Exploring West African Folk Narrative Texts using Machine Learning

Author:

Gossa Lô (2523988)
a.g.lo@vu.nl

Supervisor:

Dr. Victor de Boer
v.de.boer@vu.nl

Abstract. West African and Western European folk tales differ in style and structure, partly due to differences in historical literary traditions. As part of the Digital Humanities agenda, recent advances in Machine learning (ML) and Natural Language Processing (NLP) lead to new ways of investigation into textural heritage. This thesis examines how ML and NLP can be used to identify, analyze and generate West African folk tales. Two corpora of West African and Western European folk tales are compiled and used in three experiments. Each of these experiments investigates the applicability of ML and NLP on cross-cultural folk tale analysis.

In the text generation experiment, two types of deep learning RNN text generators are built and trained on the West African corpus. We show that although the texts range in semantic and syntactic coherence, each of them contains West African features. A survey conducted among 14 participants demonstrates that humans are well able to distinguish the generated texts trained on the West African corpus from those trained on the Western European corpus by focusing on the occurrence of context-specific characters and objects.

The text classification experiment further examines the distinction between the West African and Western European folk tales by comparing the performance of an LSTM (acc. 0.74) with a BoW classifier (acc. 0.93). Since the BoW model does not consider word order, its high accuracy confirms earlier results that the two corpora can be clearly distinguished in terms of vocabulary. An interactive T-SNE visualization of a hybrid classifier (acc. 0.85) highlights the culture-specific words for both.

The third experiment describes interviews with Ghanaian storytelling experts and an ML analysis of narrative structures that builds on this. Classifiers trained on parts of folk tales according to the three-act structure are quite capable of distinguishing these parts (acc. 0.78). Common n-grams extracted from these parts not only underline cross-cultural distinctions in narrative structures, but also show the overlap between verbal and written West African narratives.

The experiments in this thesis demonstrate the contribution of various ML and NLP techniques to the cross-cultural exploration of West African folk tales, from the extraction of culture-specific and informative features to folk tale generation and classification.

Keywords: Storytelling · Deep learning · NLP · Text generation · Text classification · West Africa · Folk tales · ICT4D

Table of Contents

1 Introduction

Storytelling is a powerful communicative interaction, in which narratives are shaped and shared with the aim to captivate and involve the audience [44]. A narrative is a perspective of a story, represented as an event or a sequence of events and actions, carried out by characters [1].

Storytelling has been a particularly popular type of communication in African cultures. It is used to orally pass down and retain information, knowledge and traditions over generations. Through this oral tradition, personal experiences and emotions, and in a broader sense human behaviors, cultural beliefs and values are taught and mirrored [16]. African storytelling is an interactive oral performance, in which the audience participates actively while a respected community member narrates a story [58]. The ending encapsulates a moral lesson about everyday life which is also reflected in the deeper narrative structure [61].

In more recent years, the verbal tradition transformed when African authors started to write down their folk tales. These written folk tales were shaped by the long oral tradition of storytelling, or as Iyasare put it more eloquently:

> "The modern African writer is to his indigenous oral tradition as a snail is to its shell. Even in a foreign habitat, a snail never leaves its shell behind" [36].

Although Europe, like Africa, has a history of oral storytelling, this has changed to what we call a "written culture", meaning that its historical narrative is recorded in printed documentations. Aesop and Aristotle are considered among the first storytellers and fabulists. In the 19th century the German brothers Grimm wrote what are now considered the most famous fairy tales of the world [27]. The differences in verbal and written traditions between West Africa and Western Europe have influenced narrative style, themes and meaning [20].

Throughout the years, many researchers have come up with narrative theories to study use of narratives in literature, films, and video games. The Soviet folklorist Propp was among the first to study narrative structures of folk tales by analyzing Russian fairy tales and identifying common narrative elements and themes [54]. Although fairy tales and myths are considered similar across cultures, narratives are believed to be culture-specific [25].

With the emergence of Natural Language Processing (NLP) techniques and Deep Learning (DL) in the field of Computational Linguistics, more advanced artificially intelligent algorithms can be applied to analyze human language. Ever more NLP tasks are being completed successfully, which is why the question arose whether it is possible to use these recent advances in an exploratory research analyzing West African and Western European folk tales. This master thesis therefore focuses on uncovering cultural differences between West African and Western European folk tales by means of Machine learning (ML) and NLP. The analysis is done by collecting folk tales from both continents to compile two corpora: a West African and a Western European one. The main interest is to examine whether culture-specific elements will emerge when we do this automatic analysis.

1.1 Problem statement and motivation

This master thesis is conducted both from a Digital Humanities and an ICT for Development (ICT4D) perspective. Digital Humanities is defined as the interdisciplinary field between computationally or digitally engaged research, and the humanities. In this field, the role of humanities is being modernized by researching human cultures using computational approaches. ICT4D research aims to bridge the digital divide that is apparent in rural and remote areas in the Global South [2]. The digital divide is the gap or uneven distribution between regions that have access to information and communication technologies (ICTs), and those that do not. This includes access to personal computers, television, mobile phones and the internet.

The aim of this thesis is to conduct preliminary research into differences in narratives and communication between West Africa and Western Europe. The project furthermore examines the effect of oral storytelling on contemporary written literature in West Africa. Folk tales reflect culture-specific knowledge, beliefs and motivations. Examining them is trivial in understanding human actions and cognition. Identifying and understanding the cultural information and knowledge underlying narratives helps us in shaping and adjusting messages to properly address individuals or groups from different cultures.

From an Artificial Intelligence point of view, we study the usefulness of automatic ML and NLP approaches over manual and more traditional ones in analyzing and extracting patterns and culture-specific features from folk tales. Some work has been done in computationally studying folk tales, but no research was found to date that specifically focuses on cultural differences between West Africa and Western Europe. With the emergence of more complex mathematical models and more robust machines, ML and DL have become ever more capable of identifying patterns in huge amounts of data. Recent trends that focus on integrating DL and NLP have allowed humans to make human language related tasks such as machine translation and speech recognition more effective. By making use of these developments, this thesis applies existing ML algorithms to investigate narratives from different angles. In addition, a field trip to Ghana is conducted to study the historical and modern use of storytelling and its effect on written narratives.

1.2 Outline

The main contribution of this thesis is the compilation of two corpora. One corpus contains West African folk tales, and the other contains Western European folk tales. These corpora are used throughout the entire thesis, which is divided into three experiments.

In the following section (i.e. section 2), the theoretical background is described. This section explains several narrative theories underlying the experiments. Section 3 elaborates on our approach with regard to the experiments, and introduces the research question. In section 4, the construction of the corpora is explained and the data is explored. The three consecutive sections each describe a different ML approach applied to examine the research question. Section 5 elaborates on the first experiment: Text Generation. In the sixth section, the Text Classification experiment is described. The third and final experiment, Narrative Structure Analysis, is explained in section 7.

Each experiment has its own introduction, related work section, and discussion. The reason for repeating these subsections is that each experiment tackles a different sub-problem using various ML and DL techniques. An extensive literature study has been done, partly described in the "theoretical background" section, and partly in the related work section of each experiment. Where the former describes the theoretical framework that underlies decisions made in the project, the latter describe technical studies that complement or are similar to the experiment in question.

Finally, sections 8 and 9 discuss and conclude the entire thesis. These sections reflect on the contribution of each experiment in answering the main research question and describe future work.

The corpora and code created for this project can be found on our GitHub repository[1]. In the "code" folder of the repository, the Python scripts associated with the three experiments are stored. The neural network models can be found in the "model" folder, and the corpora in the "data" folder. The visualization created for section 6.4 can be found on our web page[2].

2 Theoretical background

In this section, several African and European narrative theories are described. These are directly or indirectly relevant to the experiments conducted in the following sections. The theoretical framework is particularly relevant for section 7, which computationally analyzes the narrative structure of the folk tales. The other two experiments (sections 5 and 6) also benefit from the elaborate literature study, as having a deeper knowledge of culture-specific narrative structures facilitates the tracing of elements that are part of this structure in folk tales.

Narratology is the study of narrative structure, which focuses on researching similarities, differences and generalizability of narratives. A narrative structure is the structural framework that describes the story as a sequence of events, in a specific setting and plot that is appealing to its listeners or readers [9].

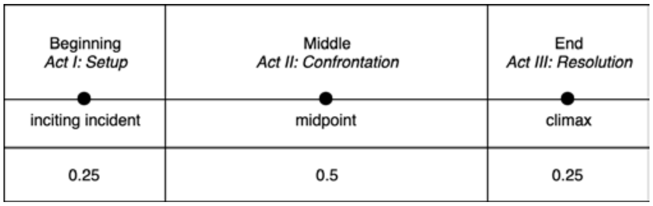
Various theories exist about narrative structures and their cultural sensitivity, some of which focus specifically on folk tales. Inderjeet describes *Computational Narratology* as the study that examines algorithmic processes required in creating, interpreting and modelling narratives and narrative structures from the point of view of computational information processing [47]. In this thesis, algorithmic Computational Narratology processes, such as the automatic generation of stories (section 5) and extraction of narrative structures (section 7), are used to create and interpret folk tales.

[1] https://GitHub.com/GossaLo/afr-neural-folktales

[2] https://gossalo.github.io/tsne-visual/

A popular concept in narrative structure is the three-act structure, which some claim originated in Aristotle's "Poetics" [43]. In this work, Aristotle remarked that a tragedy is only whole when it has a beginning, a middle, and an end. Yet others claim it was the screenwriter Syd Field who first introduced the three-act structure in film and with it a formula for successful films [7].

Figure 1 illustrates how the three-act structure is organized. It has a beginning (i.e. Setup), a body (i.e. Confrontation) and an end (i.e. Resolution), in the ratio of 1:2:1. In the first act, the main characters and their relationships are introduced, as well as the world in which they live. In the middle of this act, an inciting incident or *catalyst* takes place, which is the first turning point in the story igniting the protagonist to engage in an action. In the second act, the protagonist tries to solve the problem, which instead worsens. The worst moment of the story is during the midpoint, in the middle of the second act. In the third act, the problems are solved. The peak of the story is during the climax, when the story is most intense en the remaining questions are answered. Finally, the protagonist solves the issue and goes back to his old life with newly acquired knowledge or skills [7].



Beginning *Act I: Setup*	Middle *Act II: Confrontation*	End *Act III: Resolution*
● inciting incident	● midpoint	● climax
0.25	0.5	0.25

Fig. 1. The three-act structure

The well-arranged layout of the three-act structure make it a frequently used structure both in Computational Narratology research and in narratives used in novels, films and video games. This is why in section 7 we use the three-act structure as a basis for our classification and information extraction task.

2.1 European narrative theories

A famous theory on narrative structure that had great impact on storytelling, writing and movie-making came from Joseph Campbell. In his book "The hero with a thousand faces" Campbell identified a common theme or archetypal motif that underlies every story, which he named *the myth of the hero*, or "monomyth". He begins his book stating:

> "Whether we listen with aloof amusement to the dreamlike mumbo jumbo of some red-eyed witch doctor of the Congo, or read with cultivated rapture thin translations from the sonnets of the mistic Lao-tse; now and again crack the hard nutshell of an argument of Aquinas, or

> catch suddenly the shining meaning of a bizarre Eskimo fairy tale: it will always be the one, shape-shifting yet marvelously constant story that we find, together with a challengingly persistent suggestion of more remaining to be experienced than will ever be known or told" [8].

Campbell claims that every story told or written is transcending both in time and culture, and is a variation on a common theme. The result is an infinite set of variations with one basic form, the monomyth. Reliant on Jungian psychology, he believed that this universal theme is due to an unconscious pattern of thought that is part of every human being. In *the hero's myth*, the hero goes on an adventure according to the structure of the monomyth, which consists of three stages: Separation - Initiation - Return. These stages are further divided into 17 sub-stages [8]. An example with Prometheus is given to illustrate how this structure could be applied to a narrative:

1. **Act I: Separation (also departure):** The hero goes on an adventure, e.g. Prometheus ascends to the heavens.
   - The call to adventure;
   - refusal of the call;
   - supernatural aid;
   - crossing the first threshold;
   - belly of the whale.
2. **Act II: Initiation:** The hero wins a victory in the climax of the story, e.g. Prometheus steals fire from the gods.
   - The road of trials;
   - meeting with the goddess/love;
   - woman as temptress;
   - atonement with the Hero's father;
   - peace and fulfillment before the Hero's return;
   - the ultimate boon.
3. **Act III: Return:** The hero comes home and is changed, e.g. Prometheus descends back to earth.
   - Refusal of the return;
   - magic flight;
   - rescue from without;
   - the crossing of the return threshold;
   - master of two worlds;
   - freedom to live.

