm,

Bechstein's aim was less scholarly and more directed toward creating popular and entertaining literature for a wider readership, as he himself states in the preface of his 1853 *Märchenbuch* (Bechstein, 1853: 10).

In contrast to classical folk tales, modern fairy tales (*Kunstmärchen*) are conceived and written down by a specific author; the author is therefore known. Such tales often bear an individual signature, stylistically as well as thematically. *Kunstmärchen* are frequently more complex in structure than folk tales and may address deeper psychological themes. Among the best-known authors are Hans Christian Andersen (1805–1875), Wilhelm Hauff (1802–1827), and E. T. A. Hoffmann (1776–1822). Well-known examples include "The Ugly Duckling," "Little Mermaid," and "The Princess and the Pea". Literary fairy tales may closely resemble folk tales in their use of marvels and magic, but there are variants that unmistakably bear the personal signature of their author (Geister, 2024: 8).

3 Related Works

The state of research relevant to this thesis is divided into three main areas in the following chapter. First, folklore studies, research in the field of computational folkloristics, and the fundamental definitions of the features that constitute the fairy tale will be examined. Second, attention will be given to the current state of OpenAI's GPT models and research in the field of artificial text generation. In particular on studies focusing on investigating GPT-3.5, GPT-4o and other various LLMs on their performance and abilities in NLG. Lastly, since stylometric analyses and various other quantitative text analysis methods used in the Digital Humanities and computational linguistics will cover a significant part of my investigation, these methods and specific applications are looked at in more detail.

3.1 Folklore Studies

Folklore studies focus on the collection, preservation, and analysis of oral traditions and their cultural significance, including the classification by motif and describing form and function of folktales (Lüthi, 1986; Uther, 2004).

More recently, computational folklore studies arose as an interdisciplinary field, using methods of computational linguistics for the creation of corpora, text mining, visualisation and the analysis of various aspects (Katsios, 2024: 3), this approach allows for large-scale analysis of both narrative form and content.

It has been an enduring endeavour in this field to define the concept of the *fairy tale*, establishing its characteristic features, and distinguishing it clearly from other subgenres of folklore, such as myths, legends, and fables (Geldern-Egmond, 2000). Over time, scholars have approached this task on different levels, ranging from literary-historical definitions to content and motif related classifications (Uther, 2004), to structural models (Propp, 1968) and stylistic analyses (Lüthi, 1986).

A standard reference point for the categorisation of folktales is the Aarne–Thompson–Uther (ATU) indexing system. This system groups stories according to similar narrative concepts and themes, thereby identifying their underlying structures and enabling cross-referencing for researchers. First introduced by

Antti Aarne in 1910, the index focuses primarily on European material, as the fairy tales by the Brothers Grimm and tales by one additional Finnish writer were central to this study. The system was expanded by Stith Thompson in 1968 and 1961. Thompson also contributed the influential *Motif-Index of Folk-Literature*.

In 2004, Hans-Jörg Uther undertook a comprehensive revision, resulting in **the** Aarne–Thompson–Uther Tale Type Index (Uther, 2011). He defined motifs as the smallest recurring building blocks of a tale (Uther, 2011:10). A *tale type* is defined as a "recurring, self-sufficient plot or group of motifs" (Harvard Research guides: Library Research Guide for Folklore and Mythology: Tale-Type and Motifs, 2025), and each tale type is assigned both a number and a descriptive title. The system is organized into seven broad categories:

1-299	Animal Tales
300-749	Tales of Magic
750-849	Religious Tales
850-999	Realistic Tales
1000-1199	Tales of the Stupid Ogre
1200-1999	Anecdotes and Jokes
2000-2399	Formula Tales

All tale types are prefaced with AT or ATU, depending on whether they were originally included or later added in Uther's revision. This classification system is directly relevant to this study, as it provides a clear criterion for determining which tales can be considered as fairy tales in a folkloristic sense. Distinctions between "Animal Tales" and fables, or "Religious Tales" and legends, often times blend into each other, and distinctions are hard to make. Therefore, the index also includes tales more closely related to fables, legends, and *Schwänke* (comic tales). Within the ATU index, however, fairy tales correspond specifically to the group of "Tales of Magic" (ATU 300–749), or *Zaubermärchen* in German terminology (Carrassi, :71). By restricting my corpus to this category, it is ensured that the subsequent analysis focuses on narratives that conform to the genre-specific features of fairy tales, rather than on the wider and more heterogeneous varieties of folktales. Hans-Jörg Uther's *German Folktale Catalogue* (2015) systematises around

5,000 German-language folktales and classifies them according to the international ATU system. It provides a valuable foundation for comparative folktale research.

In their 2022 paper Werzinsky, Zhong and Zou researched on what kind of trends on themes can be found across different cultural zones. To this end, they employed the topic modelling models *BERTopic* (Grootendorst, 2022) and Latent Dirichlet Allocation to extract and cluster reoccurring topics found between regional folktale corpora. They did find that there are topics that are common across most cultures, like family, food, animals and mythological figure, but that there were also regional differences in the types of food, animal and environment. African tales for example mention jackals, Asian tales contain dragons and European tales have castles and hair colours as distinguishing themes.

While the previous approach to categorising folktales relied on the content, the Russian folklorist Vladimir Propp (1968) looked at the formal structure of this genre. In Vladimir Propp's *Morphology of the Folktale* the formalist examined the structure of Russian folktales, and, in his book, he developed a model for the systematic analysis of narrative elements. He identified 31 recurring *functions of action*, such as the hero's absence, interdiction, violation, struggle, or reward, which occur independently of the characters who perform them. While not every tale contains all of these functions, Propp observed that beneath the varying surface content of different tales there emerges an invariant deep structure of action. In addition, he defined seven central character roles (e.g., hero, villain, donor, helper, dispatcher), each of which fulfils specific narrative functions (Propp, 1968: 79f.). Like words in a language, these elements are basic units from which a story can be built (Propp, 1968: 21). The combination of functions and roles, in addition to a fixed sequence rule, thus forms a grammar that enables the analysis of fairy tales. Propp suggested that this grammar may be used to generate new tales by selecting and re-ordering functions within the grammar, all recognisable as belonging to the fairy tale genre. For the present study, Vladimir Propp's approach offers a potential method for examining both human-authored and AI-generated fairy tales with regard to their structural similarities and differences.

More recent research efforts have been put into turning this theoretical model into a practical, machine-readable, markup system. Annotation tools, like XML-Schemas, provide clear, reproducible guidelines, categories and validation rules. Once the tales are annotated, it also possible to computationally check the grammar for errors or if the permitted sequences can be verified. This can grant a valuable view into the structure of different tales and allows researchers to assess whether a given story conforms to or diverges from the structure identified by Vladimir Propp.

In 2001, Malec introduced *PftML* (Proppian fairy tale Markup Language), an early XML-Shema to manually mark Proppian functions (Malec, 2001). This work was later extended by Thierry Declerck and Hans-Jörg Scheidl (2010), who developed *APftML* (Annotated Proppian fairy tale Markup Language). Their version added linguistic layers such as syntax and semantics, which made Proppian annotation more precise and reusable. Later, Lendvai, Declerck, Darányi, and Malec refined *APftML* further in projects like CLARIN and AMICUS, showing how it could support larger, multilingual corpora in the digital humanities (Lendvai et al., 2010). In 2016, Yarlott and Finlayson introduced *ProppML*, the first fully complete annotation scheme for Proppian morphology. Unlike earlier approaches, *ProppML* was designed to capture all aspects of Propp's model, including functions, moves, and dramatis personae. The scheme was tested on fifteen tales from Propp's original corpus, producing a gold-standard dataset.

In their study *Objectivity and Reproducibility of Proppian Narrative Annotations* (2012), Bod and Fisseni conducted one of the first empirical tests of Vladimir Propp's system. They found that inter-annotator agreement was generally low, both between annotators themselves and between annotators and Vladimir Propp's original baseline annotations. The main difficulties arose from the vague descriptions of the functions and dramatis personae, which left much room for subjective interpretation. In a follow-up study, *Annotating with Propp's Morphology of the Folktale* (2014), Fisseni et al. addressed these shortcomings by providing more intensive annotator training. Their results showed, that with sufficient preparation—and under the condition that the selected tales were relatively

short—Proppian annotation can in fact be applied reproducibly. A similar conclusion was reached by Yarlott and Finlayson (2016) with the introduction of *ProppML*, a complete annotation scheme tested on fifteen tales from Vladimir Propp's original corpus. Their study likewise demonstrated that, given extensive annotator training and adjudication, Propp's system can be annotated objectively and reproducibly. As with Fisseni et al., however, the reliability of results was closely tied to the length and complexity of the tales under consideration.

These findings suggest that applying Vladimir Propp's methodology in combination with computational approaches, such as stylometry, would require a dedicated study with multiple annotators, substantial training, and a restricted focus on relatively simple tales. Only under such conditions can reproducible results be expected, which proved to be beyond the scope of this study.

While the ATU system emphasizes content and motifs, and Vladimir Propp focuses on structural functions, Max Lüthi (1909-1991) proposed a third stylistic perspective. For Max Lüthi, the fairy tale is defined less by its motifs or how its actions are sequenced, but by its form and the way it is narrated (Lüthi, 1986: 3). He identified five recurring stylistic features that set the fairy tale apart from other forms of narrative:

One Dimensionality: Max Lüthi argues that folktales portray the natural and the supernatural as a single dimension. Encounters with magical beings, enchanted objects, or otherworldly realms are narrated without hesitation or appearing strange. Characters "do not feel that an encounter with an otherworld being is an encounter with an alien dimension" (1986: 10). Whether facing witches, fairies, or talking animals, protagonists treat them as if they were no different from ordinary people (1986:6).

Depthlessness: Max Lüthi observes that folktale characters are "figures without substance, without inner life, without an environment; they lack any relation to past and future, to time altogether" (1986: 11). Sentiments are mentioned seldomly, and an atmosphere is rarely created for its own sake. Relations to family, home, or community appear only if the narrative requires them.

Abstract Style: It is argued in this work, that the folktale avoids concrete descriptions. When a hero discovers a city of iron, 'it does not waste a single word describing the iron buildings' (1986: 25). Fairy tales use a limited palette of symbolic colours, namely gold, silver, white, black, red and sometimes blue. Forests are always described as being "deep" or "large", but the colour green is very seldom mentioned. (1986: 28). Max Lüthi also notes that the folktale, as is typical for orally transmitted stories, favours formulaic numbers: one, two, three, seven, twelve (1986: 32) and strict word-for-word repetitions (1986: 33). Descriptions are often given in stark contrasted expressions in binary opposition, like ugly and beautiful, good and bad (1986: 34), aiding in the characteristic abstraction of the fairy tale tradition.

Isolations and Universal Interconnection: Characters in folktales frequently appear in isolation: heroes set out alone, helpers emerge suddenly, and encounters take place without broader social embedding. Max Lüthi notes that every encounter has a purpose, and that these isolated episodes are linked within an overarching chain of events that gives the tale cohesion. (1986: 39)

Sublimination and All-Inclusiveness: He concludes that fairy tales tend to sublimate everyday problems into a symbolic form. Everyday problems like money struggles, dangers and hunger are being translated into motifs.

Lüthi's work characterises the folk tale as a highly abstract and formulaic form of narrative. These findings will serve as one of the main pillars for the selection of genre-specific stylistic features (see SRQ1), especially in capturing the abstract style of the fairy tale genre.

3.2 GPT and Artificial Text Generation

When talking about the capabilities of any ChatGPT model, it is first important to understand what exactly happens within the model once a command is given or a question is asked. Once this input (prompt) is given to the ChatGPT model, the application sends the request to a large language model (LLM), which generates a completion. The result is then returned to the application to parse the result and then gives the user an output (cf. Berryman/ Ziegler 2023: 11). An LLM is essentially a probabilistic model. Its main task is to predict the most likely next word or sequence of words based on the input it receives (Berryman/Ziegler 2023: 4). As Berryman and Ziegler explain in their work: "At their core, LLMs are just text completion engines that mimic the text they see during their training" (2023: 4). The models learn structure, grammar, and semantics of a language through the processing of large amounts of textual data. This large-scale pre-training happens on massive, unlabelled data, from sources such as *CommonCrawl*, *Wikipedia*, and freely available digitalized books (Zhou et al., 2024: 4).

LLMs can be grouped into three main architectural categories. There are Encoder-only LLMs, which focus on only understanding tasks rather than text generation, such as *BERT* (cf. Devlin et al., 2018), *RoBERTa* (cf. Liu et al., 2019) and *DeBERTa* (cf. He et al., 2021). These LLMs are used e.g. in question answering tasks or for text classification (Bhandarkar et al., 2024). Encoder-Decoder models are built for tasks that require both understanding and text generation, such as *BART* (Lewis et al., 2019) and *ChatGLM* (Zeng et al., 2022) which are best suited for text generation and transformation, like e.g. summarising tasks, translation or text simplification. Lastly the Decoder-only models are trained on next-token-prediction (NTP) to predict the next sequence using all previously seen tokens, always working from left to right through the text (unilateral). Examples for this architecture of LLM are for one the LLaMA-series developed by Meta (Touvron et al., 2023), Gimini's Bard developed by Google and OpenAIs GPT series (cf. Zhou et al., 2024: 4).

OpenAI's first GPT model, short for Generative Pretrained Transformer, was introduced in 2018 (Radford et al., 2018; cf. Berryman/Ziegler, 2023). At this time,

the model was like a regular transformer architecture, but without its encoder, functioning solely as a decoder optimized for fine-tuned tasks.

With GPT-2, in 2019, an increase in training data and model size led to unprecedented abilities in text generation. This model often already achieved better results than state-of-the-art models that were fine-tuned specifically for the task (Berryman/Ziegler, 2023: 9).

The next generation, GPT-3, again saw an increase in training parameters and model size, which in turn led to further performance boosts in various language generation tasks (Brown et al., 2020). As Berryman highlights, this was also the point at which researchers discovered the possibilities that the purposeful modification the input could mean for the quality of the output, the prompt engineering (cf. Berryman/ Ziegler, 2023: 9). OpenAI's dialog application ChatGPT was released shortly after in November 2022, running on their GPT-3.5 model, and not too long after that, in March 2023, ChatGPT-4 was released.