Campbell's claim that all stories are instances of the monomyth was not widely accepted and received backlash. Dundes, for one, states that Campbell does not distinguish between myths, folk tales and legends. While Campbell premise is myths, his study instead was based on analyzing folk tales and legends. Regarding the universality statement, Dundes argues that Campbell could not empirically proof that a universal myth exists. He further states that the proof provided by Campbell is insufficient. When Campbell mentions a universally common motif "Birth from Virgin", he backs his claim with only three citations. Examples from Africa, Australia and many other places, have not yet been found. Dundes mentions that, contrary to Campbell's point of view, he and many other folklorists believe that any universality or parallelity can be traced back to the *monogenesis and diffusion* theory [14]. This theory is based on the belief that one parent tale has multiple descendants spread over the world and is in contrast with the *polygenesis* theory, which assumes that similarities in tales is due to independent invention in unconnected places. Supporters of the *polygenesis* theory, such as Campbell and Propp, try to refute the *monogenesis and diffusion* theory by pointing at the similarity between specific tales (e.g. the frog queen) found in different parts of the world, and trace its resemblance back to a uniform tendency of the human psyche [59].

A narrative theory from which the structure is derived from Campbell's study was Christopher Vogler's "The Writer's Journey" [10]. Vogler came up with a more contemporary variant of the monomyth, which is more in line with the three-act structure. Most of Campbell's sub-stages illustrated above can be used interchangeably. An exception to this is the "refusal of the return", which should of course come after "refusal of the call". Vogler instead proposes a more linear structure consisting of 12 stages:

1. Ordinary World
2. Call to adventure
3. Refusal of the Call
4. Meeting with the Mentor
5. Crossing the First Threshold
6. Tests, Allies, Enemies
7. Approach to the Inmost Cave
8. Ordeal
9. Reward (Seizing the Sword)
10. The Road Back
11. Resurrection
12. Return with the Elixir

In the beginning, the Hero finds himself in its usual situation, the ordinary world. Then, a problem or challenge occurs, which is why the hero has to leave the ordinary world. After a first hesitation to leave, a mentor will offer advice. During the adventure, the Hero has to solve many sub-problems, and encounters both allies and enemies. All these sub-problems prepare the Hero for the largest and most challenging problem, which takes place in the "inmost cave". At the height of the ordeal, something terrible happens after which all seems lost. However, the Hero survives and receives an award. Finally, the Hero travels back home, encountering yet another problem that is also being overcome. In the end the Hero returns back to its ordinary world. These events are interwoven in the three-act structure and are often used to shape storytelling in games and movies. Its flexibility is suitable for a wide range of stories about adventures requiring problem solving from fantasy to reality [13].

A narrative theory popular for its focus on folk tales is described by folklorist Vladimir Propp in his book "Morphology of the Folk tale". In his study, Propp systematically analyzed 100 Russian folk tales and identified common themes and character functions among them, as well as cultural characteristics. This was done by dividing the stories into morphemes and identifying narrative units, which he called "functions of Dramatis Personae" [54]. Some examples of these functions are:

- An interdiction is addressed to the hero.
- The villain receives information about his victim.
- One member of a family either lacks something or desires to have something.

Propp concluded that there are only 31 identifiable and generic functions in the Russian folk tale. He furthermore argued that the sequence of the functions in these tales is similar and that they can all be fitted into one single structure [23]. Propp's narrative theory has often been used as a basis in computational studies to annotate, analyze and generate folk tales [18][24][34].

Although most narrative theories described in this subsection are not used directly in the project, being aware of their existence and how they vary is relevant when conducting Computational Narratology research. This subsection indicates how divers and complex the different theories are. Automatically identifying let alone extracting them from folk tales is a difficult task. Section 7.2 mentions some studies that used Propp's work in their computational analysis, but contrary to our corpora, most of these were annotated. Our choice to use the three-act structure in section 7 is reinforced by the fact that both Campbell and Vogler confirm its existence and introduce a more complex variant based of this.

2.2 African narrative theories

In contrast to the wide variety of narrative theories that exist in the West, less information can be found on West African narrative structures. Instead, a considerable part of the literature focuses on the oral storytelling tradition, which heavily influenced written literature. African storytellers have the important task to memorize genealogies and events, which they then recite to chiefs, kings and other important figures in an engaging way. Good storytellers have great sense of timing, use suitable voices and are great in creating suspense and interacting with the audience [4]. Typical West African literary elements are said to be its relatively loose narrative structure, idea of time, and a lack of character delineation [36]. Contrary to European literature, African written literature has matured much later. The emergence of written literature (in English) in Anglophone West Africa goes hand in hand with colonization.

As Roland Barthes points out in his book *Le degré zéro de l'écriture*: "le langage n'est jamais innocent" [3], meaning that political, social and cultural beliefs and habits are reflected in the way language is used. Since English is a second language in Anglophone West Africa and their authors have a different mother tongue, the distance between these two is said to have left its mark on the narrative structure and how the works should be interpreted. Because these works have been translated or written in the language of the oppressor, a more Eurocentric point of view is utilized instead of a West African one. This could cause elements typical for West African oral literature, e.g. proverbs, imagery, lyrical language and riddles, to get lost [28]. Furthermore, neglecting these storytelling elements when analyzing these works is to ignore what make these narratives unique, rich and typically African [36].

Ninan et al. argue that in developing cross-cultural computational narrative models, one should be sensitive to culture-specific logic and relations. In their research on the use of narrative structures by the Yorùbá people, they identify and distinguish two human forms: the tangible (physical), and the intangible (spiritual). They furthermore state that, as seems to be the case for more West African narratives, they are focused on the communal instead of the individual [51]. These world-views are quite important to keep in mind when analyzing and creating (computational) narratives.

Edward Sacky has studied a selection of African novels to examine the extent to which written literature can be traced back to traditional oral storytelling [56]. Symbolism is often used in these novels and stems from a traditional African form of communication. To illustrate, one of the novels has 13 chapters, symbolizing the Akan traditional calender. Another book has seven episodes, which is the number often associated with perfection and unity.

The novels Sacky analyzed have an introduction, a body, and a conclusion, among which said chapters are divided. Following the Akan tradition, the introduction and conclusion are connected. Audience participation is very important in African art. In the storytelling setting, the storytellers are surrounded by the audience, and both have obligations throughout the event. The storytellers take turns telling the stories, and the story starts with a specific interactive opening formula between the two parties. The Dangme people in Ghana have a clear role for both storytellers and audience during the introduction of a story. The storyteller begins the story by saying "*i ti ha nye*", which means "I'll tell you a story", to which the audience responds with "*wa he o no*", or "we are all ears". The storyteller then says "*Ligbi ko wa nge*", or "Once upon a time", and the audience again responds with "*Abo dzemi loko wa ba*", "The story was created long before we were born". This exchange is meant to enthuse the audience, and was incorporated in some African novels. At the end of the storytelling event, the storyteller ends the tale using a closing formula, and asks the next storyteller to tell a story. The role of the audience is to interrupt throughout the tale by making remarks, correcting the storyteller, or by singing songs [4].

Folktales, and fables in particular, have often been used in ethnographic studies to examine culture-specific habits and beliefs of specific cultures. Fables lend themselves well to crosscultural comparison, because they exist in every culture and indicate what everyday life looks like and societal norms and behaviors in a simple yet engaging way [57].

The majority of the folk tales analyzed in this project are fables, which are known for their moralistic nature. Animal tales are a specific type of fables and can be defined as short narratives featuring anthropomorphized animals or plants. This means that human instincts and responses are projected on them while most of their animal behaviors have been effaced. The aim of the fable is to both be entertaining and informing at the same time by conveying a moral lesson to the reader. The best documented type of a West African fable is the trickster tale. In this tale, a weak but triumphant, intelligent and witty animal takes on a stronger animal, and defeats it with tricks. The type of trickster animal differs per country, dependent on the fauna found in the region, but the way in which it is used to portray human characteristics is always similar.

One of the most well known West African tricksters is the Ghanaian spider Kwaku Ananse, the folk tales which later spread to the West Indies and the Caribbean. In other West African countries such as Ivory Coast and Senegal the trickster is a hare, and among the Yoruba people in Nigeria it is a tortoise. Despite the trickster being a weaker prey animal, it always succeeds in tricking a stronger animal and overcoming its problems [56][53]. In addition to the tales centered on tricksters, it is typical for African tales to include mystical, supernatural beings, such as Sasabonsam, the forest monster. Another recurring actor is the Hunter, and so are twins, orphans, and children-born-to-die [4].

The studies described in this subsection identify characteristic West African narrative and literary elements, some of which find their origin in oral storytelling. Several of these, such as communal thinking, the storytelling beginning and end, the singing of songs, and the use of animals and mystical beings are also mentioned by the storytelling experts interviewed in section 7. Furthermore, this information enables us to better understand what to look for when we search for West African features in folk tales. This knowledge benefits all three of the experiments (i.e. sections 5, 6 and 7).

3 Research question and Approach

As mentioned in the introduction, West African and Western European literature differ in historical background and both contain different features. NLP has come to play an increasingly important role in ML, in particular to understand the structure of and meaning behind text. Using ML to examine a corpus of West African folk tales and comparing it with a corpus of Western European ones has not yet been done. With this scope in mind, the main research question of this thesis is therefore:

> *How can Machine learning and Natural Language Processing be used to identify, analyze and generate West African folk tales?*

The research question is exploratively examined on the basis of the three experiments mentioned in section 1.2. Figure 2 illustrates the project's process flow from data acquisition to evaluation.

The left block in the figure, named *Data acquisition and preprocessing*, shows the acquisition of the West African and Western European folk tales. These have either been scraped from the Web or were extracted from scanned books available online using Optical Character Recognition (OCR). Using these techniques, a total of 742 English narratives have been collected, 252 of which are West African, and the other 490 Western European. The West African folk tales are written by authors from Anglophone West African countries such as The Gambia, Ghana, and Nigeria. Some of the folk tales are written by West African adults as a way to preserve and read them to their children. The Western European folk tales are written and/or translated by authors from countries such as the Netherlands, Germany, France, and the UK.

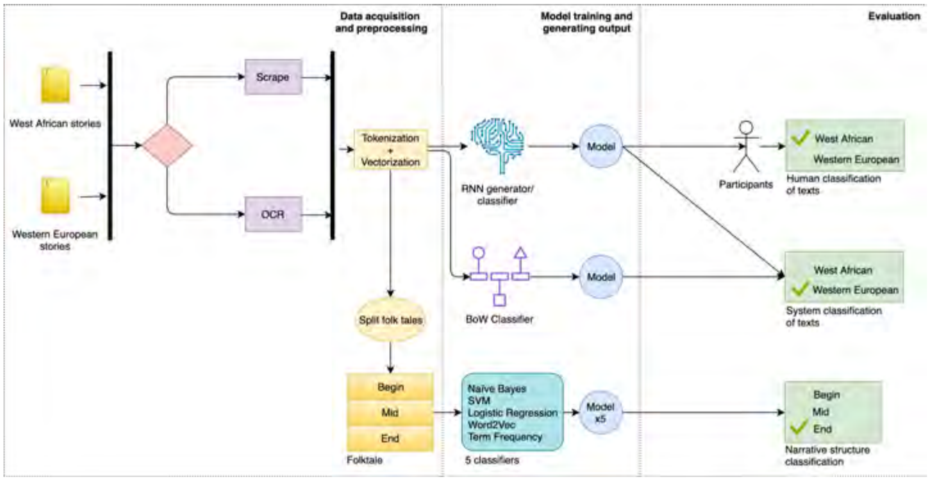


Fig. 2. Thesis project process flow

To prepare the data for training, the folk tales are preprocessed. This is done by cleaning, tokenizing and vectorizing the texts, the specific steps which depend on the model being trained. Keras is used to train the models. This is an open-source neural network library for Python that is used for building DL neural networks. It is built on top of TensorFlow and is intuitive and easy-to-use. Neural network layers are stacked on top of each other, allowing to analyze the individual layers by using several data analysis and visualization packages in Python. Additionally, a ML software library for Python called Scikit-learn is used for the ML models.