Since the ChatGPT-5-main model has only been released about 1 month ago at the date this thesis, literature concerning benchmarking studies is still lacking. As OpenAI states in their GPT-5 System Card (2025), that the new model uses a hybrid architecture and has the ability to switch between a faster *main* model and a deeper *thinking* model, unlike ChatGPT-4 (2025: 4-5). According to Open AI, it also incorporates advanced safety measures, higher benchmarks performance across certain NLP tasks, and achieves lower hallucinations (2025: 16-19). Overall, each new GPT model has introduced notable improvements, largely rooted in the increase of model size and the expansion of training data.

Brown et al. (2020) demonstrated in their introduction to the GPT-3 model that large language models, and in particular this new generation of GPT, with its 175 billion parameters (2020: 5), are capable of producing coherent and contextually appropriate text by drawing on patterns learned from massive training corpora. Because the model was exposed to an exceptionally broad range of genres and disciplines, it can reproduce a wide variety of literary styles. The outputs are often polished, fluent, and stylistically versatile, to the point that many readers find them difficult to distinguish from human writing (Amirjalili, 2024: 2).

Cai et al. (2023) investigated to what extend LLMs, in their case ChatGPT and Vicuna, can actually produce human-like language patterns. They employ twelve psycholinguistic experiments that tested, among other things, syntax, word meaning, priming, preferences for word length and what inferences LLMs make. ChatGPT was found to replicate human-like language output in ten out of twelve tasks, showing its ability to approximate many aspects of human langue through the reuse of sentence structure and its flexible adaptation to meaning in context.

For this study, LLMs ability for storytelling is especially relevant. Beguš (2024) compared in her study 250 human-authored crowdsourced texts with 80 texts generated by ChatGPT-3.5 and ChatGPT-4. The methodology combined narratological approaches with statistical analyses. The findings showed that while AI can occasionally introduce innovative twists into plots, its narratives overall tend to be less imaginative. AI-generated stories were more progressive than human ones in their treatment of gender roles and sexuality, which was discovered by investigating character roles and character relationships.

Complementing the previous study Georgiou (2024) compares five human written essays and five ChatGPT-3.5 written essays in response to the same prompt and the same instructions. For the assessment of the linguistic elements the AI tool OpenBrainAI was used to extract phonological and morphological (like the ratio of parts of speech) features, syntax, lexical diversity, readability and semantics. This study proves how these particular measures can be used to discern human-written and AI-generated texts.

In a similar effort, Herbold et al. (2023) compared argumentative student essays with ChatGPT-3.5 and -4 essays. The researchers employed teachers to evaluate the essays based on, among other measures, completeness, complexity, vocabulary, and expressiveness. To supplement these assessments quantitatively, computational linguistic analysis such as lexical diversity, discourse markers and syntactic complexity where added. The AI-generated texts often received higher ratings, as they appeared more structured and coherent, while a stylistic difference was also noted. ChatGPT tended to use more normalisations and complex sentences.

Sandler et al. (2024) also examined linguistic differences but focused on dialogues. They investigated conversations between two humans and ChatGPT dialogues and drew upon the text analysis tool LIWC (Linguistic Inquiry and Word Count) to investigate categories such as authenticity and emotion. This tool works with a large dictionary that assigns words to psychologically and linguistically meaningful categories (Boyd et al., 2022). Human conversations showed greater variability and authenticity, while ChatGPT outputs were marked by a distinctly positive emotional tone and a stronger focus on social processes.

From these previous studies, several points of added value emerge for the present research. First, the analytical methods and measures they applied - such as vocabulary complexity, lexical diversity, discourse markers, syntactic complexity, and readability - demonstrate that differences between AI-generated and human texts can be made visible and quantifiable. These methodological approaches are also central to the current study, which draws on similar markers to compare classical fairy tales with ChatGPT-generated narratives. Second, the findings reviewed so far highlight that ChatGPT's output is not only distinguishable from human writing but also carries distinctive features of its own.

These observations were further built upon by Fedoriv et al. (2023). They analysed discursive and linguo-cognitive markers in narrative and descriptive texts, comparing human-written and ChatGPT-3.5 generated content. Their goal was to work out differences and similarities, especially in the use of discourse markers, such as *however*, *therefore,* and *thus,* and found that while ChatGPT is able to build coherent sentences, it produces noticeable patterns, such as an overuse of the before mentioned discourse markers. Beyond surface-level coherence and authorship markers, the capabilities of ChatGPT can also be tested within natural language processing tasks that involve the recreation of established authorial styles. The question of authorship is also central to the work of Amirjalili et al. (2024). This study compared student essays with ChatGPT texts and used the Voice Intensity Rating Scale, introduced by Helms-Park and Stapleton (2003), which measures assertiveness, self-identification, focus, and authorial presence. While ChatGPT produced coherent texts, they frequently lacked quotations, and

references, and included textual errors such as hallucinations. It was further concluded that the successfulness of the output depends heavily on the prompting process.

Bhandarkar et al. (2024) extended the investigation of style imitation to a much larger scale. In their study, they examined twelve different state-of-the-art LLMs and assessed their ability to emulate authorial styles by using five established authorship attribution techniques. The algorithms measure author discrimination on various levels, ranging from n-gram models to *BERT*-based contextual features. Their large-scale comparison confirmed earlier findings: although ChatGPT and DeepSeek performed best, they still managed to capture only about two-thirds of the original style. One notable result was that longer text samples improved stylistic imitation, whereas overly directive prompting led to a decline in performance. The authors argue that this observation occurs because "LLMs are trained with a probabilistic objective, aiming to generate the next word that most closely aligns with the preceding context, it is plausible that these models develop an understanding of the author's writing style from the example author text […]" (Bhandarkar et al., 2024: 79), and that too much external constraint through instructions can interrupt this capacity.

Central to most of these studies are stylometric approaches that rely on features such as the most frequent words (MFW). These measures form the backbone of many authorship attribution techniques and provide a methodological bridge to the following discussion of stylometric analyses.

3.3 Stylometric measures

In digital stylometry, texts or passages are compared stylistically on the basis of statistical distributions. Stylometric analysis has been used as a stable method in various classification tasks, such as authorship attribution and verification, genre classification, AI detection and forensic linguistics. Potential authors can be selected based on the choice of various features like setting or style. It is also favourable for the result to include other authors and their texts in the corpus as a comparison, in order to minimise erroneous results. Digital stylometric analysis can determine the authorship most likely to be attributed to the text, with a high success rate and reliability (Horstmann, 2024: Anwendungsbeispiel). One of the most stable and well-established metrics for the discrimination between authors is the analysis of the most frequently occurring words (MFW). Stamatatos noted that non-content words, such as articles, prepositions, and pronouns, are usually among the best features for such an approach. Especially functions words, while they do not carry any semantic information, are largely used in an unconscious manner by authors and are topic independent (cf. Stamatatos 2008: 5).

Complementarily, in his research published in 2002, Burrows states that the 30 most frequent words are the most significant for these processes. He refers to these as "markers of potentially equal power" (Burrows 2002: 271) for stylistic differences. To measure the distance between bodies of text, the distance of vectors is measured. To ensure that the most frequent words do not dominate, and thus distort, the results of the stylometric analysis, all N most frequent words must be given equal weight in the calculation, the so-called *z-score,* to ultimately, produce a numeric value. The algorithm used is called *Burrows Delta (*Burrows, 2002*)* and is the most commonly used in computer-assisted stylometry. Over time, there have been several alternative propositions to improve on the Delta measure, e.g. the *Cosine Delta* (Smith/Aldridge, 2011), which work on the same theory, but use a different metrics to measure the vector distance and has been found to be more suited for corpora with varying text lengths, as the frequency vectors are normalised, which helps with the managing of unequal texts (cf. Mikros, 2025: 593). Smaller Delta values indicate greater stylistic proximity.

In Mikros' (2025) study, ChatGPT-4 was asked to imitate the styles of Hemingway and Shelley. Different prompting strategies were tested, some without examples, others with text snippets and detailed stylistic instructions. A stylometric analysis using, among others, MFW, measured by its *Cosine Distance*, as well as n-grams, LIWC, embeddings, and token-type-ratio revealed that ChatGPT could successfully reproduce surface-level features such as sentence length and syntax, but failed to capture the deeper stylistic signature of the authors. The study also found that, given more precise instructions, the prompts improved the resulting output produced by ChatGPT-4. Mikros further concluded that structurally consistent styles like Hemingway's are easier for GPT to replicate than more richly descriptive styles (2025: 597).

Cuéllar (2024) evaluates, in his study on authorship attribution of Spanish Golden Age Theatre, the capabilities of stylometric analysis. For this, he employed the *stylo* package by Eder et al. (2016) for R Studio, testing various core features, such as MFW, culling, several classification methods, n-grams, and cross-validation, with different parameters on their attribution performance. Most methods did classify the plays correctly (>95%), and the accuracy was not greatly impacted by the amount of the MFW, giving similar results for 100 and 5,000 MFW. As for n-grams, it was uncovered that single n-grams gave the most accurate results and that the text length is detrimental for the reliability of the results. Texts with less or under 1,000 words produced fairly unreliable results and that files with over 7,500 words performed the best on accuracy. This poses a challenge for the analysis of MFW in the fairy tale genre, seeing as the genre contains a multitude of very short (< 500 words) tales (see Table 1).

However, as remarked by Lagutina et al. (2019) in their survey of stylometric text analysis, the accuracy of stylometric features in NLP tasks such as authorship attribution, authorship verification, text classification and genre classification depend heavily on the selection and correct use of the appropriate stylometric markers. The aim of their review was to propose recommendations on the use of feature extraction and to highlight the importance of certain key classes. On the character level, features such as character n-grams and punctuation; on the word

level, MFW and function word-ratios; on the syntactic level, clause and subordinate clause ratios and on the document level sentence length distributions and dialogue ratios.

While the researchers claim that the selection of the right features for each specific task is the most important step, they still make the clear point that stylometric features cannot fully capture an individual style. Stylometry only captures quantifiable textual signals, and not the literary style. The review also makes a clear point that, for reliable classification tasks, just looking at simple features like raw word frequencies are insufficient. They recommend a multi-feature approach to control topic effects and to normalise the texts for length in order to create more robust models.

4 Operational framework

In this chapter, the previously introduced concepts are translated into reproducible metrics that allow the investigation of the research question (RQ) and the hypotheses.

4.1 Corpus

First, the AI-generated fairy tales and a collection of classical German fairy tales must be converted into feasible corpora. The German folk fairy tale corpus is built from works by the Brothers Grimm and Ludwig Bechstein, which are made available through the German *Projekt Gutenberg* website[1]. This project offers a large digital library that provides free access to thou-sands of public domain books, primarily literary classics. Its texts are openly available and may be used for scholarly research, though users should ensure compliance with copyright laws in specific jurisdictions. While the website offers a great number of German literature and works by folk tale authors, this study is only interested in representatives of folk fairy tale authors.

As defined earlier, classical folk fairy tales in this study refer to folk fairy tales and, in a narrower folkloristic sense, those in the Aarne-Thompson-Uther ATU 300-749 range (Carrassi, :72), which encompass the tales of magic. This is the range researchers like Propp (1968: 19) operated in. While the folktales in the other ATU categories do contain fairy tales as well, they overlap with other folk tale genres, such as fables, legends, and "*Schwank*" (comic tales), which are not the object of this study.

The AI generated corpora were produced by OpenAI's ChatGPT[2] models, more precisely ChatGPT-4o and its successor, ChatGPT-5-main. Both models received the same prompt, including a 150-word sample from the folk fairy tale corpus as few-shot input (see Apendix). This follows findings from the work of Berryman and Ziegler (2024: 91f.) and studies on authorship emulation (Mikros, 2025; Bhandarkar et al., 2024), that few-shot prompts promote more successful

[1] https://www.projekt-gutenberg.org
[2] https://chatgpt.com/

task completion, for example, in text generation, because LLMs generalise patterns from examples to accomplish related tasks. A further reason to include a snippet of the target author's work is the idea that implicit instructions often yield better results than explicit formulations, as they encourage the model to infer patterns directly from an author's sample text (Berryman/Ziegler, 2024: 91f.).

4.2 Folkloristic Features

Significant to this study are those folk fairy tale characteristics which can be detected by computational methods and are easily quantifiable. Not all theories on what makes a fairy tale a fairy tale proposed by folkloristic researchers lend themselves to the computational capture of the *essence* of the fairy tale genre. However, Max Lüthi's emphasis on the abstract style of the folktale provides a particularly suitable approach for the development of measurable features. His descriptions often include stylistic keywords or patterns that can be systematically detected with established methods in Digital Humanities. Accordingly, the following features are examined: The following indicators are extracted from the texts:

1. **Sparse Descriptions**: According to Max Lüthi (1986), the abstract style of the folktale is characterised by a narration that avoids elaborate detail. Objects, settings, and characters are usually described in only the broadest outlines, leaving much to the imagination of the readers. As described by Max Lüthi: "Only rarely does the folktale mention sentiments or attributes for their own sake or to create a certain atmosphere. It mentions it when it influences the plot […] Attributes and sentiments are expressed in actions." (1986: 13). This stylistic tendency can be investigated in this study by looking at the overall frequency of descriptive modifiers. This is determined by analysing the proportion of adjectives in the text (POS-tagging). A low frequency of descriptive modifiers reflects Max Lüthi's notion of abstraction.
2. **Sentiment-Sparseness**: The above-mentioned absence of sentiments, descriptions and emotions is additionally examined using opinion mining methods. While sentiment analysis for the German language especially

for historic bodies of text is not as developed as for the English language, a lexicon-based approach using SentiWS[3] (Remus, et al. 2010), a German sentiment dictionary, that assigns positive, neutral, or negative sentiments, is employed in an effort to gain further insight on the sentiment sparseness.

3. **Symbolic Colours and Materials**: Characteristic for the folktale tradition is the symbolic use of colour and material adjectives such as *gold, silver, black, white, red,* and seldomly *blue*. As proposed by Max Lüthi (1986: 27-28) they carry both stylistic and symbolic significance. Their recurrence systematically extracted using pattern-based approaches (Regex).

Beyond Max Lüthi's proposition on the fairy tales abstract stylistic tendencies, further markers arise from the oral tradition of the folktale genre, which has shaped both form and style of narration. These oral features are considered here as additional, computationally feasible indicators:

4. **Formulaic Opening Phrases**: A defining feature of the folk fairy tale is the use of recurring, highly conventionalised opening phrases such as "*Es war*…" ("There was…") or the well-known "*Es war einmal*…" ("There was once…") (Petsch, 1900). These openings immediately situate the narrative within the fairy-tale tradition and signal to the reader that what follows does not belong to realistic storytelling. Their presence and recurrence are therefore not only stylistically but also structurally significant, as they frame the tale within a recognisable pattern of narration. In this study, their occurrence is systematically examined to assess the extent to which texts reproduce this characteristic marker of the genre.
5. **Prominence of Direct Speech**: One of the most important features of folk narrative art lies in its frequent use of direct speech. Its occurrence can be quantified by measuring the proportion of text enclosed in quotation marks, thereby capturing the oral-dialogic character of the genre (Freitag, 1926: 60).

[3] https://wortschatz.uni-leipzig.de/en/download

6. **Syntactic Simplicity through Sentence Fragmentation**: In oral narration, long and complex sentence structures tend to be broken down into shorter, more manageable units (Rölleke, 1975: 33). This reflects the tendency to transform complexity into accessible, sequential narration. This stylistic simplification can be measured by analysing sentence length distribution.

To test if this tendency of simplification on the syntax level also carries over into other parts of speech and influences the linguistic complexity, the following measures are investigated:

Mean Sentence Length: This calculates the average number of words in each sentence. A higher word count can indicate more complex sentence structure with subclauses in contrast to short and simple sentences.

Sentence Length Variation: This measure indicates how much sentence lengths vary throughout the text – whether all of the sentences are long or short or both. It can also show if the text follows an even pattern or has more variance.

Mean Word Length: This calculates the average number of characters per word. A higher character count can indicate a more complex style.

Mean Word Length Variation: Measures how much word lengths fluctuate throughout the text. Low variation indicates a uniform vocabulary, while high variation points to a wider range of short and long words, signalling stylistic diversity

Type-Token Ratio (TTR): These figures track lexical diversity, or how wide the vocabulary is (Richards, 1987). TTR compares the number of unique words (types) to the total number of words (tokens). This shows whether one of the corpora shows higher or lower lexical diversity, and whether one reuses vocabulary, which may indicate a more simple, formulaic style. A higher TTR value indicates a more differentiated and richer vocabulary, while a lower TTR value indicates a greater degree of repetition and more formulaic language use. (Perkuhn et al., 2012: 2).

The value calculated by this method is however dependent on corpus size. As the text size increases, the TTR becomes smaller, as the share of new words is

much higher in a small text. This poses a problem for comparative work on different sized corpora.

As a solution various normalisation techniques have been proposed, such as the standardized TTR (STTR), which will be used in this study, the corrected TTR (CTTR), and the moving-average TTR (MATTR), which all aim at reducing the dependence on text length (Covington & McFall, 2010).

Readability Index: These formulae convert complex linguistic characteristics such as sentence length, word length, and number of syllables into numerical values, which can then be used to determine and compare the difficulty of a text. Such indexes are often easy to interpret as their calculated values are typically represented using grading systems that indicate the level of reading competence. Popular and robust are e.g. the LIX score (Björnson, 1968) or the Flesch Reading Ease (Fedin, 2025), which has also been adapted to the German language by Amstad (1978).

4.3 Stylometric Features

In addition to the above developed folkloristics indicators of the abstract style and oral tradition, established stylometric measures are incorporated in addition to uncover patterns, that are not directly accessible at the surface level of textual analysis (with the exception of sentiment analysis, which is done in previously in an effort to capture the abstract style of the fairy tale genre). An analysis of the most frequently occurring words provides a systematic basis for comparing the three corpora's stylistic tendencies. Frequent word lists are extracted, and a *Principal Component Analysis* (PCA) graph is generated with R package and the *stylo package* (Eder et al., 2016). A PCA of MFW frequencies visualizes stylistic similarities and groupings across corpora.

5 Methodology

The corpora, code, prompt, figures and tables are made available under a dedicated GitHub repository[4]. During the process of this study there were several corpora collected and generated, which will be explained in the next chapter. Then a pipeline for the extraction and analysis of the folkloristic features conducted with Python and the stylometric analysis of the most frequent words conducted with the R Studio package *stylo* is implemented.

5.1 Corpus Construction

First a reference corpus of classical German folk fairy tales is collected, which forms the basis for comparison with the AI-generated material. Drawing on the data collected in this process, the prompt for the ChatGPT-5-main and ChatGPT-4o models is designed. Then one further corpus of literary fairy tales is collected as a contrast corpus for later analysis.

585 ***Spindel, Weberschiffchen und Nadel.*** Ein Königssohn will das Mädchen heiraten, das zugleich das ärmste und das reichste ist [H1311.2]. Spindel, Weberschiffchen und Nadel helfen einem tugendhaften Mädchen aus Dankbarkeit für seinen Fleiß. Die Spindel bringt den Königssohn zu ihm [D1425.1], das Weberschiffchen läßt eine magische Straße entstehen [D1484.1, D1485.1], und die Nadel verschönt das ärmliche Zimmer [D1337.1.7]. Der Königssohn heiratet das Mädchen.

Literatur/Varianten: BP III, 355; Uther 2013, 366 f.
Deutsch ab 1800: bayr.: Aurbacher 1834, 160 ff., Grimm/KHM (1857) Nr. 188 (nach Aurbacher).

Figure 1: German Fairy Tale Index (Uther, 2025: 144)

5.1.1 Construction Classical Fairy Tale Corpus

The data for the corpus of classical folk fairy tales from the ATU index can all be found within the range "tales of magic". All entries pertaining to Brother's Grimm and Ludwig Bechstein are looked up in the German catalogue "Deutscher Märchenkatalog" by Hans-Jörg Uther (2015). While this index does not offer fairy tale titles, it offers references to numbering systems and pages in fairy tale collections by German-speaking authors (see Figure 1).

[4] All Python scripts and RStudio files are freely available in the GitHub Repository: https://github.com/SophiaHeigl/Masterarbeit-Fairy-tales-vs.-AI.git
All Python Code was created, debugged or cleaned using the assistance of ChatGPT-5-main.

For Ludwig Bechstein's fairy tales, the process was the following: First the whole catalogue was searched for the references to the Bechstein fairy tale collection published by Hans- Jörg Uther in 1997, since this was the only referenced Ludwig Bechstein source I was able to acquire. While the method using this reference book yields 30 entries that could clearly be assigned to the desired range, this method has the drawback that references to tales in older works by the author are missed. The identified entries are then downloaded by hand from the collection of Bechstein's tales that *Projekt Gutenberg* provides.

The process for the Grimm fairy tales was similar, though previous research on this author pair did reduce the workload significantly. Seeing as Anti Aarrne designed the first version of the index in 1910 for the most part on the Brother's Grimm's work, their tales have been extensively mapped on the ATU-indexing system. To accelerate the selection process of the Grimm tales, a list provided by the University of Missouri Libraries[5] is employed, entailing the 200-fairy tales found in their 1857 published seventh edition of the collection with their assigned ATU number. These are then cross referenced with their original German titles and the ATU number is checked by their reference number in Uther's German fairy tale catalogue in order to verify the assigned ATU number.

Rank	File	Tokens	Words
1	Grimm_Die zwei Brüder_303.txt	9309	7717
2	Grimm_Die beiden Wanderer_613.txt	4538	3795
3	Grimm_Märchen vom einem, der auszog ...	4364	3429
	...		
121	Grimm_Die Wassernixe_316.txt	301	259
122	Grimm_Der süße Brei_565.txt	272	223
123	Grimm_Räthselmärchen_407.txt	139	119

Table 1: Corpus German_FFT entries sorted by token count

Some of the qualified tales are not suited for the corpus, since they are written in Low German dialect (Rölleke 2024: 60) that is not suited for computational analysis, e.g. "*Dat Erdmänneken*", "*Von dem Machandelboom*". Since this process is time consuming, this paper makes use of only two authors to represent the genre of the classical German folk fairy tale.

[5] https://libraryguides.missouri.edu/c.php?g=1052498&p=7642279

The entries are then saved as text files, without titles in the plain text to simplify the task of looking at opening phrases in a later chapter and are saved with the authors name, the title and the ATU number, sectioned by underscores, e.g.: Grimm_Die beiden Wanderer_613. This naming system is off use for the classification task. In a first pre-processing step, all indicators for direct speech, which varied across the different text files, were assimilated using Python code.

5.1.2 Prompt Design and Text Generation

A few-shot prompt was designed to generate imitative text from ChatGPT-5 and its predecessor ChatGPT-4o, inspired by the methodology demonstrated in Mikros' (2025) paper. For the prompting process all previous Chats were deleted in the ChatGPT settings, and the memory function was disabled, which aids in keeping the context clean and each new conversation isolated. Before applying the prompt, a new chat session was initiated in order to avoid any priming from previous results. The prompt is written in German to elicit same language fairy tales. The original German prompt with the full 150-word sample text can be found in the appendix. The sample text in the translation was collected from the English *Project Gutenberg*[6] website, which offers various collections of Grimm fairy tales by various translators and editors.

Prompt Element	Prompt Text
Introduction	You are an emulator designed to writing classical folk fairy tales.
Task Description	You are tasked with generating fairy tales. Here is a sample from a classical folk fairy tale:
Text Excerpt	"There were once upon a time two brothers, one rich and the other poor [...]"
Instructions	No title or chapter headings. 900-1900 words.
Refocus	Write a new classical folk fairy tale.

Table 2: Few-Shot Prompt

The few-shot prompt (see Table 2) design incorporates an introductory component to establish the context, a 150-word sample from the corpus of original fairy tales, as well as a refocusing element at the end, to redirect the model's attention back to the task, particularly in the case of longer prompts (Berryman/

[6] Project Gutenberg is a digital archive that offers free access to public domain texts. The collection is openly accessible and can be used for academic research, provided that copyright regulations in the respective country are observed. https://www.gutenberg.org/

Ziegler, 2024: 125). Instructions were given to the model such as to not include headings, give the word count and to generate the story within a certain word range. A range was chosen for this instruction rather than a fixed number in order to avoid imposing an artificial limitation on the model's output. This approach allowed the model some flexibility in text generation, while still ensuring that the resulting stories remained comparable in length. The range chosen was the median of the document word count of the fairy tale corpus (approximately 1437 words, rounded to 1400) and an allowance of plus and minus 500 words was given, which put the range from 900 to 1900. The median of the German fairy tales was chosen as an average due to the corpus being heavy on outliers (see Table 1) due to one over 7717-word file ("*Die zwei Brüder*"). The prompting process was repeated 100 times per model.

The output for the prompting process for the ChatGPT-5 model was saved as text files numbered from one to 100 and given the identifier ChatGPT-5 e.g.: ChatGPT-5_01. This was then repeated for the prompting of the ChatGPT-4o model accordingly e.g.: ChatGPT-4_01, which resulted in the corpora Corpus_GPT_5 and Corpus_GPT_4 respectively.

There is also a fourth corpus that was constructed for the purpose of providing a contrast corpus for the MFW analysis conducted with the R Studio *stylo* module. The selection of the authors and texts of the literary fairy tales followed the book *Das deutsche Kunstmärchen* (The German literary fairy tale) by Paul-Wolfgang Wührl (2003). There are numerous *Kunstmärchen* listed in the table of contents and those were collected, that were available on the Projekt Gutenberg website. Since this corpus serves only as a contrast for the actual comparison of folk fairy tales and AI generated tales, the corpus was kept on the smaller side, containing only thirteen tales by authors such as Willhelm Hauff, E.T.A Hoffmann, Ludwig Thiek and Johann Wolfgang von Goethe.

5.1.3 Preliminary Corpus Description

Corpus	Number of Documents	Token Count Documents	Word Count Documents	Mean Tokens/ Documents	Mean Words/ Documents	Median Words/ Documents
German_FFT	123	241159	200782	1960.64	1632.37	1437
ChatGPT_4o	100	139923	114701	1399.23	1147.01	1152
ChatGPT_5	100	152570	126151	1525.7	1261.51	1266.5

Table 3: Key Linguistic Features Corpora

Before turning to the methodology, a preliminary description of the three corpora is provided in order to present measures such as corpus size and central tendencies. The AI-generated corpora are 18.7% smaller than the corpus of original fairy tales with 123 tales and approximately 201.000 words compared to 114,701 words for the ChatGPT-4o corpus and 126,151 words in the ChatGPT-5 corpus, since the prompting process was only implemented 100 times per model. ChatGPT-4 reaches about 58% of the German_FFT corpus's token, whereas ChatGPT-5 reaches about 63% of the token count. A closer look at average document length reveals that ChatGPT-4 generated noticeably shorter tales (mean ≈ 1,147 words, 1,399 tokens) than ChatGPT-5 (mean ≈ 1,262 words, 1,526 tokens), despite both models being prompted in the same way. This suggests that ChatGPT-5 tends to produce slightly longer, and more elaborated narratives compared to its predecessor. It can also be inferred that, prompting a range of text length instead of a set number, will lead the models to produce texts on the shorter end of the spectrum. ChatGPT-4o mean document length is ~70% of the German fairy tale corpus' mean of ~1632 words, while ChatGPT-5's mean reaches ~77%.

5.2 Methode Pipeline

Parts-of-Speech Tagging

The first step in the proposed pipeline is a closer look at the parts of speech. Descriptions of characters, places and feelings play a major role in Max Lüthi's theory on the style of the folk tale, as they are decidedly lacking in this genre (1868: 34). This is expected to be reflected in the frequency and occurrence of adjectives. To extract the adjectives POS-Tagging is employed. In this process words are assigned their corresponding part of speech and then clustered in tag sets. The tag set most relevant to this study is the adjective category, though other tags may also hold significance in differentiating AI from human writing, as observed in.

For this purpose the words first have to be broken down into their basic form, so an algorithm can cluster the right words together. This can either be done by stemming or lemming (Nicky, 2020), though the later one is found to be more suitable for this pipeline, as the context of the words is important. While there are numerous modules for lemming in the English language, the selection for German language models is narrower. The *SpaCy*[7] module was chosen over alternatives like the English NLTK's *WordNetLemmtizer*[8] or the German *HanTa*[9], as it is a more common and stable module for the target language (cf. Einführung in stemming und Lemmatisierung Deutscher Texte mit Python, 2022). First, the relative distribution of POS categories (e.g., nouns, verbs, adjectives, adverbs) was measured, then the adjective forms are extracted and compared across corpora in side-by-side rankings. To visualize the grammatical composition of the three corpora a heatmap of the POS (percent scale) is created to highlight systematic deviations at a glance. This provides insight into the descriptive vocabulary used by human and machine authors, particularly since adjectives in folktales often carry symbolic weight (e.g., colour terms such as *golden*, *black*, *white*) (Lüthi, 1986: 27-28).