Training the models (see figure 2: *Model training and generating output*), depends on the experiment at hand and is explained in the associated sections 5, 6 and 7. To speed up training, the Lisa cluster from SURFsara[3] is used. This system offers a number of multi-core nodes, which can be used by those who need large computing capacities. We train both on CPU (8) and GPU (4) nodes, the latter which significantly sped up the training. The drawback of using Lisa is

[3] https://userinfo.surfsara.nl/systems/lisa

that for each job there is a queue. This means that in some cases, we had to wait for days to train a model. As an alternative, some models were trained on 8 CPU nodes on the Google Cloud Platform[4].

In the first experiment, *Text Generation* in section 5, a story generator is built using two Recurrent Neural Network (RNN) models (see figure 2). The model outputs generated texts based on predictions it learns during training. Part of the evaluation is done by the author, who compared the syntactic and semantic strength of the output texts. The second part of the evaluation is done by conducting a survey with human participants. They were asked to complete a set of tasks concerning the generated texts. More on this evaluation is found in section 5.5.

The second experiment, *Text Classification* in section 6, uses classifiers to categorize the folk tales from the corpora according to their geographical background (i.e. West African or Western European). Both a Bag-of-Words based classifier and a DL RNN classifier were trained (see figure 2). The metric used to measure the performance is the accuracy score. Furthermore, the models were evaluated by computing the predicted origin of eight unseen texts and comparing these with their true origin.

In the final experiment, *Narrative Structure Analysis* in section 7, the narrative structures of both geographical backgrounds are further analyzed by training classifiers to distinguish between parts of folk tales. Additionally, frequently occurring n-grams are extracted from the parts of the folk tales to examine if we can find recurring patterns. As can be seen in figure 2, each folk tale is divided into three parts: begin, mid, and end. The reason behind this is further explained in section 7.4. Multiple ML classifiers are trained, and their models are used to classify the texts. The metric used to measure the performance is the accuracy score.

4 Corpus construction

As mentioned, both a West African and a Western European corpus have been compiled for this project. These corpora contain folk tales collected from various online sources. The corpora can be found in the “data” folder in our GitHub repository[5].

Part of the folk tales come from the collection of folk tales published online by the University of Pittsburgh[6]. An example snippet of one of the Western European folk tales is:

> “Her parents did not think about it for long. ’Birds of a feather, flock together,’ they thought, and gave their consent. So Fat Trina became Heinz’s wife, and drove out both of the goats. Heinz now enjoyed life, having no work to rest from, but his own laziness. He went out with her

[4] https://cloud.google.com/

[5] https://GitHub.com/GossaLo/afr-neural-folktales

[6] https://www.pitt.edu/ dash/folktexts.html

only now and then, saying, 'I'm doing this so that afterwards I will enjoy resting more. Otherwise I shall lose all feeling for it.'"

An example snippet of one of the West African folk tales is the following:

> "Ansa was King of Calabar for fifty years. He had a very faithful cat as a housekeeper, and a rat was his house-boy. The king was an obstinate, headstrong man, but was very fond of the cat, who had been in his store for many years. The rat, who was very poor, fell in love with one of the king's servant girls, but was unable to give her any presents, as he had no money."

In the experiments, we make a distinction between two types of corpora, A and B.

- Type A corpora: 0.5 MB West African, 0.5 MB Western European
- Type B corpora: 1.1 MB West African, 1.1 MB Western European

Both of these types contain a West African and Western European variant. However, the corpora of type A both are approximately half the size of the corpora of type B (i.e. 1.1 MB each). The reason for this difference is that we decided to train and evaluate a text generating RNN model by means of a survey at the start of the project as a baseline (see section 5.5). At this time, the corpora were still relatively small. Later, when more folk tales had been collected, the other experiments were performed on the larger sized corpora of type B. This means that all experiments except for the survey conducted in section 5.5 use the corpora of (2 * 1.1) 2.2 MB of type B.

As can be seen in table 1, the number of folk tales in the West African corpus differs from those in the Western European corpus, while the file sizes and total number of words and characters are almost the same. This is due to the fact that a significant part of the Western European folk tales is shorter than the West African folk tales. This is reflected in the lower average word count of the Western European folk tales i.e. 421 words, compared to the West African ones (i.e. 804 words).

Table 1. Statistics for the corpora of type B

	West African	**Western European**
File size in MB	1.1	1.1
Total no. folk tales	252	490
Total word count	203,537	202,866
Total no. characters	857,590	855,097
Min. word count	53	34
Max. word count	7,878	3,536
Avg. word count	804	421

5 Experiment 1: text generation

5.1 Introduction

At first glance, West African folk tales may look like Western European ones. It is only when the narratives are read, that differences become apparent. Pinpointing the concrete differences between the two is more difficult.

In recent years, the focus within NLP research has shifted from more knowledge-driven, pattern-based approaches to data-driven, or statistics-based ones, with some applying hybrids to understand and analyze human language [32]. Although both approaches have advantages and disadvantages, in this section a data-driven approach is used, because we aim to recognize and extract patterns rather than applying them.

Natural Language Generation (NLG) is an NLP task that uses large amounts of text to compose new texts. The aim of AI researchers is to get NLG-technologies to the point that they generate texts that seem to be written by humans. Current progressions in DL have made the training of text generation neural networks relatively easy. One of the challenges is that even though generated texts may appear correct at first glance, they are often semantically and syntactically incorrect.

In this section, we train Deep Neural Networks using both corpora of type B (see section 4). The main goal is to analyze generated texts to examine whether they contain West African features and what these look like. The generated narratives are evaluated on their level of semantic and syntactic coherence. We investigate whether we can generate narratives with West African features that are semantically and syntactically coherent.

First, the corpus of West African folk tales is fed to two neural networks, which are trained to generate new narratives in West African style. The assumption is that more explicit West African features such as culture-specific protagonists, other characters or objects will appear in the generated texts. Identifying a West African narrative structure, however, is arguably more difficult as this is more implicit and hidden in the tale. Finally, we conduct a qualitative human evaluation to examine whether participants are able to identify culture-specific features and how they would assess the semantic and syntactic coherence of the generated texts.

5.2 Related work

In the early days of Artificial Intelligence, text generation used to be knowledge-driven through rule-based reasoning or by using templates. First attempts to build story generating systems used narrative theories such as those described in the theoretical background (section 2) [24]. However, defining these theories in AI terms is quite difficult, and writing a “good” story is more than establishing a formula and simply following a set of rules alone. The types of (human) intelligence required, think of creativity and intuition, are difficult to mimic in machines [22].

Some of the earlier, classical story generation systems build are TALE-SPIN (1977) and MINSTREL (1984), which both use user-submitted requirements to generate stories. The former requires a large set of logical inferences to set up a plan for the main character to reach his/her goal. MINSTREL, on the other hand, uses a Case-Based Reasoning approach to generate stories set in King Arthur's universe by reusing parts of old stories stored in memory [22][52]. A couple of years later, narrative and storytelling returned in the field of AI, this time to create interactive narratives for virtual video games [62][49].

It is only in more recent years that the focus has shifted and has become a Language Modelling problem, in which a language model is trained to predict the likelihood of the next word given a sequence of words. Neural networks in particular have become frequently used solutions to NTG tasks such as machine translation and dialogue generation [15]. There is, however, a trade-off in using data-driven approaches over knowledge-driven ones. Where the former tend to be more flexible and expressive, knowledge-driven approaches are more controllable and predictable [67]. The fact that in data-driven NTG approaches, the outcome is difficult to control often yields unexpected, creative, but unsolicited outcomes.

NTG is a sequence generation task, which is currently solved best by neural networks that work through sequence-to-sequence (Seq2Seq) learning. In Seq2Seq learning, models that contain an encoder-decoder architecture are trained to convert sentences from one domain to another, as in the case of machine translation. The reason why this is a suitable option for text generation is that input and output sequences are not required to have a similar, fixed size [17]. Since in human written texts each sentence has a different amount of words, and previous words should be considered when predicting the next, this requires a more advanced setup than a fixed-sized network.

The most frequently used type of Seq2Seq neural networks for NTG are Recurrent Neural Networks (RNN). RNNs are dynamic models that are able to generate sequences in multiple domains such as machine translation, image/video captioning, and music [5][37]. In 2015, Karpathy published a blog post on how to build character-level RNNs for text generation[7]. He furthermore released his version of a character-level RNN model named Char-rnn on GitHub. Corpora used to train this model are the complete works of Shakespeare (4.6MB) and Tolstoy's "War and Peace" (3.3 MB)[8]. His post became a huge success and encouraged others to try and follow his example. Another easy-to-use character-level text generator that came from this is TextGenRNN[9], published by Woolf in 2017. This module allows the user to input some text, tune the network parameters and train the RNN to output text in a matter of minutes. Another well known example is that of an RNN model trained on the first four books of

[7] http://karpathy.GitHub.io/2015/05/21/rnn-effectiveness/

[8] https://cs.stanford.edu/people/karpathy/char-rnn/

[9] https://GitHub.com/minimaxir/textgenrnn

the Harry Potter book series (2.8 MB)[10]. This resulted in a new Harry Potter chapter, which is semantically but especially syntactically well written.

Although RNNs are capable of generating new, complex sequences and maintaining short-term information in memory, on their own they are incapable of storing long-term past inputs for too long, because of the so-called vanishing gradient problem. In RNNs, information travels through time, meaning that information obtained in previous time points is used for the next time points. For each of these time points the error, or cost function, is calculated. The vanishing gradient problem arises when the cost function has to be propagated back to update the weights. Because of the multiplications involved in this process, the values get closer to zero, and could eventually vanish. This quick decreasing of the values makes it harder for the network to update its weights and to receive the final result [30].

Since texts are filled with long-term dependencies between sentences or even chapters, a model should be able to process these. If the model is capable of looking back in time to previously made predictions, instead of just looking at recent inputs, it would be capable of formulating better predictions [26]. In order to solve the vanishing gradient problem, Hochreiter and Schmidhuber in 1997 came up with Long-Short Term Memory (LSTM) [31]. The LSTM is a type of RNN that includes memory cells to store long-term information. A set of gate units (i.e. input gate, forget gate, output gate), decides when to open or close access to the cells. This allows for a better overall control and maintenance of long-term dependencies, and takes care of the vanishing gradient problem [21].

Even more recently, the boundaries of RNNs were pushed further by the introduction of attention. The attention mechanism in neural networks aims to solve the memory issue of previous models completely and allows for even more long-term dependencies [65]. In previous networks, intermediate states of the encoder are castaway and only the final vector is used to initialize the decoder. Although this works well for shorter sequences, it becomes problematic when sentences are too long to be summarized in a vector. In attention mechanisms, instead of discarding the intermediate states, these are used as context vectors to generate output sequences. This is done by attending to the parts of the sequence relevant for the output. Attention is frequently used in fields such as computer vision [68] and machine translation [45].

The most recent attempt of building a text generator using attention was made by AI research organization *OpenAI*. They released their transformer-based text generating model GPT-2, trained on 40 GB of text scraped from outbound links from Reddit [55]. On their website they state that “Due to our concerns about malicious applications of the technology, we are not releasing the trained model.”[11]. The texts generated by this model are impressive and difficult to distinguish from human written.

[10] https://medium.com/deep-writing/harry-potter-written-by-artificial-intelligence-8a9431803da6

[11] https://openai.com/blog/better-language-models/

This section shows that the history of text generation is subject to rapid developments. In this section, we build data-driven, DL text generation models instead of knowledge-driven ones. This because instead of having to manually create templates or logical rules and inferences, our focus is on automating the process to extract patterns. The aim is to gain information about and comparing the structures of the folk tales from the compiled corpora, instead of building the most state-of-the-art language model. Since models that use attention came to light only recently and their effect is still being studied, we opted for the most frequently researched option with RNN.

5.3 Technical implementation

In order for a West African story structure to emerge from generated text, a large amount of training data is required. The question that logically follows is: How much data is considered enough? The sizes of the corpora mentioned in the previous section (i.e. Shakespeare, Tolstoy, Harry Potter and Reddit) range from a few MBs to several GBs. Although a consensus does not exist on the amount of training data required, the general idea is that the more data is used, the better the generated text will be. Compared with the mentioned examples, the corpora of type B (see section 4) we used in this experiment are relatively small (i.e. 1.1 MB each).