[7] https://spacy.io/models/de
[8] https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html
[9] https://github.com/wartaal/HanTa

Sentiment Analysis

While transformer-based approaches like BERT are state of the art (Chan et al., 2020) in the field of sentiment analysis and achieve very high accuracy in the assignment of the correct labels in specific German target corpora (Schmidt et al., 2022), they do have their drawbacks. While there are big data sets for fine-tuning large German transformer-based models like *deepset*'s gbert-base (Chan et al., 2020), such as GermEval 2017 (Wojatzki et al., 2017), which contains ~28,000 annotated post for this task, predominantly succeeded in shorter contemporary texts such as the analysis of tweets (Schmidt et al., 2022) and news (Zielinksi et al., 2023), and are not suited for unlabelled historic texts. Using this methodology the corpus was assigned mostly neutral sentiments. While there is the possibility of training *gBERT* for a few epochs on unlabelled historical texts to improve performance and adapt the model to the domain in question (Gururangan et al., 2020) it was ultimately decided against this approach. For cost efficiency a lexicon-based approach is employed using the German sentiment dictionary SentiWS[10] (Remus, et al. 2010). This lexicon was chosen because it is easily available and achieves a fairly satisfying performance, being the second best in a performance evaluation on dictionary-based sentiment review (Fehle et al., 2021: Lexicon Performance with Modifications). The study also revealed that preprocessing and lemmatizing is important for the performance, and it is recommended to add valence shifters like intensifiers (e.g. "*sehr*", "*gar*"), negations (e.g. "*nicht*", "*ohne*") and diminishers (e.g. "*kaum*") have to be added to the code as well. For a better word recognition, a normalisation pipeline is coded, which converts archaism from the text into identifiable words, such as

```
"thun":"tun"
"seyn":"sein"
"thräne":"träne"
```

Key Word Extraction and Regular Expressions

In the next step, a further look will be taken into the above-mentioned colour and material adjectives as proposed by Max Lüthi (1986). The colour scheme

[10] https://wortschatz.uni-leipzig.de/en/download

characteristic for the folk tale genre is according to Max Lüthi *gold*, *silver*, *white*, *black*, *red*, *blue*, while *green*, even though forests and nature are at times described, is not used (Lüthi, 1986: 27-28). To see if these symbolic colour patterns are reproduced by the AI-corpora, or if they can even successfully be identified in the German folk fairy tale corpus the corpora were searched for the colours "*silber*", "*gold*", "*weiß*", "*schwarz*", "*rot*", "*blau*", "*grün*", "*gelb*", "*pink*", "*lila*". Unexpected outliers are further investigated by extracting the sentences containing the outliers from all corpora.

Tri-Gramm Extraction

Opening phrases are extracted by counting the first trigrams (the first three tokens of the first sentence) in every document by corpus. To draw a comparison, alongside raw counts, the relative frequency (share) of the trigrams in every corpus is calculated by their count the number of occurrences divided by the total number of sentences that begin with the start phrase, multiplied by 100. A simple Regex expression is used to segment the sentences and token. During the process it was determined that bigrams do not capture significant enough information, while greater n-grams did not serve any more insight than the trigrams.

Direct Speech

The share of direct speech is extracted from the texts using Regex to count all tokens found between quotation marks. The share is measured by the percentage of text that is recognized as direct speech within the file. This method does not look at indirect speech. No further pre-processing is necessary here since quotation marks are already normalised in the first processing step in the corpus description (see 5.1.3). While there are several more complex machine learning algorithms trained for this task (Byszuk et al., 2020; Jannidis et al. 2018; Brunner 2013) these are mostly relevant for bodies of text for which the identification of direct speech cannot be reliably found by the use of quotation marks and punctuation alone, due to it being a historic novel for example or badly edited (Jannidis et al., 2018: Introduction). This is not the case for the corpora of this study, so a simple Regex can reliably catch all instances of direct speech.

Complexity and Readability

To investigate the assumed simplicity of FFTs tied into their roots in oral tradition. First the average of each document is calculated on the word and syntax level. Additionally, the within document standard deviation (*sd*) is calculated, which measures how much word or sentence length varies inside a typical document. This is complimented by the *sd* between documents. This will indicate how much word or sentence length varies within a document and between documents. Additionally, for a comparison of lexical richness and text length by looking at the types and tokens in the corpora. To measure the lexical diversity the calculation relies on more length robust indicator, since the corpus sizes deviate from each other. For this the standardized type-token ratio (sTTR) is chosen, which calculates the TTR over a fixed window of 1,000 tokens. It also takes into account documents, which are shorter than 1,000 tokens, which is the case with my corpus, by using the whole document for the calculation, however long it is.
To measure text difficulty the Flesch-Reading Score, accounted for the German language by Toni Amstadt (1978)

Most-Frequent-Words

The Investigation of most frequent words follows an established approach in stylometry and enables systematic identification of differences in language use. This metric has also been used by various research to determine genre or author identification (Mikros, 2025; Stamatatos, 2008). The distance metric applied was *Cosine Delta*, which has proven particularly robust in previous studies with frequency-based stylometric features in comparative text analyses (Mikros, 2025: 593; Evert et al., 2017).
Sampling and chunking was deliberately avoided, in order not to lose these short tales through segmentation or exclusion. Likewise, *culling* was set to 0. Culling refers to the exclusion of words that appear only in a small portion of the corpus (e.g., fewer than 20% of the texts). While higher culling values can help filter out noise, in my case they would have led to the loss of valuable lexical information, especially in shorter texts. For the visualization of results, I used principal component analysis (PCA) on the distance matrix. PCA reduces the large set of word

frequency features to a few main axes that capture most of the variation between texts. By plotting the first two components, the texts can be shown in a two-dimensional space: stories with similar styles appear close together, while those that differ more clearly are placed further apart. This makes it possible to see general patterns and tendencies in the corpus, for example whether human and AI-generated tales' group separately or overlap.

6 Analysis Results

The following chapter documents the result of the previously introduced methodology and proposes how these interpretations fit in with related findings in the field of computational text analysis and ties in with folkloristic theories

6.1 Folkloristic Features – Abstract style

Parts-of-Speech Tagging

Looking at the parts-of-speech distribution as seen in the heat map (see Figure 2) several trends can be observed. Both GPT corpora contain more nouns (NOUN) than the folk tale corpus, 18.2 % and 18.9 % compared to 16.3% in the FFT corpus, which is also reflected in the use of proper nouns (PROPN), which
are personal names, characters and places. GPT corpora use the units about twice as much (2% and 2.2% compared to 1%). This ties in with Max Lüthi's statement, that occurrences of real-world places and people's names are only sparsely utilized in the genre (1986: 52).

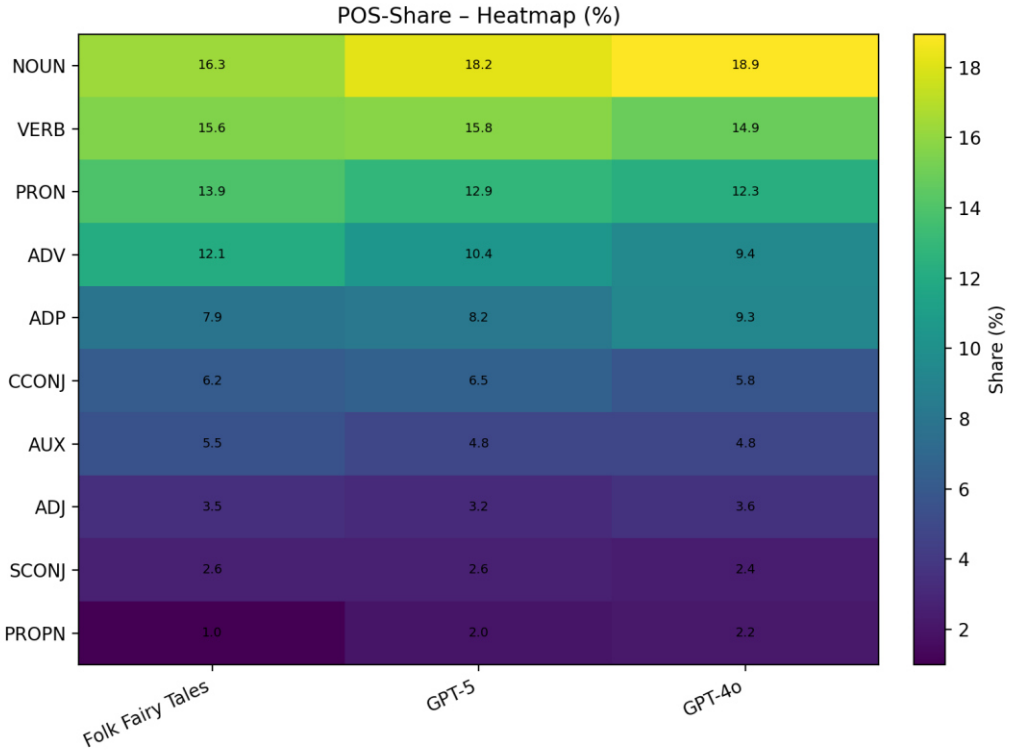


Figure 2: POS Share Heatmap (by author)

However auxiliary verbs (AUX) and pronouns (PRON) are most prevalent in the classical fairy tale corpus. The observation that GPT models tend to use normalisations was also encountered by Herbold, Hautli-Janisz, Heuer, Kikteva and Trautsch in their study on ChatGPT-3 and -4 written student essays, noting that "while ChatGPT models write more complex sentences and use more norminalizations, humans use more modals and epistemic markers" (2023: Results). It can however be observed that the more recent GPT-5 model slightly improved on this notion, being closer to the FFT corpus's noun percentage than its predecessor GPT-4o (0.7 % lower). Prepositions (ADP) are higher in ChatGPT-4o, which is consistent with the nominal style due to more prepositional phrases.

The proportion of adjectives does not offer such clear distinctions, seeing as ChatGPT-4 uses more adjectives (ADJ) than, and is closer to, the FFT corpus (3.6% versus 3.5%), while GPTT-5 uses fewer than both (3.2%). However, the type of adjectives is decisive (see Table 4). The German folk tales corpus is dominated by evaluative and typifying adjectives such as *beautiful, good, dear* ("*schön*", "*gut*", "*lieb*") as well as oposing indications of size and age *big/small, old/young ("gross"/"klein", "alt"/"jung").* These formulaic attributes furhter support Lüthi's theory on extreme contrasts in descriptions (1986:34) and of an abstract and schematic style (1986: 13).

rank	FFT	Share FFT	Count FFT	GPT-5	Share GPT-5	Count GPT-5	GPT-4o	Share GPT-4o	Count GPT-4o
1	gross	0.002275	455	alt	0.002553	322	alt	0.003801	436
2	alt	0.001725	345	klein	0.001974	249	klein	0.002058	236
3	schön	0.001605	321	gross	0.001427	180	golden	0.001221	140
4	anderer	0.00145	290	golden	0.001324	167	silbern	0.001151	132
5	klein	0.001255	251	schwarz	0.000904	114	dunkel	0.001037	119
6	gut	0.001155	231	rein	0.000832	105	jung	0.00102	117
7	jung	0.00111	222	tief	0.000674	85	schwarz	0.001003	115
8	ganz	0.001025	205	dritter	0.000658	83	erster	0.000968	111
9	golden	0.00102	204	gut	0.000626	79	nächster	0.000959	110
10	lieb	0.000855	171	nächster	0.000618	78	dritter	0.000785	90

Table 4: Most Frequent Adjectives

GPT-4o and GPT-5 partially imitate this trend. Words such as *old*, *big*, *small*, and *good* are also present in their most used adjectives, but they significantly expand on these with ordinal markers, such as *first*, *next*, *third* ("*erster*", "*nächster*", "*dritter*"). While both models use *old* and *small* as their most frequent adjectives, it is notable that proportionally, GPT-4o uses *old* and *small* very frequently, even

compared to the GPT-5 model. This may indicate a less variation in their use of adjectives.

Sentiment Analysis

In the next step, a text mining approach is used, namely the designation of positive or negative sentiments, using SentiWS, on the three corpora. The goal is to ascertain if Max Lüthi's proposed sentiment sparseness is reflected on the FFT corpus and if noticeable insights in trends between both GPT models can be drawn.

corpus	Pos Label %	Neg Label %	Confident Pos Label %	Confident Neg Lable %
Folk Fairy Tales	26,22	22,56	0,17	0,40
GPT-4o	19,48	18,87	0,15	0,56
GPT-5	24,97	21,42	0,20	0,56

Table 5: SentiWS Sentiment Summary across all Corpora

It should be noted that there is a slight positivity in all three corpora. The proportion of positive tokens exceeds that of negative tokens by ~1.0% - 3.5% in each case. Furthermore, there are only a few extreme values, as the words that could be confidently assigned to either the positive or negative category are mostly well below 1%. Thus, the choice of words is not particularly emotionally charged. With regard to LLM specifics, it can be observed that GPT corpora show slightly higher proportions of strongly negative tokens (0.56%) compared to FFTs (0.40%), which means that when LLMs use negative ratings, these are relatively more often clearly coded as negative (e.g. "*arm*", "poor", "*verschwinden*", "disappear", "*brechen*", "break") (see Table 6), but at the same time less frequent overall than in FFTs. Overall, GPT-5 is closer to German fairy tales than GPT-4o, suggesting that it is closer to the reference register.

According to Max Lüthi (1986), folk tales are typically written in an abstract style in which feelings are rarely explicitly named but rather conveyed through the plot and actions (Lüthi, 1986: 38), and the low confidence rates observed support this characterisation, suggesting that extreme evaluative adjectives are rare.

	Folk Fairy Tales	GPT-5	GPT-4o
1	arm	klein	klein
2	klein	arm	verschwinden
3	schloss	verschwinden	arm
4	fallen	brechen	fremde
5	brechen	schwach	brechen
6	angst	schloss	fallen
7	schlecht	grau	grau
8	verschwinden	fremde	verlieren
9	falsch	fallen	schwach
10	fehlen	verlieren	krank

Table 6: Top 10 Negative Sentiments

At the same time, folk fairy tales have a higher proportion of positive/negative lexicon than GPT-4o, which is consistent with formulaic evaluation patterns (*good/evil*, *big/small*) and moral division into two clearly separate groups (Lüthi: 1986: 38). GPT-5 is closer to the FFT proportions than GPT-4o and thus agrees with the other stylometric findings regarding better stylistic imitation. While these results may indicate some tendencies, this is by far not a very efficient method for this investigation as lexicon-based approach on sentiment analysis are lacking behind its transformer-based competitors (Fehle, 2021).

Colour scheme

Max Lüthi's proposed colour scheme (1986: 27-28) can already be detected in the most frequently used adjectives (*"golden"*, *"schwarz"*, *"silbern"*) (see Table 4). A closer look at occurrences of colour adjectives in the corpora (see Table 8) reveal that, while all three corpora do mostly adhere to *silver*, *gold*, *black*, *white* and *red*, there are also occurrences of *green* and *yellow*. To interpret these deviations from the expected findings, the sentences where these lemmas were detected are extracted from the text[11]. The colour *yellow*, especially in Grimm's fairy tales is used to visualise emotions: "*Da erschrak die Königin und wurd grün und gelb vor Neid*" (Grimm, KHM 19), where the Queen turns green and yellow in envy, while both GPT models use it to denote eye colour[12] or the colour of leaves[13].

[11] Find the tables here in the GitHub Repository:
[12] ChatGPT-5_86, ChatGPT-4o_80
[13] ChatGPT-4o_09

The use of colour words shows clear distinctions between the corpora. While all three corpora use the colour *gold* the most often, the AI models use the term significantly more often (0.17% for the original fairy tale corpus versus 0.24% for GPT-4o and 0.32% for GPT-5.) This overrepresentation of certain colours can also be found in the colour *silver* (0.02% for FFTs versus 0.11% for GPT-5 and 0.16% for GPT-4o) and the colour *black* (0.04% for FFTs versus 0.11% for GPT-5 and 0.14% for GPT-4o), which is also represented in the most frequently used adjectives (see Table 4). The only colour for which this is the opposite is the colours *red*

corpus	tokens	silber Count	silber %	gold Count	gold %	schwarz Count	schwarz %	weiss Count	weiss %	rot Count	rot %	blau Count	blau %	gruen Count	gruen %	gelb count	gelb %
FFT	200034	46	0,02	338	0,17	81	0,04	100	0,05	72	0,04	24	0,01	40	0,02	7	0,0
GPT-4o	114699	181	0,16	273	0,24	164	0,14	49	0,04	32	0,03	25	0,02	46	0,04	3	0,0
GPT-5	126146	133	0,11	402	0,32	141	0,11	78	0,06	29	0,02	15	0,01	31	0,02	2	0,0

Table 7: Colour Lemma Counts and Percent normalized by 1K

(0.04% for FFTs versus 0.03% for GPT-4o and 0.02% for GPT-5). In light of Max Lüthi's (1986) findings, this profile supports two points: First, the formulaic symbolism of fairy tale colours and a reduced colour pallet imitated, but exaggerated ornamentally, especially through the strong preference for *gold*/*silver* in the AI texts. Taken together, GPT texts use around twice as many metallic colour references as folk tales (≈ 0.40–0.43% vs. 0.19%).

And secondly, the data suggests that, while the GPT models mirror the stylistic colour scheme, they do not capture the symbolic meaning and merely overinflate the words that are "typical" for the fairy tale genre. This is also supported by the data from the most frequent adjectives (see Table 4), seeing as it is very top-heavy.

6.2 Folkloristic Features – Oral tradition

The first three tokens from the first sentence of each text were extracted as a trigram and frequency-ranked across the corpus; for the two canonical formulas "*Es war einmal*" ("once upon a time") and "*Es waren einmal*" ("there once were") the

trigram	Folk Fairy Tales	Folk Fairy Tales (%)	GPT-5	GPT-5 (%)	GPT-4o	GPT-4o (%)
es war einmal	54	44.6	43	43.0	53	53.0
es waren einmal	5	4.1	57	57.0	47	47.0
es war ein	6	5.0	0	0.0	0	0.0
es lebte einmal	3	2.5	0	0.0	0	0.0
vor alten zeiten	2	1.7	0	0.0	0	0.0
es war eine	2	1.7	0	0.0	0	0.0
es hatte ein	2	1.7	0	0.0	0	0.0
es war vor	1	0.8	0	0.0	0	0.0
in der schweiz	1	0.8	0	0.0	0	0.0
ein schneider und	1	0.8	0	0.0	0	0.0

Table 8: Top 10 excerpt from document start trigrams

immediate continuation (the next three tokens) was also profiled. The results show a clear fixation on the canon in the AI corpora (see Table 9). In GPT-5, 43% of the texts begin with "Once upon a time" and 57% with "There was once." In GPT-4o, the figures are 53% and 47%, respectively. The situation is different for fairy tales: the formula is used frequently, but not exclusively; "once upon a time" occurs in 44.6% of cases, "there were once" in 4.1%. The difference (51.3%) demonstrates a wide variety of alternative beginnings such as "there once lived," "in ancient times," and "there was a." The AI texts thus narrow the repertoire to almost exclusively canonical openings, while historical fairy tales exhibit a significantly more variable opening pattern.

next trigrams	GPT-5 (n)	GPT-5 (%)	GPT-4o (n)	GPT-4o (%)
ein alter müller	12	27.9	22	41.5
ein müller der	10	23.3	17	32.1
ein armer müller	2	4.7	4	7.5
ein könig der	3	7.0	0	0.0
ein bauer der	3	7.0	0	0.0
ein altes weib	0	0.0	2	3.8
ein altes ehepaar	1	2.3	1	1.9
ein armer köhler	1	2.3	1	1.9
ein armer hirte	0	0.0	2	3.8
ein alter bauer	0	0.0	2	3.8

Table 9: Top 10 following Trigrams "*Es war einmal*"

The continuations after the canonical beginnings are also highly stereotyped in the AI corpora. "Once upon a time" is immediately followed by typical role

assignments: an “old miller” (GPT-5: 27.9%; GPT-4o: 41.5%) and a “miller” (GPT-5: 23.3%; GPT-4o: 32.1%), deviating (“poor”) variants e.g. a “poor miller”, “poor charcoal burner” or a “poor shepherd” and isolated short introductions “a king that” or "a farmer that”.

next trigrams	GPT-5 (n)	GPT-5 (%)	GPT-4o (n)	GPT-4o (%)
ein müller und	13	22.8	22	46.8
ein alter müller	9	15.8	22	46.8
drei schwestern die	6	10.5	1	2.1
ein armer müller	6	10.5	0	0.0
zwei schwestern die	6	10.5	0	0.0
ein König und	3	5.3	1	2.1
ein mann und	3	5.3	0	0.0
in einem stillen	2	3.5	0	0.0
in einem fernen	2	3.5	0	0.0
in einem tiefen	2	3.5	0	0.0

Table 10: Top 10 following Trigrams "*Es waren einmal*"

In the GPT-4o corpus, a miller and an old miller each account for 46.8%. In the GPT-5 corpus, additional opening sequences appear, such as “three sisters” and “two sisters” (10.5% each). Two more opening sequences appear in the GPT-4o generated tales as well, but overall, the opening formula in all corpora follow a highly repetitive pattern.

corpus	docs	Mean Direct Speech %	Min %	Max %
Folk Fairy Tales	122	26,32	1,15	71,99
GPT-4o	100	19,24	5,40	35,38
GPT-5	100	20,80	8,66	32,69

Table 11: Prominence of Direct Speech

A comparison of the corpora reveals a clear pattern in favour of fairy tales in terms of the proportion of direct speech. The percentage of characters within quotation marks per document was measured (after standardising typographical variants) and then combined. In the folk tale corpus, the mean value for direct speech (see Table 12) is 26.32% with a considerable in corpus dispersion, as the percentage occurs in a range between only 1.15% and up to 71.99% of direct speech percentage within one tale. In the GPT-4o corpus it is 19.24% with a much smaller dispersion (5.40–35.38%). GPT-5 performs close to its predecessor with 20.80% and a similar dispersion range (8.66–32.69%). Folk tales thus outperform LLM texts by an average of +7.08 percentage points compared to GPT-4o and +5.52

percentage points compared to GPT-5; in relative terms, this corresponds to approximately +37% and +26% more direct speech, respectively.

In addition, the greater range in the folk tale corpus indicating higher stylistic heterogeneity, compared to both GPT models, which can be observed in a dispersion box plot (see Appendix). This result supports the thesis of a stronger oral-narrative orientation of folk tales: direct speech reduces the role of a commenting narrator and externalises actions and speech acts, which is for one consistent with Lüthi's description that sentiments but manifest themselves in action and speech (cf. Lüthi 1986: 38).

Syntactic Simplicity

On average, FFTs have significantly longer sentence structures than the two AI models (GPT-4o: 14.03 words; GPT-5: 14.68 words), with 20.33 words per sentence (see Table 13). The variance is also significantly higher within German fairy tales (11.27 compared to 7.54 and 6.10, respectively). This indicates greater heterogeneity: the texts contain short, fragmentary sentences as well as significantly longer, nested sentence structures. In comparison, GPT models work with more consistent sentence lengths and thus with a more uniform syntactic structure. There is a stronger tendency towards stylistic uniformity and readability in the AI texts, while the fairy tales fluctuate more between simple and complex syntax.

This supports the finding that Sandler, Choung Ross and David (2024: 1) published on the comparison of human and ChatGPT conversations, where they discovered that conversations between humans showed greater variability and authenticity, while the GPT excelled at positive emotional tone and analytical style. (2024:3).

Author/Group	N_docs	Mean word length (chars) (mean ± sd)	Word length variation (sd)	Mean sentence length (words) (mean ± sd)	Sentence length variation (sd)
Overall mean	322	4.58 ± 0.13	2.12	16.62 ± 3.55	8.76
Folk Fairy Tales	122	4.62 ± 0.19	2.25	20.33 ± 2.72	11.27
GPT-4o	100	4.54 ± 0.08	2.01	14.03 ± 1.53	7.54
GPT-5	100	4.59 ± 0.08	2.08	14.68 ± 1.36	6.10

Table 12: Sentence and Word Level complexity

The average word length for all corpora is similar, ranging between 4.54 and 4.62 characters per word (See table 14). However, it is noticeable that FFTs show a slightly higher variance in word length compared to Chat-GPT (2.25 compared

to 2.0 - 2.1 for GPT texts). This could be interpreted as an indication of a wider range of vocabulary used, from short, simple words to longer and more complex lexemes. On average, folk tales are significantly longer at the document level, with around 1,636 words, than the texts generated by AI models, GPT-4o with approximately 1,147 words and GPT-5 approximately 1,261 words. The absolute number of different word forms (types) is also highest in folk tales, at 555 compared to 484 and 527 respectively. This means that AI texts have a more even distribution of vocabulary in relation to text length and repeat the same words less frequently.

Author/Group	N_docs	Tokens (mean ± sd)	Types (mean ± sd)	stTTR % (mean ± sd)
Overall mean	322.0	1367.63 ± 622.34	524.07 ± 136.86	53.12 ± 7.61
Folk Fairy Tales	122.0	1635.79 ± 937.51	555.03 ± 210.86	47.36 ± 6.94
GPT-4o	100.0	1146.79 ± 99.41	483.77 ± 32.68	57.87 ± 5.95
GPT-5	100.0	1261.31 ± 149.65	526.60 ± 49.93	55.39 ± 4.98

Table 13: Lexical Diversity

The established readability indices show and confirm the impression of greater linguistic simplicity in the AI-generated texts. The *Flesch Reading Ease* (FRE) score is 81.71 for GPT-4o and 79.48 for GPT-5, while the German FFTs score significantly lower at 70.41. The LIX value also shows a clear difference with FFTs at a 38.93, GPT-4o with 28.82 and GPT-5 with 31.24. The lower the value, the easier the texts are to understand, so the generated texts are easier to understand, while original German fairy tales have a higher linguistic complexity.

corpus	n_docs	Avrg.Sentence Lenght (mean)	FRE (mean)	LIX (mean)	Words/ doc (mean)	Sentences/ doc (mean)
Folk Fairy Tales	122	22.75	70.41	38.93	1639.07	74.43
GPT-4o	100	14.19	81.71	28.82	1146.99	81.98
GPT-5	100	15.41	79.48	31.24	1261.46	82.8

Table 14: Readability

It can therefore be concluded that folk tales are characterised by longer and more variable sentence structures, as well as a higher repetition of vocabulary. This finding is consistent with the traditional, oral style of storytelling, which is characterised by formulaic language and episodic variability. In contrast, the AI models produced texts that are shorter, more consistent and easier to read. Despite the example fairy tale in the prompt used to generate the AI corpus, the

output texts are more modern in terms of language and are designed for consistency and comprehensibility. This reflects the fundamental differences between traditional oral storytelling forms and the practice of language models

6.3 Stylometric Features

The results of the MFW analysis reveal clear differences between human written folk fairy tales and the AI generated ones. Within the PCA visualisation (see Figure 3) the texts generated by GPT-4o stand out as a separate group in the upper part of the plot, while the texts from GPT-5 are more strongly concentrated in the middle. The fairy tales form a clearly distinct group in the lower part of the plot. There is a zone of overlap between the German FFTs and the ChatGPT generated ones, indicating that have a partially shared stylistic feature distribution, while there is a comparatively high discrimination between the classic human fairy

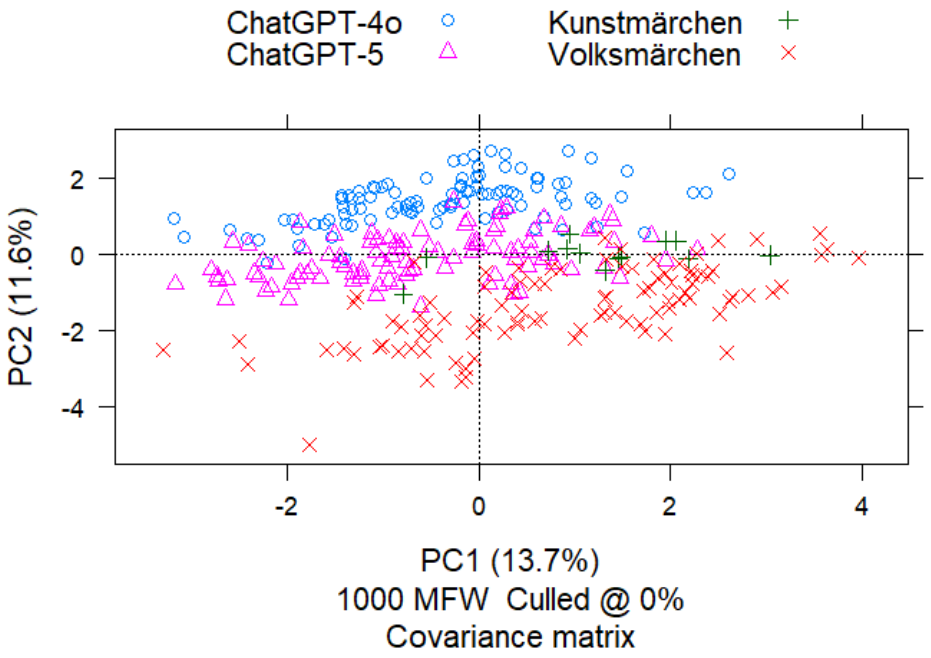


Figure 3: Principle Component Analysis MFW (by auhtor)

tales and the GPT-4o ones. Furthermore, the less compact literary fairy tales are situated in the middle, overlapping with all other corpora. This shows that the AI texts can be clearly distinguished stylistically from the folk fairy tales, while GPT-5 share similarities with the folk fairy tales.