When building a text generating RNN, one can choose to predict either by character or word-level. The advantage of the former is that it does not have to deal with computational complexity which increases with the size of its vocabulary. This is because in the character-level model, only a few dozen different characters exist, i.e. the letters of the English alphabet and some other symbols [42]. Furthermore, character-level models are better equipped at dealing with noisy texts. This is the case in tweets, which on average contain more spelling mistakes and abbreviations than books [11]. The drawback of character-level modelling, however, is that it takes longer to train the model since the sequence length is increased. Additionally, word-level RNNs are better at capturing long-term dependencies, because they have to make less predictions than the alternative.

Both a character-level and a word-level RNN have been built for this experiment, to compare their results. This was done by feeding both models seed sentences, and comparing the words they generate to complete them. The building, training and results of the models are explained on the basis of the West African corpus only. In section 5.5, which describes our human evaluation, the corpora of type A are used, which means a decrease by more than a half in training data size compared to when the corpora of type B were used (see section 4). In the structure of this section it makes more sense to first explain the steps needed to build a story generator before evaluating it, which is why we start with explaining this part of the experiment.

Data preprocessing The character-level model did not require much data preparation. The corpora are stored in two separate files, with each folk tale separated from the other by a blank line (see the GitHub repository[12]). Feature selection was conducted by removing the blank lines, such that one long sequence of characters separated by spaces remains. These were then split into input-output sequences of length 8, the first seven of which are input characters (X), used to predict the eighth character (y). The result is a total of 1,060,956 sequences.

The word-level model required more elaborate data preparation, even though the steps are similar. Since we now had to deal with words instead of characters, the part in which the sequences were created differs. After splitting the file into tokens, these were converted to lowercase and punctuation was removed. This reduced the number of unique tokens, and with that the size of the vocabulary. For instance, "Tree?", "Tree." and "Tree" would now be considered as one and the same (tree). Reducing the size of the vocabulary reduced the training time. Then the remaining non-alphabetic characters were removed. Finally, input-output sequences were made of length 50, the first 49 of which are input tokens (X) that predict the 50th token (y).

The final step before training the model was encoding the sequences by mapping the characters or words to integers. This was done by means of the Keras Tokenizer. The difference between the character and word-level models becomes clear in this part. While the encoded vocabulary of characters contained only 93 unique characters, the word-level vocabulary had 8,027 tokens and was thus significantly larger. The target word y in the word-level model was one-hot encoded. This means that for each word in the vocabulary, a 1-dimensional vector of size 8,027 is created filled with zeros, except for a one on the vocabulary position of the predicted target word. This rules out any similarity between words and allows the model to better predict the next word.

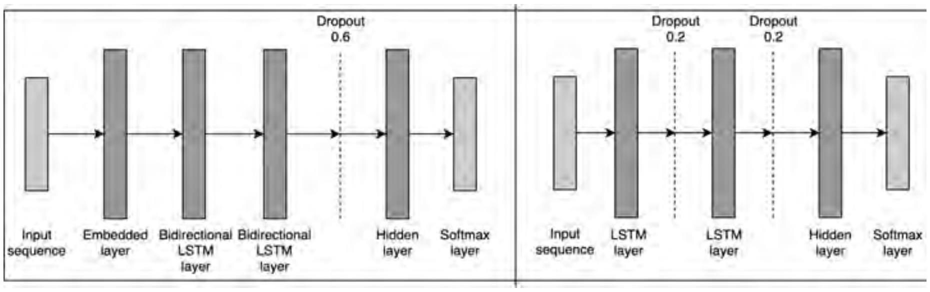


Fig. 3. RNN network architecture of word-level (left) and character-level (right) models

Training the model Figure 3 (left) illustrates the network architecture of the word-level RNN model. The input sequence is first fed to the Embedded layer, where each word has 50 dimensions. The next two layers are Bidirectional

[12] https://github.com/GossaLo/afr-neural-folktales/tree/master/data

LSTMs, each containing 200 memory cells. In a Bidirectional LSTM, two independent RNNs are merged, allowing the network to receive both forward and backward information at every time step. This may result in a higher accuracy compared to when a single LSTM is used. The Dropout layer prevents overfitting on the training data by randomly dropping out nodes during training. This is then followed by a Hidden layer and the final predictions are made using a Softmax activation function. This function is used in multi-classification tasks and chooses the next word based on the highest probability.

The values belonging to the layers and the architecture of the model have been chosen through trial and error. By tweaking the values each time new results came in, the generated texts became more coherent and fluent, while overfitting was prevented. The structure of the network was deliberately kept simple, since we are working with a limited amount of data. The loss function chosen is the categorical cross-entropy, which, like Softmax, is used in multi-classification tasks. The optimization algorithm is Adam, since it requires little memory and is computationally efficient.

Figure 3 (right) shows the network architecture of the character-level RNN model. Since this model does not use word embeddings, the Embedded layer is not part of the architecture. Furthermore, the LSTM layers are not Bidirectional, and two Dropout-layers instead of one are added. These settings yielded the best results from those we tried. Apart from these differences, the architectures are the same. Both models were trained for 200 epochs, where each epoch equals to the dataset getting passed both forward and backward through the network once. However, since the loss function of the character-level model did not further improve after 96 epochs, the output of this point was used.

One main difference between the models is that training the character-level model took a lot longer than the word-level model. In the character-level model, each epoch was trained for 2.5 hours on Google Cloud using 8 CPU nodes, resulting in approximately (200 * 2.5=) 500 hours, or over 15 days. For comparison, the word-level trained under similar circumstances only took 13.3 minutes to train one epoch. This means that to train a similar amount of epochs this model requires only 44.33 hours, or a little less than two days. Training on GPU instead of CPU nodes would significantly improve training time[13]. However, the gap in training time between the two types of models using GPU nodes will likely remain similar. Depending on the situation, this would be a reason to consider the word-level over the character-level model.

5.4 Results

In this section we compare the outcomes of the two models. For each, we use two seed sequences derived from Manu's "Six Ananse Stories" book as input [48]. The models both predicted the next words to continue the narrative based on the input seed.

[13] https://azure.microsoft.com/en-us/blog/gpus-vs-cpus-for-deployment-of-deep-learning-models/

The following snippets were used, where the bold words form the seed input and the normal text is how the narrative continues according to the book:

> **The snake went nearer and Ananse cleverly tied him to the stick with a rope. Then he** carried him home. Ananse boasted about his powers to his wife and children.
>
> and:
>
> **Long ago, there lived two very close friends, Ananse and Pig. They did everything together. They** dined, played, worked and moved about together. Wherever Ananse went, Pig was sure to go.

Now that we established the seed text and have seen how the story normally continues, we investigate how the models perform. The following examples have been generated by the character-level model:

> **Pig was thinking of Ananse's welfare while Ananse was planning Pig's death.** When the son was a large tortoise the son was a large son and said to himself, "I will be a little batehte to the same things to the shark."
>
> and:
>
> **Long ago, there lived two very close friends, Ananse and Pig. They did everything together. They** were all the same thing that he was a stranger and the sultan said to himself, "i will be a little the same thing to make the souther of the stone and the sun was a little the same.

The character-level model does not look too impressive. The first example repeats the use of the word "son", and the second example repeatedly uses "the same". Both examples are difficult to interpret, partly because the model does not capture long-term dependencies well. It is, however, interesting to see that the model is capable of generating words, when it is trained on a character-level and has a limited dataset to train on. There are two spelling errors, and the model has learned to generate quotation marks whenever something is said. Unfortunately, in the second example the quotation marks are not closed. The model uses a correct word in the second example to come after "They". When trying to identify West African features in the story, we consider the occurrence of the words *tortoise* and *sultan*.

The following examples have been generated by the word-level model:

> **Pig was thinking of Ananse's welfare while Ananse was planning Pig's death.** He told him the other side of the row, and the other birds was standing near the ground sounded beautiful firmly on the ground. And they could not get home and the animals was sitting with him on his back.
>
> and:
>
> **Long ago, there lived two very close friends, Ananse and Pig. They did everything together. They** went and placed a visit on by his subjects. And the other birds were still picking up the keeper of a few grains of corn, which by Ejuqua they all became angry and set out to pick out shouting.

This model creates sentences that are quite correct and more syntactically and semantically coherent. Both examples are well supplemented and there are no spelling errors. In the first example, two verbs use the wrong conjugations, but other than that it performs quite well in capturing long-term dependencies. If we pay attention to characteristic West African features, we highlight the occurrence of the words *corn*, *the animals*, and the character *Ejuqua*.

We can clearly see that the word-level outcome is easier and more pleasant to read than the character-level one. The former outperforming the latter is probably due to the fact that the word-level model has to make less predictions and is better equipped at dealing with long-term dependencies. Furthermore, no spelling errors occur in this model, because it does not have to learn to create words from scratch. This all adds to the readability of the text, its coherence, and we could even argue that it is slightly poetic. Furthermore, the word-level texts score better on semantic and syntactic coherence. However, in both types of models the semantic coherence is quite low. Some West African features emerge in the texts, which are mentions of food, animals, and character types and names.

5.5 Human evaluation

As mentioned, a human evaluation in the form of a survey was conducted on an experiment with data trained on roughly half the corpus (i.e. type A, see section 4). In this survey, participants were asked to rank generated texts by coherence, and to classify them according to geographical background. A total of 13 participants completed the ranking assignment, and 14 made the classification task. The texts were generated by a character-level model similar to the one previously described. The difference in network structure, however, is that this model uses only one LSTM layer. The main difference, however, is in the training data. At the moment of the evaluation, both corpora had a size of 500 KB. Since the size of the data used for training greatly impacts the outcome, they are considered to be of lower quality than those described in section 5.3.

Survey setup For each trained epoch, an example snippet of generated text is output in three temperatures (i.e. 0.1, 0.5, and 1.0). The temperature resembles the freedom of creativity, in which higher temperature texts use less of the input corpus. On the one hand this makes the higher temperature texts more creative, but it also makes it more syntactically error-prone. For this human evaluation, ten texts were selected, half of which are West African, the other half which are Western European. Eight texts are RNN generated with different temperatures (e.g. 0.5 or 1.0), and the remaining two are extracted from the original corpora.

Before distributing the survey, a list of expected ranks was made, in which the fact whether a text was generated or not and its temperature were both taken into account. Higher temperature texts are placed at a lower expected rank, and so were generated texts compared to the original ones. This list, which is illustrated in table 2, shows three ranking categories, where those placed in group 3 were expected to perform best, and those placed in group 1 worst. This makes it easier to compare the expectations with the outcomes of the survey. According to this grouping, the expectation is that the original texts (i.e. texts 5 and 6 in table 2) would be ranked as most coherent.

Table 2. Expected ranking of the texts

Nr.	**Background**	**Temperature**	**Ranking**
1	European	0.5	2
2	African	0.5	2
3	African	1.0	1
4	European	1.0	1
5	African	original	3
6	European	original	3
7	African	1.0	1
8	European	1.0	1
9	European	0.5	2
10	African	0.5	2

A text snippet that has been used in the survey is the following (text 10):

> “A long time ago, the people were starving. Very soon he had a very fine flock of sheep. The oracle revealed that he was his proper wife, he took to travel for the punishment. They kept walking through the wood until at last, on one could go and wait when he found one another.”

The same ten texts have been used in the classification task. However, in this task participants were asked to categorize each text either as *West-African*, *Western European*, or *Unclear*. Additionally, participants were asked to leave behind a comment explaining each choice.

Survey results The difference between the expected and true results are illustrated in figure 4, where the numbers of the individual texts are shown on the x-axis and the y-axis represents the values of the assigned ranking groups (similar to those in table 2). As expected, original texts *5* and *6* (high coherence) were ranked best in terms of coherence. The results of texts *1, 2,* and *10* (medium coherence) also corresponded to the expectations, as were texts *3, 4,* and *8* (low coherence). Only texts *7* and *9* were ranked differently than expected. Overall, eight out of ten texts were ranked in accordance to the categorization made prior to publication.

In the comment section of the ranking assignment, most participants admitted it was difficult to rank the texts, since most were rather incoherent. Some texts were more coherent and contained less spelling mistakes, and so they were ranked higher. Some participants described using their "gut feeling", because it was unclear to them how else to complete the task.

One of the results of the classification task is illustrated in figure 5. More than 90% of the participants categorized the text as being West African, which is correct in this case. For each classification, we consider the class with the majority of the votes. Using this method, seven out of ten texts were correctly classified. In the other three cases, the majority chose *Unclear*, followed by the correct option.

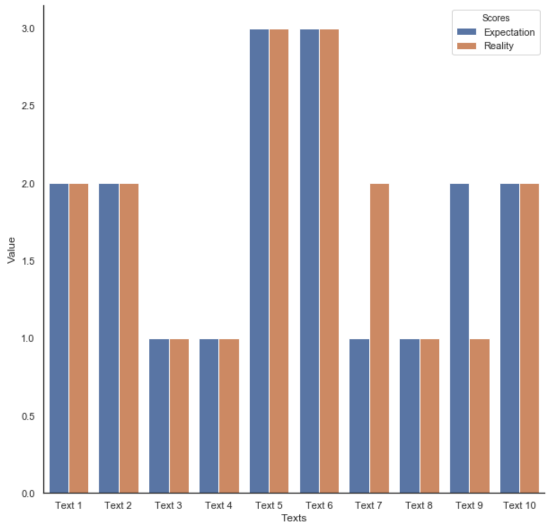


Fig. 4. Survey ranking task result

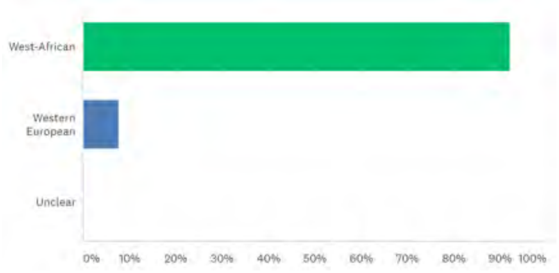


Fig. 5. Survey classification task result

The comment section below each classification provided more information about the choices made by the participants. This clarified that in most cases a choice was made based on names of main characters, animals and objects mentioned. For instance, panthers do not live in Western Europe, and so a story about a panther was classified as West African. "Reynard", on the other hand, was recognized by all participants as being the main character in Western European folklore, and a text containing this character was thus classified as such. In yet other cases the texts were classified based on their resemblance with instances from Greek Mythology, Snow White, and Shakespeare. "Once upon a time", for instance, was identified as Western European, because of its frequent occurrence in European fairy tales.

In the texts where *Unclear* received the majority of the votes, the texts were deemed unreadable and participants found it difficult to extract any relevant cues about a geographical or cultural base. Another reason why this option was chosen was in cases where animals or objects occurred in texts that exist on both continents, e.g. sheep, and oracles.

Survey discussion After reviewing the results, the conclusion to be drawn is that the generated texts are difficult to interpret. This is due to the fact that the semantic and syntactic coherence between sentences is poor. Geographical and cultural features occurring in the texts enhanced participants' belief that it originates from one of the two continents. Texts that resemble original texts are generally seen as being more coherent than those that are composed more creatively.

5.6 Discussion experiment 1

In this section, a corpus of West African folk tales was used to train a neural network in generating narratives with West African features both on a character and a word-level. Furthermore, a human evaluation was conducted to evaluate a model trained on West African and Western European input.

The RNNs with an LSTM layer proved capable of generating new words and sentences. Increasing the size of the dataset would most likely improve the semantic and syntactic coherence of the output.

The texts generated by the word-level model outperformed those generated by the character-level one in syntactic and semantic coherence, capture of long-term dependencies, and overall readability. Given the small size of the data and the fact that the models learn to generate text from scratch, the results can be called quite impressive. In both outcomes, characteristic West African features emerged, which were either African character names or types, animals or types of food.

Distinguishing texts by geographical background based on narrative structure only, which is relatively implicit, proved difficult. However, when participants of our survey focused on more explicit culture-specific words and sequences of words that are usually associated with the continent in question, such as character names, animals, and objects, this distinction was well executed. This is confirmed by the fact that participants in the human evaluation classified a majority of the texts correctly according to their geographical background.

Most narratives contain long-term dependencies and consistency in use of subjects and objects between sentences. Unfortunately, most neural networks that generate text have a hard time reproducing these, which make them difficult to interpret. Using part of the original data to generate new narratives improved the perceived syntactic and semantic coherence of the narratives.

6 Experiment 2: text classification

6.1 Introduction

The feedback obtained in the human evaluation of the previous section suggested that distinctions between geographical backgrounds can be made in use of vocabulary. Words that are deemed West African were classified as such, as were Western European ones. This motivated us to look into machine-identifiable differences between the two corpora. Is it enough to make a distinction purely based on differences in use of words, or are their less explicit differences? Is a machine capable of making a similar distinction based on geographical background by focusing on features such as main characters and environmental context? Or can we find other typically West African words or structures?

Text classification is the task of assigning documents of texts formed in natural language to one or more predefined categories. This can be done either manually or automatically (algorithmically), for instance by arranging books in a library or classifying recipes based on meal type. Text classification has several prominent NLP use cases in information and computer science, such as spam filtering, sentiment analysis, and genre classification. Furthermore, there exist various text classification approaches, each differing in complexity and efficiency. Since we identify only two classes in this project, i.e. West African and Western European, the supervised task at hand is binary classification.

In this part of the project, the performance of a DL classifier (i.e. LSTM) is compared to that of a non-neural Bag-of-Words (BoW) model, both trained on the corpora of type B (see section 4). The BoW model is relatively simple and fast to train [38] compared to the LSTM classifier. The main difference between the two models is that while the LSTM model focuses on word order in sequences, this order is completely absent in the BoW model.

The main interest in building the classifiers is to research whether we can get them to distinguish between the West African and Western European folk tales from the corpora compiled for this project. Furthermore, the distinctions made by the neural network are illustrated by means of an interactive visualization. This is done to gain more insight in the data and make the decision-making process of neural networks less abstract. Several text classification algorithms exist, some more suitable than others. The expectation is that, similar to the first experiment, the frequent occurrence of culture-specific words increase the performance of the classifiers.

6.2 Related work

Traditional classifiers relied heavily on feature engineering and feature selection. Popular feature engineering options include BoW, part-of-speech (PoS) tags, or topic models. In feature selection, features that make data noisy, such as stop words and punctuation, are removed or altered. BoW is often used on top of a simple ML algorithm, such as Logistic Regression (LG) or Support Vector Machine (SVM) [40]. In the BoW approach, a vocabulary of all the unique words in the corpus is created. Then, each document is transformed into a vector in which each value represents a term in the vocabulary. This value represents for instance the frequency of occurrence, TF-IDF value or n-gram [66].

More recently, with DL techniques becoming more advanced, text classification models based on neural networks are gaining ground. After their success in Computer Vision and Speech Recognition tasks, Convolutional and Recurrent Neural Networks are referred to frequently when it comes to text classification. These DL models have shown good results for a variety of classification tasks, such as sentiment analysis and sentence classification [69][39].

Contrary to traditional models such as BoW, RNN classifiers take word order and semantics into account. With the simple example "man bites dog != dog bites man", the difference between the two types quickly becomes clear. Since BoW captures neither the meaning of text, nor the word order, the representations in this example would be considered identical.

RNN classifiers analyze a sentence word by word, and store past information in a hidden layer. This both ensures that the models capture contextual information, and it makes them suitable for sentence classification tasks. Although they are more time-complex than more traditional classifiers, this is not too bad if the dataset is kept small. The main drawback of RNN classifiers, however, is that they are biased in assigning greater importance to words appearing later in the text than those appearing earlier. This is problematic when the semantics of the entire text is considered instead of just the end [40].

Although DL, such as those previously described, are used more and more frequently, more traditional models that use BoW are still considered relevant [6]. Simply put, not all tasks have the complexity that is required to train DL models. In some cases, traditional models outperform deep learning models, or they slightly underperform in terms of accuracy but are faster. *FastText* is an example of such a model, which uses a variation of BoW and is shown to outperform various DL models in terms of speed and accuracy[38].

Not many studies have been conducted on the classification of folk tales. Merely two papers were found that use folk tales as a basis for text classification, with differing objectives. Where the first paper aims to categorize folk tale genres, the second focuses on identifying language dialects. No research was found on identifying the geographical background of folk tales.

In the experiment described by Nguyen et al., Dutch folk narrative genres such as *legend*, *fairy tale*, and *riddle* [50] were classified using an SVM classifier. The goal was to test distinctiveness of genres and improve accessibility of folk tales. Their corpus contains approximately 15,000 manually annotated Dutch narratives written over a time span between the 16th and 21st century. Some features they used were unigrams, character n-grams (i.e. n-grams from length 2-5), punctuation and PoS patterns, the most effective which were the character n-grams. Even though the achieved results are not bad, the fact that a significant amount of narratives were classified under multiple genres proves that it is a difficult task.

The second paper by Trieschnigg et al. used the Dutch Folk tale Database to do a language identification task on Dutch folk tales [60]. Over 39,000 documents from the database were used written in 16 Dutch dialects. Trieschnigg et al. compared a number of classifiers, e.g. nearest neighbour and nearest prototype, with an n-gram baseline. Their results indicate that their input corpus made language identification difficult, at least partly due to the fact that it was annotated by over 50 annotators and it remains unclear whether each of them had used the same annotation method.

As mentioned, not every task requires a complex and DL model. Our assumption is that in this project, particular words occur more frequently in West African folk tales than in Western European ones. This is why we build a classifier using BoW feature engineering. On the other hand, examining the effect that semantics and word order have on identifying the geographical background is interesting too. Because of this, and since we would like to know whether our model improves if we use a neural network, the BoW model is compared with an LSTM classifier.

6.3 Technical implementation

Before moving on to explain the specifics of the classifiers, some data exploration is done to compare the corpora on a basic level. Table 3 gives more insight into word use of both corpora by showing the top 10 most frequently occurring words and their term frequencies. What stands out at first glance is that both lists include animals that are associated with their geographical background, e.g. lions and tortoises in West Africa, and foxes and wolves in Western Europe. This adds to our motivation to compare the LSTM classifier with the simpler, unigram-based BoW model.

Table 3. Top 10 most frequently occurring words per corpus (type B)

	West African		**Western European**	
1	day	564	little	585
2	time	533	time	513
3	tortoise	526	king	466
4	king	485	day	379
5	little	482	fox	344
6	lion	430	wolf	309
7	told	393	home	279
8	water	390	house	270
9	people	385	wife	256
10	jackal	379	reynard	241

LSTM classifier Using the LSTM to do the classification task is not too different from using it to do text generation. Each folk tale in the corpus can be seen as a sequence of words. In this task, the geographical background of the complete narrative is predicted, instead of the next word. This is therefore called a binary classification task, and not a multi-classification one as in the case of text generation.

First, the data was preprocessed by performing feature selection, in order to prepare the data for training. This was done through cleaning the texts by removing short (n<2), non-alphabetic and stop words. Subsequently, the texts were vectorized using Keras Tokenizer. This tokenizer uses a maximum number of words used, in this case 5000, based on word frequency, and cleans the data by filtering out punctuation and transforming words into lowercase.

Labels were then mapped to the folk tales: ones for West African and zeros for Western European folk tales. In addition, the maximum amount of words per tale was set to 500 to equalize the vectors. Folk tales that contain more than 500 words were padded and those with less words were truncated.

Once the preprocessing had been performed, the sequences were trained. The network consists of a set of layers through which the data is passed. The first layer of the network is the Embedded layer. Word embeddings are created, where the distance equals the similarity in meaning. Each word is represented as a 32 length real-valued vector. Using word embeddings is where the LSTM differs from the BoW model. The next layer is the LSTM layer with 100 memory units. A Dropout layer was added to prevent overfitting, followed by a fully-connected Hidden layer. The Sigmoid activation function was used to make a 0 or 1 prediction for both classes. Figure 4 illustrates the network architecture.