This observation is supported by the distance measures of the cosine delta values. The smallest distance is between literary and folk tales (0.7367), underlining the relative proximity within the human text groups. Overall, the distances between the GPT models and the two fairy tale corpora are higher, but still within a similar range (0.79-0.82). It is striking that GPT-4o is slightly closer to folk tales (0.8202) than GPT-5 (0.8019), while the values remain almost identical in comparison to literary fairy tales (0.8117 vs. 0.7920).

GroupA	GroupB	Cosine Delta Mean
ChatGPT-4o	ChatGPT-5	0.7581081
ChatGPT-4o	Kunstmärchen	0.8117171
ChatGPT-4o	Volksmärchen	0.8201548
ChatGPT-5	Kunstmärchen	0.7920126
ChatGPT-5	Volksmärchen	0.8018917
Kunstmärchen	Volksmärchen	0.7367486

Table 15: Cosine Delta

The results of the MFW analysis reveal clear differences between the corpora examined. Within the PCA visualisation, the texts generated by GPT-4o stand out as a separate group in the upper part of the plot, while the texts from GPT-5 are more strongly concentrated in the middle. The folk tales form a clearly distinct group in the lower part of the plot. Furthermore, there are differences between the less compact. The results suggest that AI-generated texts develop linguistic patterns that are more closely aligned with literary fairy tales. The initial assumption that the example text in the prompt for generating AI texts would result in an imitation of the folk narrative style could not be confirmed. The models adopt stylistic characteristics from literary text traditions, but the oral-traditional structures of folk tales are less well captured. This confirms that GPT-generated fairy tales do establish a certain stylistic proximity to literary fairy tales but maintain a significant distance from folk tales.

7 Discussion

If the present study proofs one point, then that it is not an easy endeavour to capture the essence of the fairy tale. Various approaches have been undertaken by scholars to break this genre down into understandable and discernible features, from reducing them to their smallest building brick: the motifs (Uther, 2004), to structural-morphological ones (Propp, 1968) to stylistic ones that describe the essence of its subliminal and abstract style. The indicators selected for this study, while of course not be able covering all aspects and that also was never the aim.

The computational indicators used in this study, namely the proportion of direct speech, variance in sentence length, parts-of-speech profiles, readability indices (FRE/LIX), most frequent words, and stylometric distances (cosine delta, PCA) translate these folkloric descriptions into replicable metrics. The combination of dialogue proportion and measures of dispersion (sentence length, word length) in particular shows that folk tales not only have different mean values, but above all greater heterogeneity. This supports the thesis that oral tradition does not merely produce conventionalized formulas, but rather narratively varying practices that are reflected in high intra- and intertextual variance.

When compared with current research on LLM texts, the findings are consistent: AI texts are highly readable, stylistically smooth, and more homogeneous. The higher readability and lower variance (especially in sentence length) demonstrated by the GPT corpora are consistent with reports that characterize LLM texts as less variable and creative (Herbold et al., 2023) At the same time, stylometric methods confirm the continuing differentiability: in PCA, GPT-4o and to a lesser extent GPT-5 form independent, model-typical clusters. This work thus confirms a widespread finding: LLMs reliably emulate surface patterns, but do not achieve the dialogical density and varied design of human speech and writing (Sandler et al., 2024).

Reporting on the key differences worked out through the proposed pipeline it can be said that folk tales show a higher proportion of direct speech (on average > 26%, with a very wide range up to > 70%), while GPT-4o and GPT-5 clearly fall

below this (≈ 19–21%). It is not only the mean value that is decisive, but also the dispersion, showing more between document creativity.

The folk fairy tales have longer and more variable sentences (Ø ≈ 20 words; significantly increased standard deviation). AI texts are more uniform (Ø ≈ 14–15 words), which contributes to higher readability but also to rhythm smoothing: The PCA places GPT-4o clearly separated, while GPT-5 more centrally with partial overlap with folk tales. Cosine delta relativizes this and as expected, the distances are smallest within the human corpora. The AI reflects key contrasting adjectives and color schemes (e.g., old, small; gold, silver, white) but tends to overemphasize individual high-frequency markers (especially GPT-4o). In a similar fashion ChatGPT 4 not only heavily overuses canonical colors, but it also heavily overuses its most frequent adjective.

The work shows with consistent evidence: LLMs capture the form, but not the oral nature. They reproduce iconic surface markers of the genre like formulaic beginnings, elementary contrasting adjectives, color fields and increase readability; at the same time, they remain more stylistically homogeneous, rhythmically smoothed, and lacking in dialogue compared to folk tales. Stylometrically, this results in a measurable distance that GPT-5 reduces in some aspects without eliminating it. Methodologically, the study establishes a transferable toolkit for digital humanities that operationalizes folkloric theory and thus precisely reveals the limits and possibilities of AI-based genre imitation. For the practice of creative writing, this results in concrete levers (prompt design, dialogue induction, entry variation, post-editing) to bring AI texts closer to the richly varied orality of folk tales—a path that should now be pursued further in breadth (corpora, languages) and depth (structure, effect, perception).

8 Conclusion

This study investigated the capabilities of ChatGPT-4o and the, in August 2025 released, ChatGPT-5 model on their capabilities in reproducing specific genre style, in this case German folk fairy tales. To assess if the metrics derived from folklore research yield significant results, the research questions will be addressed.

The first (SRQ1) line of inquiry was to ascertain which genre-specific stylistic markers of abstraction, as identified in folkloristic theory by Max Lüthi (1986), are preserved or altered in AI-generated fairy tales compared to German folk fairy tales. AI adopts typical surface features e.g., formulaic beginnings ("*Es war einmal*", "Once upon a time"), and canonical color and adjective patterns but does not fully achieve the condensation of variation the original tales achieve. GPT models also tend to overuse these features.

The next question was to investigate to what extend ChatGPT-4o and -5 generated fairy tales reflect features of oral tradition characteristic of the genre (SQ2)? The frequency of direct speech and the greater variation in sentence length remain hallmarks of folk tales; AI texts reproduce these markers of orality only to a limited extent.

The last point of investigation was to measure to what extent do ChatGPT-4o and ChatGPT-4 generated fairy tales statistically diverge from or resemble classical German fairy tales when compared through word frequency analysis (SRQ3)? The PCA and Cosine Delta reliably separate the corpora; the smallest distance is found between the two human corpora, while GPT-4o and GPT-5 remain at a distance (range ≈ 0.79–0.82). However, in the PCA graph there is an overlap between the ChatGPT-5 and the FFT corpus, signifying that

The main research question remains**:** To what extent do fairy tales generated by ChatGPT-4o and ChatGPT-5 reproduce the stylistic markers and oral-tradition features characteristic of the classical German folk fairy tale, and how do they compare to human-written tales on a deeper stylistic level?

The study reveals that folk tales are more dialogical and rhythmically variable, measured by their share of direct speech in the text and by looking at the variance

in sentence, word and token length. AI texts remain syntactically smooth and more uniform. The pipeline also addressed Readability, measuring the difficulty and reading level of the corpora. It was shown that GPT-4o and -5 systematically generate more readable texts with a higher FRE and lower LIX than the reference corpora of folk tales. To further quantify the distance between the three corpora, the targets were loaded into the RStudio package *stylo*. Looking at MFW, models form model-typical clusters. While the GPT-5 cluster moves partially toward folk tales, the GPT-4o model remains more clearly separated, but clearly overlaps with the ChatGPT-5 models' cluster. The cosine delta distances confirm a measurable difference between AI and human texts. In summary: The models reliably reproduce surface-level genre markers but fail to capture the dialogic density and heterogeneous narrative rhythms of folk tales.

The key findings are for one, that both GPTs match canonical signals (opening formulas, color fields, frequent adjectives), but at the same time produce stylistic homogenization: sentence rhythm, dialogue proportions, and variation are reduced. The models thus appear "smoother" rather than oral storytelling tone.

The results highlight the difference between superficial imitation and oral storytelling practice: ChatGPT-4o and ChatGPT-5 reliably reproduces iconic markers (colours, contrasting adjectives, formulaic beginnings), the folk tales' notion to situational variability including in dialogue and sentence rhythm remains underrepresented. This results in an empirically reliable contrast between genre profiles that quantitatively supports the theoretical assumptions of folklore studies.

To gain further insights, in the following the hypotheses that were set at the beginning of this thesis will be validated or falsified. The first hypothesis stated that since LLMs excel at recognizing patterns (Brown et al., 2020: 34; Berryman/ Ziegler, 2024: 4), it is expected that the generated tales will reproduce formulaic features of fairy tale style, such opening phrases like "There once was…" ("*Es war einmal*…"). Looking at the data collected, the hypothesis can be positively verified, seeing as both GPT corpora started with the canonical story openers 100%

of the time. they also adhere to the restricted colour palette, although the overused “gold”, “silver” and “black”.

The second hypothesis assumed that the more recent ChatGPT model would perform better, or in the context of the emulation task closer to the target corpus. This was true for most of the measures investigated in this study, like e.g. sentence length variability and nominalisation. This would indicate that the newer model outperforms its predecessor on several tasks, however, when looking at the cosine delta, ChatGPT-4o is measurably closer to the target author. Seeing as the cosine is the measure set for this decision, the hypothesis is falsified for now.

In light of this, when looking at the third hypothesis, we will measure the sTTR and the Readability. When looking at the token and types, AI texts have a more even distribution of vocabulary in relation to text length and repeat the same words less frequently. They work with more consistent word lengths which would indicate a lower variability than the original fairy tales. If we are taking the Flesch Reading Ease and LIX into account to additionally answer the last hypothesis, it can be claimed that the AI produced texts are shorter, more consistent and easier to read, which would verify both hypotheses.

As for limitations that were encountered during the process: the decision to work with German language fairy tales did influence the process a lot. Seeing as most English language libraries and coding packages are more established, accessible and successful, it would be interesting to re-conduct the pipeline with English translated international fairy tales, especially the sentiment lexica are only partially optimal for historical German genres. By narrowing the corpus to only fairy tales from a single region and cultural zone, other narratives are missed.
The biggest challenge to work within the pipeline was the length dependency of many tools and workarounds that had to be implemented : Despite careful parameterization, short texts remain a challenge for the stability and generalizability of stylometric measures.

One further aspect that could not get the attention it deserved within this thesis was the inclusion of the morphological level for narrative analysis. While the

challenges of conducting a proper annotation study deserves its own dedicated work, it was ultimately decided against integrating the formal-structural level into this work. A multi-annotator setting with extensive training, guidelines and reliability checks supplement studies into the form and structure of AI-generated fairy tales. It would be interesting to see if LLMs pattern recognition is developed to catch such underlying actions. Setting up automated approaches (e.g., sequence labeling for Propp functions) would be a medium-term prospect.

While this study focused on the German fairy tale tradition, a multilingual, ATU-wide fairy tale corpus beyond the German canon can clarify whether the observed patterns are language or culture-specific or inherent to the model. Traditions with different dialogue conventions and narrative rhythms would be particularly interesting. Werzinsky et al. (2022) approached this idea from a content perspective: their topic modeling approach revealed separate and shared thematic trend across different cultures. It would be interesting to observe differences on a lexical level or readability level.

Lastly, a reader study on the perception of such AI fairy tales would be of great value, seeing as the accessibility of such story generators through apps and web application the endless possibilities could also either pose a great advantage for early childhood development or children's education.

9 References

Aarne, Antti (1928): *The Types of the Folk-Tale: A Classification and Bibliography*, Translated Stith Thompson, Helsinki: Suomalainen Tiedeakatemia (Folklore Fellows' Communications 74), [online] https://en.wikisource.org/wiki/File:Antti_Aarne_and_Stith_Thompson_-_The_Types_of_the_Folk-Tale_(1928).pdf [accessed on 18.09.2025].

Amirjalili, Forough / Neysani, Masoud and Nikbakht, Ahmadreza (2024): "Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature", in: *Frontiers in Education*, vol. 9, DOI: https://doi.org/10.3389/feduc.2024.1347421.

Amstad, T. (1978): *Wie verständlich sind unsere Zeitungen?* Dissertation, Universität Zürich. Zürich: Studenten-Schreib-Service.

Baumgärtner, Alfred Clemens (1968): *Märchen und Sage. Grundzüge ihrer Struktur und ihrer Behandlung im Unterricht*, 2.Auflage, Frankfurt am Main: Moritz Diesterweg.

Bechstein, Ludwig (1853): *Ludwig Bechstein's Märchenbuch*, Leipzig: Georg Wigand, [online] https://archive.org/details/ludwigbechsteins00bech/page/n9/mode/2up [accessed on 16.09.2025].

Beguš, Nina (2024): "Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling", in: *Humanities and Social Sciences Communications*, vol. 11, artical. 1392. DOI: https://doi.org/10.1057/s41599-024-03868-8.

Bensaid, Eden / Martino, Mauro / Hoover, Benjamin and Strobelt, Hendrik (2021): "FairyTailor: A multimodal generative Framework for storytelling", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2108.04324.