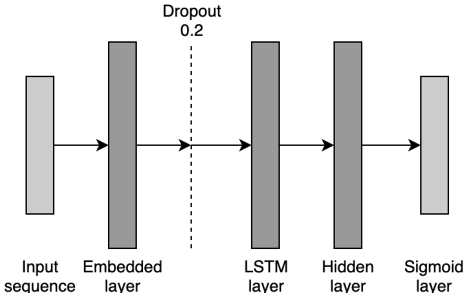


Fig. 6. LSTM network architecture

The binary cross-entropy loss function was used, as well as the Adam optimizer. The data was then trained on only five epochs. This and the fact that the network structure was kept simple was done to prevent overfitting. Finally, 10-fold Cross-validation was applied to give a more robust, less biased estimate of the performance of the model on unseen data. The metric used to assess the performance is the accuracy score, which is 0.74.

Furthermore, the model was assessed by predicting the geographical background of ten new snippets of texts. 6 out of 10 texts were classified correctly by the model. The four misclassifications were West African tales predicted as Western European ones.

Bag-of-Words classifier BoW is a method used in NLP to extract features from text documents. In this method, a document is considered an unordered bag filled with words. The frequency of each word is counted, not taking into account interrelationships between words. The main drawbacks of BoW are that it ignores long-term dependencies and does not deal well with negation [12]. This is particularly inconvenient when doing sentiment analysis. For instance, [*not great*] should be given a negative sentiment, but will most likely be classified as positive, since word order is not maintained.

The words were preprocessed by transforming them into lowercase, and removing punctuation and numbers. Additionally, the text was divided such that each line contains one sentence. After the data preparation, a vocabulary was defined containing all the words of the folk tales that occur at least twice. This reduced the vocabulary size from 13713 to 6599. Subsequently, each folk tale was modeled by counting the number of times a word from the vocabulary appears in it.

The network used for training is a simple feedforward Multilayer Perceptron (MLP). This network consists of an input layer, a single hidden layer with 50 neurons and a Rectified Linear activation function (ReLU). A Dropout layer (0.6) prevents overfitting and the output layer of one neuron has a Sigmoid activation function to make the 0 or 1 prediction for the two classes. Figure 7 illustrates the complete network architecture.

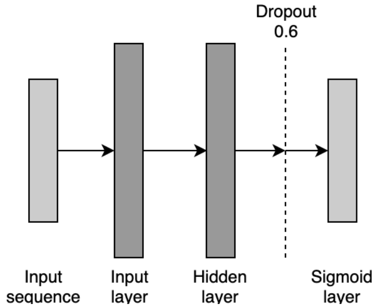


Fig. 7. BoW network architecture

Similar to the LSTM classifier, the binary cross-entropy loss function was used, as well as the Adam optimizer. The model was trained on only 50 epochs to avoid overfitting. Then 10-fold Cross Validation was applied to decrease bias and generalize the results on the complete dataset. This yielded an accuracy score of 0.93.

When testing the performance of the model on the unseen texts, 9/10 were correctly classified. This roughly corresponds to the high accuracy score acquired on the training data. The only misclassified case is a West African text classified as Western European.

6.4 Comparison of the results

The BoW (+ MLP) model (acc. 0.93) outperformed the LSTM (acc. 0.74) significantly both in terms of accuracy and in predicting the origin of the unseen texts. The results obtained using the LSTM with word embeddings show that it handles sequential data well. The fact that it performed worse than the BoW model could be due to the fact that too little training data was used. If more data would have been available, the model is capable of capturing the long-term structure. The fact that the BoW model takes word occurrences into account instead of word order or long-term dependencies resulted in its high accuracy. The previous section gave an indication of how culture-specific some words are and how this helped distinguish between geographical backgrounds of narratives. This assumption is further confirmed by the success of this model.

6.5 T-SNE visualization

One way to investigate how sentences are represented in hidden layers is by visualizing them after applying the T-SNE dimensionality reduction technique. T-SNE, short for T-distributed Stochastic Neighbor Embedding, was used to reduce and visualize high-dimensional datasets in two or three dimensions. This preserves not only the local structure but also the global structure of the high-dimensional data in the low-dimensional data. Representations of similar data points are kept close together [46].

We applied T-SNE to a hybrid model of both classifiers previously described (i.e. BoW + LSTM). First, the words were converted into numbers using a BoW model. These BoW matrices were then fed as sequences to the embedding layer of the LSTM classifier. Using this network, an accuracy score of 0.85 was achieved on the test set. Each data point was preprocessed by transforming the words into lowercase and stripping away punctuation and stop words. This resulted in a set of 1682 sequences.

The interactive T-SNE visualization can be found on our web page[14]. Figures 8 and 9 show stills of the interactive T-SNE. In this plot, each data point represents a sentence, and the colour shows its true geographical origin. Red equals the West African sequences, and blue the Western European ones. The small size of most of the data points indicates that they have been classified correctly, meaning that the true and predicted class align.

[14] https://gossalo.github.io/tsne-visual/

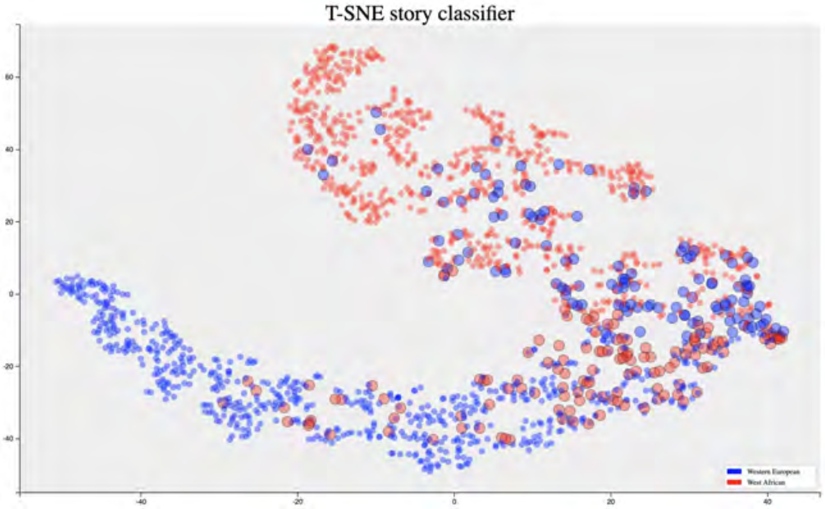
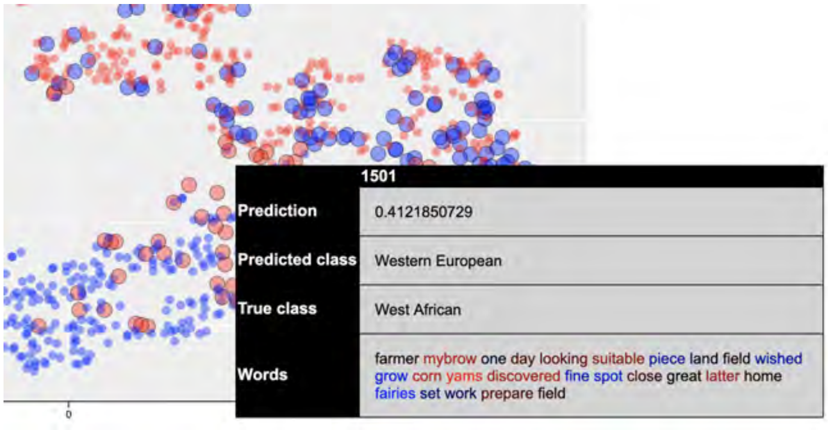


Fig. 8. T-SNE visualization



	1501
Prediction	0.4121850729
Predicted class	Western European
True class	West African
Words	farmer mybrow one day looking suitable piece land field wished grow corn yams discovered fine spot close great latter home fairies set work prepare field

Fig. 9. T-SNE table of data point

When hovering over a data point in the interactive plot, a table pops up showing the prediction value, predicted class, true class and the classified sequence (see figure 9). The words shown in the table are coloured according to their prediction value. This value is similar to the overall prediction value, where $p<0.5$ is blue for Western European and $p \geq 0.5$ is West African and thus red. The more likely that a word is classified as coming from one of these origins, the brighter its corresponding colour is. Blackish coloured words have a prediction value closer or equal to 0.5.

As shown in figure 9, some points are larger, indicating that there is a discrepancy between the predicted and true class, and that the sequence has thus been wrongly classified. As can be seen in the “Words” row, the number of red and blue words are the same. However, since the ratio of bright blue words is higher than that of bright red words, the overall sentence has a prediction value of p=0.41 and is thus classified as Western European (i.e. p<0.5).

Interesting is that the visualization clearly shows what the classifications are based on. The words that make a sentence highly likely to be West African are for instance “chief”, “kweku ananse”, and “crocodile”, whereas for Western Europe these are “reynard”, “fox”, and “castle”. The interactive nature of the plot allows to easily hover over the data and get an idea of the differences between the two corpora on a word and sequence level.

Most misclassifications are found in the center of the figure. These sequences are often short, containing less than five words. Since in a short sequence each individual word plays a larger role in changing the overall prediction value, this makes the sequence more prone to be classified incorrectly. Furthermore, in the longer misclassified sequences (i.e. n>5), the number of more vaguely colored words are more or less equal, with one bright outlier word causing the entire sequence to end up in the wrong class.

Misclassified sequences with a high prediction value, e.g. when we consider the sequences with prediction value p>0.95 for West African and p<0.05 for Western European, are more predictable. “Ant”, “hunter” and “lion” are presumed to be West African but belong to Western European folk tales. This is an understandable misclassification, as these would be words that one associates with West Africa rather than Western Europe. Similar cases exists the other way around, for instance when “King”, “kingdom”, and “pudding” are classified as Western European, but are actually West African.

6.6 Discussion experiment 2

This section compared a BoW-based and an LSTM classifier, using two corpora of folk tales from West Africa and Western Europe as input. The main interest was to research whether a machine would be capable of distinguishing between these narratives. The approaches used by the classifiers have been analyzed and compared. Furthermore, the predictions made by a deep learning classification layer have been visualized by means of a T-SNE interactive visualization.

The reason why the two classifiers were compared is that one counts word occurrences (i.e. BoW), while the other predicts on word order in sequences (i.e. LSTM). Both classifiers proved sufficiently capable of distinguishing between the classes. The BoW-based classifier (acc. 0.93) outperformed the LSTM classifier (acc. 0.74) significantly. Although the training set is arguably too small for the LSTM to perform better, the high score obtained by the BoW model indicates that distinguishing in use of words might be all it takes. More generally speaking we could state that West African and Western European tales use quite distinctive and culture-specific vocabulary.

Creating the T-SNE visualization, which uses a hybrid model comprised of a combination of the BoW and LSTM network, proved the ideal approach to demonstrate and confirm this distinction. Sentences of the folk tales were represented in a high-dimensional fashion in the hidden layer of the LSTM. The T-SNE showed that words that we would associate with West Africa or Western Europe had a high likelihood to be placed in their expected class. Misclassifications occurred with sequences containing words that occur similarly often in both corpora. The results confirm the observations described in section 5.5, in which participants of the survey made similar choices based on use of characters, animals, and objects.

7 Experiment 3: narrative structure analysis

7.1 Introduction

The previous section compared the West African and Western European corpora on occurrences of words. Instead of researching differences by directly comparing the corpora, the emphasis in this section lies on identifying and further exploring the narrative structure of individual narratives and their components. This is done in order to examine whether ML and NLP technologies can be used to identify and extract culture-specific patterns from different parts of folk tales.

The "theoretical background" section described a diverse set of narrative theories, that have been used in the past to enable categorization and identification of different narrative parts or acts. First attempts to use narrative theories to find a common structure using NLP algorithms proved difficult. This was mainly due to the fact that there is little consensus between researchers about the theory most suitable to be applied across a set of different narratives. Another challenge is that the knowledge about narrative theories is mostly implicit and is quite difficult to capture, let alone generalize, over a set of rules. Applying these theories to develop AI story systems or to computationally analyze narratives often requires input from narrative theory experts [24].

For this part of the research, a field trip to Ghana was conducted during which several story(telling) experts were interviewed. These experts either work or have worked for prominent radio and television stations throughout the country. Furthermore, they have years of experience with storytelling and applying narratives in their work. These interviews together with the literature study provide the basis for the technical implementation of the experiment.

In the first part of this experiment, we use one of the aforementioned narrative theories to divide and categorize the parts of folk tales from the corpora of type B (1.1 MB of text per corpus). These parts are then labeled and fed to a set of classifiers. Furthermore we will dive into the narrative parts and their structures to analyze them on a sentence level. One objective in this section is to extract narrative structures from folktales and compare them. Additionally, our aim is to search for recurring patterns indicative for the existence of a distinctive narrative structure between West African and Western European in parts of folk tales. To date, no comparable research that uses ML classifiers to identify narrative structure components has been conducted yet.

7.2 Related work

As early as 1986, Habel wrote that AI research mainly concerned story understanding and generation instead of considering stories as linguistic entities on their own. The main emphasis is on the content and knowledge level, instead of on form and structure. He furthermore argued that the AI models could increase in power if they would consider and integrate linguistic markers (e.g. indicative words such as 'despite' and 'nevertheless') and their function in story processing.

If knowledge and language levels were to be integrated, this would benefit the theory of narratives in AI [29].

Some AI researchers used specific narrative theories to examine folk tales. Propp's theory proved particularly popular and was frequently used to analyze narrative structures. Finlayson, for instance, applied ML to extract Propp's functions from a subset of semantically annotated text, claiming it to be "the first demonstration of a computational system learning a real theory of narrative structure" and "the largest, most deeply-annotated narrative corpus assembled to date." His objectives were to improve the understanding of the higher-level meaning of natural language, advance our ability to extract deep structure from complex texts, and to increase computational understanding of cultural cognition, influences and differences in stories. By adapting Propp's descriptions, he set up rules to annotate 15 folk tales from the corpus. One of the main contributions was the development of an algorithm to extract narrative structures from annotated stories, called *Analogical Story Merging*). He clustered semantically similar events involving specific characters, as well as the position of the events in the story. After training the model, three important function groups emerged with high accuracy scores [19].

Since our corpora are not annotated, we are also interested in studies that use unannotated texts. Valls-Vargas et al. used NLP and ML tools on Propp's corpus to automatically extract narrative information from unannotated text. They created a non-linear pipeline framework to extract different layers of explicit domain knowledge from stories. This information was then fed back to the system to improve the performance of earlier completed tasks. The main information extracted were mentions of entities and characters, such as verbs used to describe and identify characters and their role (e.g. villain or hero). The final goal was to use the information to generate new stories. The system performed particularly well at identifying characters and mentions of entities, but failed to identify subjects and objects of verbs [63][64].

Over the years, narratology has come to play an increasingly important role in (interactive) storytelling for video games. A common narrative structure used in computer games and films is the previously described three-act structure. As was mentioned in the "Theoretical background" section, these acts are named *setup*, *confrontation*, and *resolution*. Even when narratives can be further subdivided into smaller parts, there still exists a more general three-act plot [35]. Several computational systems whose contents are based on narratives, such as the hypertext reading system StorySpinner and interactive storytelling system Fabulator, use a three-act structure [33][41].

The use of linguistic markers mentioned by Habel is in line with our aim to investigate the folk tales on a word level. The words he mentions, such as 'despite' and 'nevertheless' express contrast of texts. In this experiment we will equally search for linguistic markers in folk tales that are indicative for narrative parts. Similar to Finlayson and Valls-Vargas et al., ML classifiers will be used to classify texts and NLP tools are applied to extract knowledge from the folk tales. In this experiment, however, we follow the ratio of the three-act structure as a way to divide the folk tales and define the classes. This structure is chosen because of its frequent use across many domains and because it is less complex than other structures described in section 2. Since our tales have not been annotated, we prefer to keep the structure as simple as possible.

7.3 Fieldtrip to Ghana

A large part of West African written literature finds its basis in oral storytelling traditions. In order to learn more about these traditions, a field trip to Ghana has been conducted. During this field trip, interviews with storytelling experts were conducted on three separate occasions. The names of the interviewees have been removed to guarantee their privacy. The first interviewee is a radio broadcaster for a farming radio station. The second interview was held at a local radio station and involved a group of five employees and volunteers. The final interviewee is a former television and radio broadcaster. The following sections describe the key points extracted from the interviews.

The first interview involved a former employee of local radio stations, who is currently working at a farming radio station as a consultant. He assists radio stations in designing their programs according to international standards.

The interviewee clarified that storytelling is an event mostly occurring in rural areas, whereas citizens are generally too occupied. The storytelling format is that elderly teach younger people important topics through stories. These stories are often interspersed with musical songs, where someone start singing, and the whole audience joins in while clapping. The event is meant both as a teaching moment and for entertainment. The stories are formulated around a person encountering a problem, which is eventually solved.

The stories collected by the radio station where the interviewee is employed are created by producers in close collaboration with the community. An example story of a program broadcast is about farmer whose land is ravaged by a plague. The farmer decides not to use pesticide, and his land worsens. Another farmer, who does use pesticide, has healthy crops. The solution is for the first farmer to start using the pesticide.

The second interview took place at a local radio station in a small town on the southeast coast of Ghana. The station has stored a large number of tapes containing stories about the spider Kwaku Ananse in the local language. To record these tapes, the employees went to the neighboring villages to attend storytelling events taking place in the evening.

The interviewees, who are all employees and volunteers at the radio station, stated that Ghana is still for a part a non-literate and communal society, and that storytelling replaces literary matters. Storytelling is used to “make people understand, accept, and also become communal in thinking in relations.” The stories are used to build and maintain relationships and cultural values and to learn more about history and the environment. Besides serving an informative purpose, it is also used to entertain listeners. The characters, animals and plants, represent human beings, and teach "both the good side and then the bad side of society." Stories have been transformed into idiomatic expressions, proverbs and songs, to provide people with information.

Most of the stories try to instill fear in the listeners as a way of prohibiting certain actions. Popular Ghanaian belief is that there exist two worlds: the physical world and the spiritual one. One's wrongdoings in the physical world are to be punished in the spiritual one. In the physical world, animals and plants do not talk or fly, but in the spiritual realm they do. This way, the spiritual world instills discipline and "fires your imagination, just as in your literary world in the West, people are able to imagine by drawing and putting things together." Most of the proverbs and idiomatic expressions that are used nowadays are derived from these stories.

The interviewees also explained that defining for Ghanaian storytelling are the intermezzos in the form of singing, either acapella or by using instruments such as drums. In fact, some stories were originally songs, and were only later transformed into stories. The singing of songs is seen as a form of language preservation. An important distinction between storytelling in the West and in Ghana is that where the former speaks of *rights*, the latter is more concerned with *relations*. The way in which a particular character relates to the others often brings up rights issues. Relations are a recurring theme in almost every Ghanaian story. Additionally, the interviewees pointed out that storytelling is a dying tradition, and is less common in modern education. Parents, because of work, lack the time to tell their children stories. Another reason for the dying tradition is the migration from rural areas to cities.

Storytelling is often used as a conflict resolution strategy to diffuse tension, which is useful e.g. during meetings, election time, and in court. The use of animals in stories is to avoid directness. Being indirect stimulates freedom of expression, as it is a way of insulting or expressing your dissatisfaction with a person, even if it is someone prominent like a chief, without there following direct consequences.

Even with the presence of other media such as movies and series on television, oral storytelling still effects young people. The reason for this is that "stories are imaginable, they put fear in you, they help you to exercise self-control".

Some of the main characters that appear in the stories are animals such as the elephant, crocodile, python, lion, and the spider. Most Ghanaian stories are animal trickster tales, in which the spider Ananse is the main character. Then there are the spiritual beings, such as fear-instilling *Agbesiakoornye*, a long haired tall woman walking on one leg, who grabs small children who are too far from home, cuts them into pieces and eats them. Technology and machines are not part of these stories. The actions usually performed by machines are instead played by spiritual beings with magical powers.

The context in which the stories take place are usually the forest, the market place, the community, or migration. Environmental elements such as trees, the sky, and rivers, come alive and have the ability to talk. Another important element in Ghanaian stories is the hierarchy in the relationships among people, which is used to show that "Not all people are the same. There are princes, chiefs, slaves, and a lot of stories make the vulnerable important."

The structure of the stories is dependent on the theme of the story, the target audience, and the region in which it is performed. Tragedies have a different beginning and end than comedies. In general, the story structure is as follows:

- The storyteller asks the audience for permission to tell the story;
- the audience reacts in verbal assent;
- the storyteller introduces the characters;
- the storyteller tells the story;
- if the story line is not fluent or to re-energize the audience, one of the members of the audience interjects with a song;
- the storyteller continues to tell the story;
- the storyteller summarizes the story and explains the reason why it was told (e.g. "and that is why the zebra has different colors").

The third interview was held with a former broadcast journalist of a large and popular government-funded Ghanaian public radio and television network. He used to broadcast the news in the local language, presented a popular youth and a traditional, cultural program, and did sports and ceremonial commentary.

The interviewee explained that the main purpose of storytelling is to educate children by teaching them moral lessons. The stories are about animals and relate to human life. The important thing in teaching lessons through stories is that they help (re)shape the life of the listener in either a direct or indirect way. This indirectness is achieved by using stories to change the characters from humans to animals. For instance, by changing the name of the person concerned to Kwaku Ananse to prevent offending someone directly. Additionally, stories bring guidelines in life and teach people how to behave.

The way storytelling has changed over the years in Ghana is that while it is still regarded as important in rural areas, it is much less popular in the cities. However, some television stations want to bring storytelling back. The reason behind this is that "they are seeing a lot of lessons in it and our culture and tradition is dying. Apart from the storytelling, there is a lot of our culture in it, which the children learn. Putting on clothes, and how to dress traditionally, they put it on in this setup." In the rural areas, radio stations still broadcast storytelling shows to bring informative messages across.

A storytelling event in a rural community takes place in the evening. While a story is told, a member of the audience can start a musical intermezzo. Most of the times, these are folklore songs known by everybody, and so the audience sings along. The topics of the songs are not necessarily related to the story itself. After this intermezzo, the storyteller is asked to continue telling. A reason for this intermezzo is to help the storyteller catch his breath when he is tired of talking and to "add colour" to the story. Besides singing, the audience can ask questions about the story to make it more interactive and participatory. Once the storyteller has finished talking, the villagers discuss the events recounted in the story.

In some stories, moral lessons are taught by instilling fear into children. An important saying in Ghana is that you should not sing while bathing, which could cause your mother to die. This is derived from the fact that the type of soap usually used in Ghana contains a poisonous element. By singing while bathing, one would have to open one's mouth, which is how the soap could enter and kill the person. Other stories teach children about laws and the importance of common sense.

According to the interviewee, a Ghanaian story starts with the storyteller shouting a slogan to let the audience know that he is about to begin. In response, the audience shouts back a slogan indicating that they are ready. The story involves a specific character, the problems he encounters, and the actions taken to overcome them. The story ends with a summarizing advice on why (not) to engage in the event. Finally, once the story has been told, the children will sing, dance, and argue about its message. Once they get back home, they may discuss it with their parents, who link daily events or rules to the stories in order to generate a broader understanding.

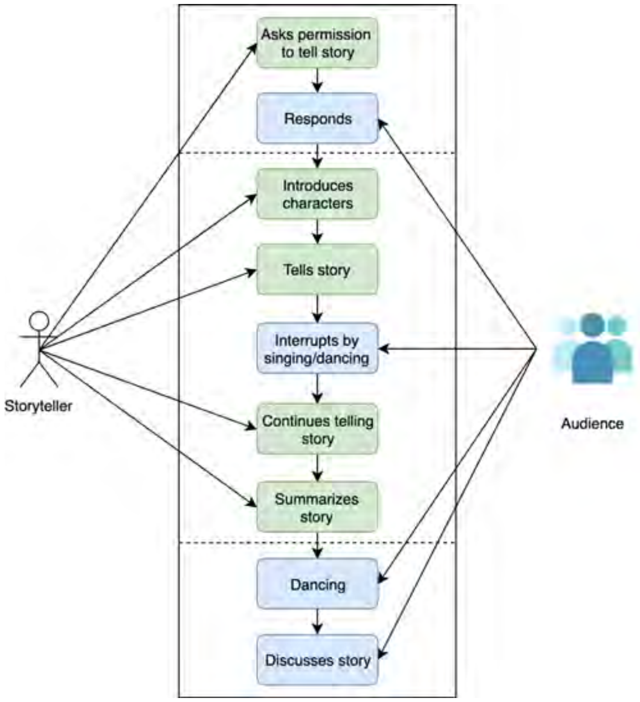


Fig. 10. Ghanaian storytelling structure

Synthesis. Figure 10 illustrates the storytelling structure as aggregated from the interviews. The storytelling structure has been determined to make a comparison between the storytelling structure and that of the written folk tales. The part above the upper dotted line illustrates the initial call-and-response between storyteller and audience. In the middle part of the event, the main characters are introduced and the story is told, interrupted with songs and dances performed by the audience. Towards the end of the middle part, the story is summarized. The part below the lower dotted line illustrates the events taking place after the storytelling. The audience dances and reflects on the content and moral message of the story, after which the storytelling event is brought to an end.