Berryman, John and Ziegler, Albert (2024): *Prompt Engineering for LLMs: The Art and Science of Building Large Language Model-Based Applications*, Sebastopol: O'Reilly Media, Inc.

Bhandarkar, Avanti / Wilson, Ronald / Swarup, Anushka and Woodard, Damon (2024): "Emulating Author Style: A Feasibility Study of Prompt-enabled Text Stylization with Off-the-Shelf LLMs", in: Ameet, Deshpande / Eun-Jeong, Hwang / Vishvak, Murahari / Joon Sung, Park / Diyi, Yang / Ashish, Sabharwal / Karthik, Narasimhan and Ashwin, Kalyan (editors), Proceedings of the 1st Workshop on Personalization of Generative AI Systems (Personalize 2024), St. Julians, Malta: Association for Computational Linguistics, pp. 76–82. [online] https://aclanthology.org/2024.personalize-1.6/ [accessed on 17.09.2025].

Björnson, Carl-Hugo (1968): *Läsbarhet*. Stockholm: Liber

Bleek, Wilhelm (1986): "Die Brüder Grimm und die deutsche Politik", in: *Aus Politik und Zeitgeschichte*, no. 1/1986, [online] https://www.bpb.de/shop/zeitschriften/apuz/archiv/533630/die-brueder-grimm-und-die-deutsche-politik/ [accessed on 19.09.2025].

Bod, Rens / Fisseni, Bernhard / Kurji, Aadil and Löwe, Benedikt (2012): "Objectivity and Reproducibility of Proppian Narrative Annotations", in: Mark A., Finlayson (editor.), *Proceedings of the Third Workshop on Computational Models of Narrative (CMN '12)*, Istanbul, pp. 15–19. [online] https://d-nb.info/1171212291/34 [accessed on 20.09.2025].

Boyd, Ryan L. / Ashokkumar, Ashwini / Seraj, Sarah and Pennebaker, James W. (2022): "The Development and Psychometric Properties of LIWC-22", Austin, TX: University of Texas at Austin, [online] https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf [accessed on 20.09.2025].

Brown, Tom B. / Mann, Benjamin / Ryder, Nick / Subbiah, Melanie / Kaplan, Jared / Dhariwal, Prafulla / Neelakantan, Arvind / Shyam, Pranav / Sastry, Girish / Askell, Amanda / Agarwal, Sandhini / Herbert-Voss, Ariel / Krueger, Gretchen / Henighan, Tom / Child, Rewon / Ramesh, Aditya / Ziegler, Daniel M. / Wu, Jeffrey / Winter, Clemens / Hesse, Christopher / Chen, Mark / Sigler, Eric / Litwin, Mateusz / Gray, Scott / Chess, Benjamin /

Clark, Jack / Berner, Christopher / McCandlish, Sam / Radford, Alec / Sutskever, Ilya and Amodei, Dario (2020): "Language Models are Few-Shot Learners", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2005.14165.

Brunner, Annelen (2013): "Automatic recognition of speech, thought, and writing representation in German narrative texts", in: *Literary and Linguistic Computing*, vol. 28, iss. 4, pp. 563–575, DOI: https://doi.org/10.1093/llc/fqt024.

Burrows, John (2002): "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing*, vol. 17, iss .3, pp. 267–287. DOI: https://doi.org/10.1093/llc/17.3.267.

Byszuk, Joanna / Woźniak, Mikołaj / Kestemont, Mike / Leśniak, Agata / Łukasik, Wojciech / Šeļa, Artjoms and Eder, Maciej (2020): "Detecting direct speech in multilingual collection of 19th century novels", in: Rachele, Sprugnoli / Marco, Passarotti (editors), *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages (satellite event to LREC 2020, Marseille, 11–16 May 2020)*, Paris: European Language Resources Association (ELRA), pp. 100–104, [online] https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/LT4HALAbook.pdf [accessed on 21.09.2025].

Cai, Zhenguang G. / Duan, Xufeng / Haslett, David A. / Wang, Shuqi and J. Pickering, Martin (2023): "Do large language models resemble humans in language use?", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2303.08014.

Carrassi, Vito (2016): "A Broader and Deeper Idea of Fairy Tale: Reassessing Concept, Meaning, and Function of the Most Debated Genre in Folk Narrative Research, in: *Folklore: Electronic Journal of Folklore*, vol. 65, pp. 69 - 88, DOI: http://dx.doi.org/10.7592/FEJF2016.65.carrassi.

Chan, Branden / Schweter, Stefan and Möller, Timo (2020): "German's Next Language Model", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2010.10906.

Covington, Michael A. and McFall, Joe D. (2010): "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)", in: *Journal of Quantitative*

Linguistics, vol. 17, iss. 2, pp. 94–100, DOI: https://doi.org/10.1080/09296171003643098.

Cuéllar, Álvaro (2024): „Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays“, in: **Hesselbach, Robert / Calvo Tello, José / Henny-Krahmer, Ulrike / Schöch, Christof / Schlör, Daniel** (Hrsg.), *Digital Stylistics in Romance Studies and Beyond*, Heidelberg: Heidelberg University Publishing, S. 101–117. https://doi.org/10.17885/heiup.1157.c19368

Declerck, Thierry / Scheidel, Antonia and Lendvai, Piroska (2011): "Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales", in: Caroline, Sporleder / Antal, Bosch and Kalliopi, Zervanou (editors), Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing, Berlin, Heidelberg: Springer, DOI: https://doi.org/10.1007/978-3-642-20227-8_9.

Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton and Toutanova, Kristina (2018): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv.org*, DIO: https://doi.org/10.48550/arXiv.1810.04805.

Eder, Maciej / Rybicki, Jan and Kestemont, Mike (2016): "Stylometry with R: A Package for Computational Text Analysis", in: *The R Journal*, vol.8, no. 1, pp. 107–121. DOI: https://doi.org/10.32614/RJ-2016-007.

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Reger, Isabella / Pielström, Steffen / Schöch, Christof and Vitt, Thorsten (2017): "Understanding and explaining Delta measures for authorship attribution", in: *Digital Scholarship in the Humanities*, vol. 31, iss. 2, pp. ii4–ii16. DOI: https://doi.org/10.1093/llc/fqx023.

Fedin, Kai (2025): "Sprachindizes berechnen mit dem Lesbarkeitsindex Rechner", *Fair Text*, [online] https://fair-text.com/lesbarkeitsindex-textanalyse-tool/#:~:text=Der%20Lesbarkeitsindex%20LIX%20(im%20Original,Messung%20der%20Lesbarkeit%20von%20Texten [accessed on 19.09.2025].

Fedoriv, Yaroslava / Shuhai, Alla and Pirozhenko, Iryna (2023): "Linguo-Cognitive Markers in Human vs AI Text Attribution: A Case Study of Narrative and Descriptive Discourse", in: *Актуальні питання гуманітарних наук* [*Current Issues in Humanities*], vol. 3, pp. 130–145. DOI: 10.24919/2308-4863/66-3-20.

Fehle, Jakob / Schmidt, Thomas and Wolff, Christian (2021): "Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques", in: Kilian, Evang / Laura, Kallmeyer / Rainer Osswald / Jakub, Waszczuk and Torsten Zesch (editors), *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, Düsseldorf: KONVENS 2021 Organizers, pp. 86–103. [online] https://aclanthology.org/2021.konvens-1.8/ [accessed on 21.09.2025].

Fisseni, Bernhard / Kurji, Aadil and Löwe, Benedikt (2014): "Annotating with Propp's Morphology of the Folktale: Reproducibility and Trainability", in: *Literary and Linguistic Computing*, vol. 29, no. 4, pp. 488–510, DOI: https://doi.org/10.1093/llc/fqu050.

Flesch, Rudolf (1948): "A New Readability Yardstick", in: *Journal of Applied Psychology*, vol. 32, iss. 3, pp. 221–233, DOI: https://psycnet.apa.org/doi/10.1037/h0057532.

Freitag, Elisabeth (1929): *Die Kinder- u. Hausmärchen der Brüder Grimm im ersten Stadium ihrer stilgeschichtlichen Entwicklung*, Vergleich der Urform (Oelenberger Handschrift) mit dem Erstdruck (1. Band) von 1812. im Rheingau: Adam Etienne, Buch- und Steindruckerei.

Geldern-Egmond, Irene (2000): *Märchen und Behinderung. Ein Beitrag zur Resilienzforschung bei Kindern und Jugendlichen mit Lernbehinderungen*, Baltmannsweiler: Schneider-Verl. Hohengehren.

Geister, Oliver (2013): *Kleine Pädagogik des Märchens. Begriff – Geschichte – Ideen für Erziehung und Unterricht.* 3., überarb. u. erw. Aufl. Baltmannsweiler: Schneider-Verlag Hohengehren. (Mit Beiträgen von Christian Peitz.)

Georgiou, Georgios P. (2024): „Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool“, *arXiv*, arXiv:2407.03646 [cs.CL], Version v2 (11.07.2024). DOI: 10.48550/arXiv.2407.03646. [Zugriff am 30.09.2025].

Google (2025): „Gemini – Google's AI assistant“, *Gemini*, [online] https://gemini.google.com/ [accessed on 22.09.2025].

Grimm, Jacob and Grimm, Wilhelm (1900): *Grimm's Complete Fairy Tales*, Garden City, N.Y.: International Collectors Library, [online] https://dn790008.ca.archive.org/0/items/grimmscompletefa00grim/grimmscompletefa00grim.pdf [accessed on 17.09.2025].

Grootendorst, Maarten (2022): "BERTopic: Neural topic modeling with a class-based TF-IDF procedure", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2203.05794.

Gururangan, Suchin / Marasović, Ana / Swayamdipta, Swabha / Lo, Kyle / Beltagy, Iz / Downey, Doug / Smith, Noah A. (2020): "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2004.10964.

Harvard Library (2025): "Tale-Type and Motif Indices – Library Research Guide for Folklore and Mythology", *Harvard Library Research Guides*, [online] https://guides.library.harvard.edu/folk_and_myth/indices [accessed on 17.09.2025].

He, Pengcheng / Liu, Xiaodong / Gao, Jianfeng and Chen, Weizhu (2021): "DeBERTa: Decoding-enhanced BERT with Disentangled Attention", in: *International Conference on Learning Representations (ICLR 2021)*, [online] https://openreview.net/pdf?id=XPZIaotutsD [accessed on 19.09.2025].

Helms-Park, Rena and Stapleton, Paul (2003): "Questioning the importance of individualized voice in undergraduate L2 argumentative writing: An empirical study with pedagogical implications", in: *Journal of Second Language Writing*, vol. 12, iss. 3, DOI: https://doi.org/10.1016/j.jslw.2003.08.001.

Herbold, Steffen / Hautli-Janisz, Annette / Heuer, Ute / Kikteva, Zlata and Trautsch, Alexander (2023): "A large-scale comparison of human-written versus ChatGPT-generated essays", in: *Scientific Reports*, vol. 13, article no. 18617, DOI: https://doi.org/10.1038/s41598-023-45644-9.

Horstmann, Jan (2024): "Methodeneintrag: Stilometrie", in: *forTEXT*, vol. 1, iss. 1, DOI: https://doi.org/10.48694/fortext.3769.

Jannidis, Fotis / Konle, Leonard / Zehe, Albin / Hotho, Andreas and Krug, Markus (2018): "Analysing Direct Speech in German Novels", in: Georg Vogeler (editor), *Kritik der digitalen Vernunft. Konferenzabstracts: Universität zu Köln, 26. Februar – 2. März 2018* (DHd 2018 Book of Abstracts), Köln: Universitäts- und Stadtbibliothek Köln, pp. 114–118. DOI: 10.5281/zenodo.4622454.

Jannidis, Fotis and Lauer, Gerhard (2014): "Burrows's Delta and Its Use in German Literary History", in: Matt, Erlin and Lynne, Tatlock (editors), *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Rochester, NY: Camden House (Boydell & Brewer), pp. 27–54. DOI: 10.1515/9781571138903-003.

Javaid, Mohd; Haleem, Abid; Singh, Ravi Pratap; Khan, Shahbaz; Haleem Khan, Ibrahim (2023): „Unlocking the opportunities through ChatGPT tool towards ameliorating the education system", *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, **3**(2), 100115. https://doi.org/10.1016/j.tbench.2023.100115

Jurčáková, Edita (2006): "Ludwig Bechstein und seine Märchensammlungen", in: *Studia Germanistica*, no. 1/2006, pp. 99–108, [online] https://dokumenty.osu.cz/ff/journals/studiagermanistica/2006-1/SG_1_10_Jurcakova.pdf [accessed on 16.09.2025].

Katsios, Gregorios A. (2024): *Folktale Story Generation and Automatic Evaluation of Generated Text*, PhD thesis, Rensselaer Polytechnic Institute, Troy (NY), in: ProQuest Dissertations & Theses Global. [online]

https://www.proquest.com/open-view/a1dedb96e3b751220eba4d87fdf3c0da/1?cbl=18750&diss=y&pq-origsite=gscholar [accessed on 20.09.2025].

Kohnke, Lucas / Moorhouse, Benjamin Luke and Zou, Di (2023): "ChatGPT for language teaching and learning", in: *RELC Journal*, vol. 54, iss. 2, pp. 537 – 550, DOI: https://doi.org/10.1177/00336882231162868.

Kumarage, Tharindu / Garland, Joshua / Bhattacharjee, Amrita / Trapeznikov, Kirill / Ruston, Scott and Liu, Huan (2023): "Stylometric detection of AI-Generated text in Twitter timelines", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2303.03697.

Lagutina, Ksenia; Lagutina, Nadezhda; Boychuk, Elena; Vorontsova, Inna; Shliakhtina, Elena; Belyaeva, Olga; Paramonov, Ilya (2019): „A Survey on Stylometric Text Features". In: *Proceedings of the 25th Conference of Open Innovations Association FRUCT (FRUCT'25)*, Helsinki, 5.–8. November 2019, S. 184–195. https://doi.org/10.23919/FRUCT48121.2019.8981504

Lendvai, Piroska / Declerck, Thierry / Darányi, Sándor and Malec, Scott (2010): "Propp Revisited: Integration of Linguistic Markup into Structured Content Descriptors of Tales", in: *Digital Humanities 2010: Conference Abstracts*, London: Centre for Computing in the Humanities, King's College London, [online] https://www.dfki.de/fileadmin/user_upload/import/4720_Lendvaietal_DH2010_final%5B1%5D.pdf [accessed on 20.09.2025].

Lewis, Mike / Liu, Yinhan / Goyal, Naman / Ghazvininejad, Marjan / Mohamed, Abdelrahman / Levy, Omer / Stoyanov, Ves and Zettlemoyer, Luke (2019): "BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.1910.13461.

Liu, Yinhan / Ott, Myle / Goyal, Naman / Du, Jingfei / Joshi, Mandar / Chen, Danqi / Levy, Omer / Lewis, Mike / Zettlemoyer, Luke and Stoyanov, Veselin

(2019): "ROBERTA: A robustly optimized BERT pretraining approach", *arXiv.org*, DIO: https://doi.org/10.48550/arXiv.1907.11692.

Lüthi, Max (1986), *The European Folktale: Form and Nature*, Translated by Niles, John D., Bloomington: Indiana University Press

Makridis, Georgios / Oikonomou, Athanasios and Koukos, Vasileios (2024): "FairyLandAI: Personalized Fairy Tales utilizing ChatGPT and DALLE-3", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2407.09467.

Malec, Scott A. (2001): "Proppian Structural Analysis and XML Modeling", in: *Proceedings of Computers, Literature and Philology (CLiP 2001)*, Gerhard-Mercator-Universität Duisburg, pp. 161-175, [online] https://www.researchgate.net/publication/247286265_Proppian_Structural_Analysis_and_XML_Modeling [accessed on 18.09.2025].

Melliti, Mimoun (2024): "Using genre analysis to detect AI-Generated academic texts", in: *Diá-logos*, vol. 16, no. 29, pp. 09 – 27, DOI: 10.61604/dl.v16i29.377.

Mikros, George (2025): "Beyond the surface: stylometric analysis of GPT-4o's capacity for literary style imitation", in: *Digital Scholarship in the Humanities*, vol. 40, iss. 2, pp. 587–600. DOI: https://doi.org/10.1093/llc/fqaf035.

OpenAI (2025): "GPT-5 System Card", *OpenAI*, [online] https://cdn.openai.com/gpt-5-system-card.pdf [accessed on 22.09.2025].

Perkuhn, Rainer / Keibel, Holger and Kupietz, Marc (2012): *Korpuslinguistik*. Paderborn: Wilhelm Fink (UTB 3433 / LIBAC – Linguistik für Bachelor). DOI: 10.36198/9783838534336.

Petsch, Robert (1900): *Formelhafte Schlüsse im Volksmärchen*. Berlin: Weidmann.

Propp, Vladimir (1968 [1928]): *Morphology of the Folktale. Second Edition.* Übers. Laurence Scott; revidiert und hrsg. von Louis A. Wagner; neue Einleitung von Alan Dundes. Austin: University of Texas Press. https://doi.org/10.7560/783911

Radford, Alec / Narasimhan, Karthik / Salimans, Tim and Sutskever, Ilya (2018): "Improving Language Understanding by Generative Pre-Training",

OpenAI, [online] https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed on 22.09.2025].

Reinert, Nicky (2020): „Einführung in Stemming und Lemmatisierung deutscher Texte mit Python“, *nickyreinert.de* (Blog), [online]

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): „SentiWS – A Publicly Available German-language Resource for Sentiment Analysis“. In: Calzolari, Nicoletta u. a. (Hrsg.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta (Malta): ELRA, S. 1168–1171. [Online] https://aclanthology.org/L10-1339/https://nickyreinert.de/2020/2020-12-09-einfuehrung-in-stemming-und-lemmatisierung-deutscher-texte-mit-python/ [accessed on 21.09.2025].

Richards, Brian (1987): "Type/Token Ratios: what do they really tell us?", in: *Journal of Child Language*, vol. 14, iss. 2, pp. 201–209. DOI: 10.1017/S0305000900012885.

Rölleke, Heinz (1975): *Die älteste Märchensammlung der Brüder Grimm. Synopse der handschriftlichen Urfassung von 1810 und der Erstdrucke von 1812.* Cologny-Genève: Fondation Martin Bodmer.

Rölleke, Heinz (2024): *Die Märchen der Brüder Grimm. Eine Einführung.* 8., überarb. Aufl. Ditzingen: Reclam (Reclams Universal-Bibliothek 17650). ISBN 978-3-15-017650-4.

Sandler, Morgan / Choung, Hyesun / Ross, Arun and David, Prabu (2024): "A Linguistic Comparison between Human and ChatGPT-Generated Conversations", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2401.16587.

Shidiq, Muhammad (2023): „The Use of Artificial Intelligence-Based Chat-GPT and Its Challenges for the World of Education; From the Viewpoint of the Development of Creative Writing Skills“, in: *Proceeding of International Conference on Education, Society and Humanity (ICESH)*, **1**(1), S. 353–357.

[Online] https://ejournal.unuja.ac.id/index.php/icesh/article/view/5614/2065

Scheidel, Antonia and Declerck, Thierry (2010): "APftML – Augmented Proppian fairy tale Markup Language", in: Sándor, Darányi and Piroska Lendvai (editors), *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, Vienna: AMICUS Workshop [online] https://www.researchgate.net/publication/228495435_APftML-Augmented_Proppian_fairy_tale_Markup_Language [accessed on 18.09.2025].

Smith, Peter W. H. and Aldridge, William (2011): "Improving Authorship Attribution: Optimizing Burrows' Delta Method", in: *Journal of Quantitative Linguistics*, vol.18, iss. 1, pp. 63–88. DOI: https://doi.org/10.1080/09296174.2011.533591.

Stamatatos, Efstathios (2008): "A survey of modern authorship attribution methods", *Journal of the American Society for Information Science and Technology*, 60(3), 538–556. DOI: https://doi.org/10.1002/asi.21001

Touvron, Hugo / Lavril, Thibaut / Izacard, Gautier / Martinet, Xavier / Lachaux, Marie-Anne / Lacroix, Timothée / Rozière, Baptiste / Goyal, Naman / Hambro, Eric / Azhar, Faisal / Rodriguez, Aurélien / Joulin, Armand / Grave, Edouard and Lample, Guillaume (2023): "LLaMA: Open and Efficient Foundation Language Models", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2302.13971.

Uther, Hans-Jörg (2011): *The Types of International Folktales – A Classification and Bibliography. Part I: Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction*. Helsinki: The Kalevala Society Foundation (Folklore Fellows' Communications 284, 2nd printing).

Uther, Hans-Jörg (2015): *Deutscher Märchenkatalog. Ein Typenverzeichnis*. Münster; New York: Waxmann.

Vanetik, Natalia / Tiamanova, Margarita / Kogan, Genady and Litvak, Marina (2024): "Genre Classification of Books in Russian with Stylometric Features: A Case Study", in: *Information*, vol. 15, iss. 6, artical 340, DOI: https://doi.org/10.3390/info15060340.

Werzinsky, Jacob / Zhong, Zhiyan / Zou, Xuedan (2022): „Analyzing Folktales of Different Regions Using Topic Modeling and Clustering", *arXiv* (cs.CL), arXiv:2206.04221, 5 S., 2 Abb. https://doi.org/10.48550/arXiv.2206.04221

Wojatzki, Michael / Ruppert, Eugen / Holschneider, Sarah / Zesch, Torsten and Biemann, Chris (2017): "GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback", in: *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, Berlin, pp. 1–12. DOI: https://doi.org/10.17185/duepublico/72074.

Wührl, Paul-Wolfgang (2003): *Das deutsche Kunstmärchen: Geschichte, Botschaft und Erzählstrukturen*, Baltmannsweiler: Schneider Verlag Hohengehren.

Yarlott, W. Victor H. and Finlayson, Mark A. (2016): "ProppML: A Complete Annotation Scheme for Proppian Morphologies", in: Ben, Miller / Antonio, Lieto / Rémi, Ronfard / Stephen G., Ware and Mark A., Finlayson (editors), *Proceedings of the 7th Workshop on Computational Models of Narrative (CMN 2016)*, Dagstuhl: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, pp. 8:1–8:19, DOI: https://doi.org/10.4230/OASIcs.CMN.2016.8.

Zeng, Aohan / Liu, Xiao / Du, Zhengxiao / Wang, Zihan / Lai, Hanyu / Ding, Ming / Yang, Zhuoyi / Xu, Yifan / Zheng, Wendi / Xia, Xiao / Lam Tam, Weng / Ma, Zixuan / Xue, Yufei / Zhai, Jidong / Chen, Wenguang / Zhang, Peng / Dong, Yuxiao / Tang, Jie (2022): "GLM-130B: an open bilingual pre-trained model", *arXiv.org*, DIO: https://doi.org/10.48550/arXiv.2210.02414.

Zhou, Hongjian / Liu, Fenglin / Gu, Boyang / Zou, Xinyu / Huang, Jinfa / Wu, Jinge / Li, Yiru / Chen, Sam S. / Zhou, Peilin / Liu, Junling / Hua, Yining / Mao, Chengfeng / You, Chenyu / Wu, Xian / Zheng, Yefeng / Clifton, Lei / Li, Zheng / Luo, Jiebo and Clifton, David A. (2023): "A Survey of Large

Language Models in Medicine: Progress, Application, and Challenge", *arXiv.org*, DOI: https://doi.org/10.48550/arXiv.2311.05112.

Zielinski, Andrea / Spolwind, Calvin / Grimm, Anna and Kroll, Henning (2023): "A Dataset for Explainable Sentiment Analysis in the German Automotive Industry", in: Jeremy, Barnes / Orphée, De Clercq and Roman Klinger (editors), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (WASSA 2023, ACL'23)*, Toronto: Association for Computational Linguistics, pp. 138–148. DOI: 10.18653/v1/2023.wassa-1.13.

Zitzlsperger, Helga (1984): *Kinder spielen Märchen. Schöpferisches Ausgestalten und Nacherleben*, Weihnheim: Beltz.

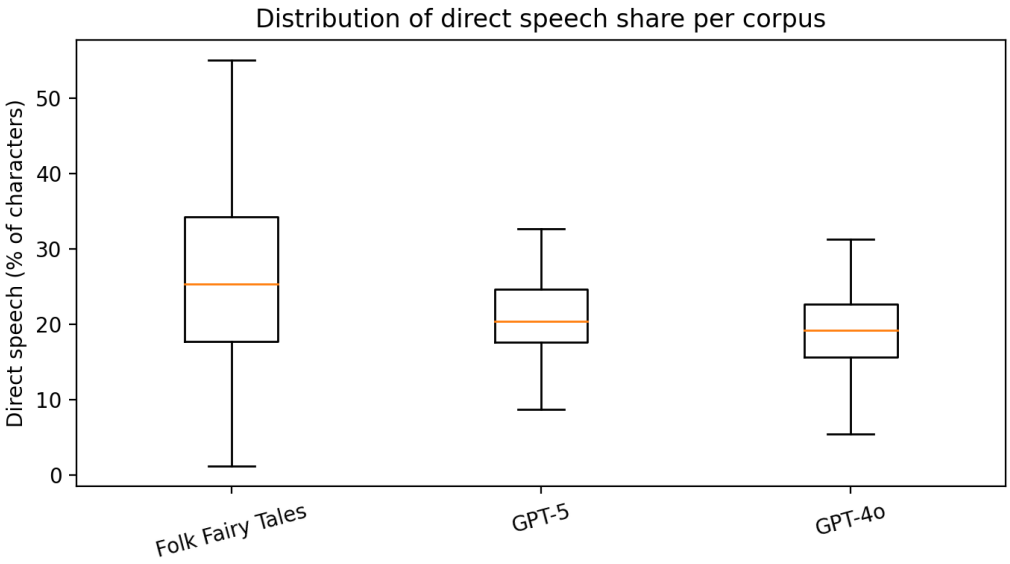
10 List of Figures

11 List of Tables

12 Appendix

1. Github Repository:
   https://github.com/SophiaHeigl/Masterarbeit-Fairy-tales-vs.-AI.git

2. Figure: Destribution of direct speech share per corpus



3. German Prompt:

```
Du bist ein Emulator, der klassische Volksmärchen
schreibt.
Deine Aufgabe ist es, neue Märchen zu erzeugen.
Hier ist ein Beispieltext aus einem klassischen deut-
schen Volksmärchen:

"Es waren einmal zwei Brüder, ein reicher und ein armer.
Der reiche war ein Goldschmied und bös von Herzen; der
arme nährte sich davon, daß er Besen band und war gut
und redlich. Der arme hatte zwei Kinder, das waren Zwil-
lingsbrüder und sich so ähnlich wie ein Tropfen Wasser
dem anderen. Die zwei Knaben gingen in des Reichen Haus
ab und zu und erhielten von dem Abfall manchmal etwas zu
essen. Es trug sich zu, daß der arme Mann, als er in den
Wald ging Reisig zu holen, einen Vogel sah, der ganz
golden war und so schön, wie ihm noch niemals einer vor
Augen gekommen war. Da hob er ein Steinchen auf, warf
nach ihm und traf ihn auch glücklich; es fiel aber nur
eine goldene Feder herab und der Vogel flog fort. Der
Mann nahm die Feder und brachte sie seinem Bruder, der
sah sie an und sprach"
```

Keine Überschrift oder Kapitelüberschriften. 900 bis 1900 Wörter.
Schreibe ein neues klassisches Volksmärchen.

Plagiatserklärung

Hiermit erkläre ich, dass die vorgelegten Druckexemplare und die vorgelegte digitale Version der Arbeit identisch sind.
Ich habe die Arbeit selbständig verfasst, keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die Arbeit nicht bereits an einer anderen Hochschule zur Erlangung eines akademischen Grades eingereicht.
Ich bestätige, dass ich von den in §26 Abs 6. der Prüfungsordnung vorgesehene Rechtsfolgen Kenntnis genommen habe.

Transparenzhinweis zum Einsatz digitaler Werkzeuge:
Zur Übersetzung, zur Programmier-/Coding-Unterstützung und zum sprachlichen Umformulieren habe ich das KI-Werkzeug Chat GPT unterstützend eingesetzt. Die Verantwortung für Inhalt, Methodik und Auswertung liegt vollständig bei mir, sämtliche verwendeten Quellen sind im Text bzw. im Literaturverzeichnis nachgewiesen.

Germering, den 29.09.2025