7.4 Technical implementation

Because of the lack of consensus in narratology in appointing a universal narrative theory for folk tales, and the fact that the interviews and literature mentioned the existence of a beginning, middle and ending in West African folk tales, each folk tale was divided into three parts. The most well known and implemented structure is the three-act structure, which is why it was used in this experiment to establish a baseline.

The three-act structure generally follows a 1:2:1 ratio for the begin:mid:end parts respectively. For the first part of the technical implementation, we used ML classifiers to examine whether the three-act structure is a correct way to divide the narratives. In the second part, we examined the occurrence of specific sequences of words to evaluate whether they match what was said in the interviews.

Classification of narrative structures In this experiment we investigated whether there are words or word sequences characteristic of individual narrative parts. The input provided by the storytelling experts (see figure 10) served as a guideline as what to look for. Since both the interviews and the literature did not clearly indicate whether the rules that apply to storytelling are similar to those for written folk tales, this assumption was tested in this experiment.

Both corpora were used to train the classifiers. However, in contrast to the previous section, in this experiment separate models were trained for both corpora instead of a merged one. The first step in preparing the data for training was to divide each story into multiple parts according to information acquired from the literature study. Each part was then assigned a label and was prepared for training. The data was cleaned by changing the words to lowercase, removing non-alphabetic symbols and stop words.

Since many supervised text classification models exist, a number of frequently used ones were trained and their accuracy scores were compared. To generalize the results and make optimal use of the data, 10-fold cross validation was performed to calculate the accuracy of the classifiers instead of splitting the data into train/test/validation sets. Furthermore, each 10-Fold Cross Validation was performed ten times, each time using different folds, and the mean accuracy was used as a metric to reduce variability. This was done because, despite using k-Fold Cross Validation, the results still deviated for each run.

Multiple ML classifiers were trained on the data. The reason behind this is that in text classification there is no perfect fit. The performance of each classifier is very much dependent on the size and structure of the texts and how they were preprocessed. Instead of seeking a task-specific best algorithm, they were compared to see whether the different splits would effect the algorithms similarly.

- **Naïve Bayes (NB)** is the most simple classifier based on Bayes' rule, calculating the fraction of times a word appears among all words in a set of documents. Although NB is seen as relatively old compared to newer and more complex algorithms, it still performs well in many text classification tasks.
- **Linear Support Vector Machine (SVM)** is a popular algorithm of choice whenever the sample size is small. SVM is based on finding an ideal hyperplane between two or more vectors. The features that are used to determine the final position of a document in the feature space, are words, of which the occurrences are counted. Since longer documents will have higher average count values than shorter ones, tf-idf values for each word are used as input, to place more emphasis on distinctive words.
- **Logistic Regression (LR)** can be used for binary classification, but is also well applicable in multi-class classification tasks. The method learns the probability of a sample belonging to a class by finding the optimal decision boundary that is best at separating the classes. It is similar to the NB classifier in that both aim to predict target y given x. The difference is that NB is a generative model, it first models the joint distribution of x and y before predicting P(y|x). LR, on the other hand, is a discriminative model, which directly predicts P(y|x) by learning the input to output mapping.
- **Word2Vec Logistic Regression (Word2Vec)** is a pretrained model, using the Gensim model available online. In Word2Vec, semantics play a role such that similar meaning words have similar word embeddings. In this case, the words are tokenized and word vector averaging is applied to each word to find the importance of each word in a document. Then, the averages are fed to a Logistic Regression algorithm.
- **Term Frequency Logistic Regression (TF)** counts the occurrence of each token and considers tokens occurring twice or more, with the exception of stop words. This is fed to a simple Logistic Regression classifier to perform the classification task.

As mentioned in the beginning of this section, the training data was divided in multiple ways, and the performance of each of the classifiers was compared. In half of the cases, the data was divided into three parts: begin, mid, and end. This version is hereafter referred to as the "split in three". However, since no information emerged from the literature and the interviews about the specific structure of the middle part, in the other half of the cases this part was removed.

This was done to test whether the assumption that narratives in oral storytelling have a distinctive beginning and ending can be extended to written narratives. The training data was therefore divided into two classes: *begin* and *end*. This version is hereafter referred to as the "split in two".

Tables 4 and 5 illustrate the difference in mean accuracy between the classifiers for both corpora. Furthermore, it shows three ways in which the folk tales were divided: "split in half", where both parts of the folk tale (i.e. begin and end), include 50% of the sentences. In the "split at 25%", the beginning and end part both make up 25% of the sentences. In the "split at 10%", the beginning and end part both contain only 10% of the sentences. The ratios illustrated in the tables indicate the portion of the training data provided for each part.

Table 4. Mean accuracy of classifiers for different splits - West Africa

	Split in half		**Split at 25%**		**Split at 10%**	
	begin:mid:end	begin:end	begin:mid:end	begin:end	begin:mid:end	begin:end
	-	1:1	1:2:1	1:1	1:8:1	1:1
NB	-	37.0%	34.1%	57.4%	43.9%	73.2%
SVM	-	45.4%	44.5%	66.2%	72.0%	77.8%
LR	-	48.8%	45.0%	67.3%	71.4%	77.3%
Word2Vec	-	60.1%	48.0%	70.4%	67.0%	73.3%
TF	-	55.0%	53.0%	70.5%	74.8%	77.9%

Table 5. Mean accuracy of classifiers for different splits - Western Europe

	Split in half		**Split at 25%**		**Split at 10%**	
	begin:mid:end	begin:end	begin:mid:end	begin:end	begin:mid:end	begin:end
	-	1:1	1:2:1	1:1	1:8:1	1:1
NB	-	59.5%	43.5%	63.4%	50.4%	66.1%
SVM	-	65.5%	51.3%	75.2%	66.3%	77.0%
LR	-	64.7%	50.6%	73.5%	67.6%	79.1%
Word2Vec	-	64.2%	52.6%	69.0%	69.4%	75.6%
TF	-	65.8%	53.0%	74.4%	66.1%	75.5%

We first consider the performance of the classifiers on the West African corpus, or table 4. One thing that stands out when comparing the *split in two* with the *split in three* is that in both cases and for each classifier, the *split in two* has a significantly higher mean accuracy. Furthermore, in all cases, we see that the smaller the beginning and end parts get, the more the mean accuracy increases. The highest accuracy scores are found in the 10% begin:end group. This indicates that the assumption made in the interviews about there being a clear beginning and ending in storytelling, where most of the difference is found in the first and last sentences, is also the case for the written folk tales from the West African corpus.

When we compare the classifiers, Word2Vec seems to outperform most other classifiers up until the 10% split. This makes sense, since Word2Vec requires more training data. As is, only 20% (i.e. 10% begin, 10% end) of the corpus is used, making the dataset too small for Word2Vec to perform well. NB performs worst in all circumstances, with one outlier (i.e. 43.9%) in the *split in three* for the "split at 10%". Besides this, the classifiers perform quite similarly, with no major outliers. The overall highest mean accuracy was obtained by the TF algorithm and is 77.9%.

The classifiers were also trained on the Western European corpus, to compare results. At first glance, the results seem similar. Just like in the previous case, the smaller the beginning and ending class get, the more the mean accuracy increases. Moreover, the NB algorithm is again performing worst. The overall mean accuracy is a bit higher for the Western European corpus i.e. 64.8%) compared to the West African one (i.e. 60.5%). This is mostly due to the fact that the "split in half" performs quite well both compared to the other splits and to the other corpus.

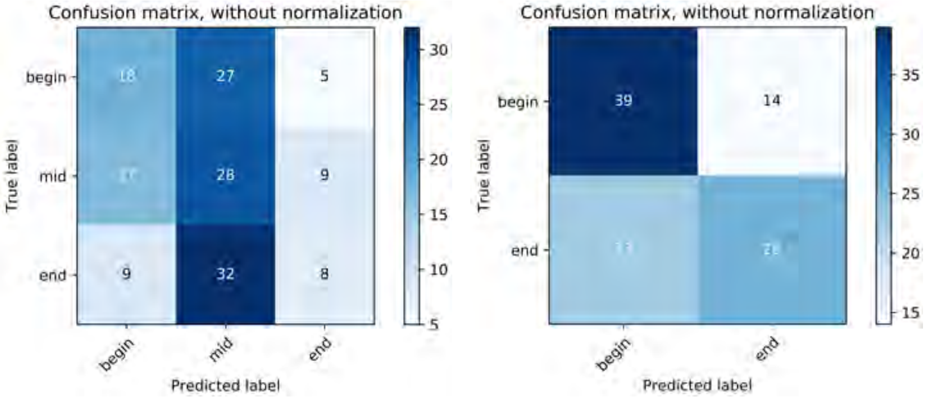


Fig. 11. Confusion matrix "split at 25%" for NB - begin:mid:end vs. begin:end

To understand what happens when the middle part is removed, two confusion matrices of the "split at 25%" group were compared in figure 11. 20% of the data was used for the test set where the confusion matrices are based on. As can be seen, in the *split in three* case on the left, most incorrect classifications occur in the "end" class, where only 8 out of 39 cases have been correctly classified. The number of correct classifications in the "begin" class is also relatively low. The "mid" class performs better, with 28 out of 54 of the predictions being correct. The low recall rate for "begin" and "end" is due to the fact that most of their instances are incorrectly placed in the "mid" class. Either this could have been caused by a bias in favor of the majority class, or because the "mid" class is less distinctive than the other two classes, causing doubtful cases to be classified as belonging there.

The confusion matrix on the right shows the *split in two* case, which shows more promising results. As was the case in the confusion matrix on the left, the recall in the "begin" class is higher (i.e. 39 out of 53) than in the "end" class (i.e. 26 out of 49). This suggests that the vocabulary used in the beginning part of West African folk tales is a better predictor for its class than in the case of the "end" class.

Figure 12 shows the confusion matrices for the "split at 10%". The recall rate of the mid class is quite high here, almost all its instances have been classified correctly. Since the mid class now occupies 80% of the data and most of the instances of the other classes are placed here, this strengthens our previous assumption that the algorithm is biased because of the imbalance in class size. Little to no instances from the beginning class have been incorrectly placed in the end class and vice versa.

When we consider the figure on the right in which the middle class is left out, we see that the two remaining classes are indeed quite distinctive. The "beginning" has a higher recall rate (i.e. 48 out of 53) than the "end" class (i.e. 15 out of 34). Since both classes contain the same amount of sentences and are therefore well balanced, this is a good indicator that we are likely to find a recurring pattern in both the beginning and end sentences of the West African folk tales, especially in the beginning.

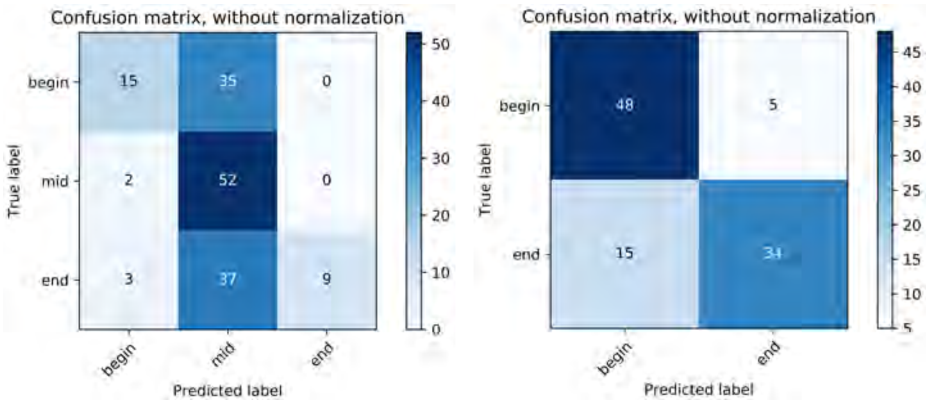


Fig. 12. Confusion matrix "split at 10%" for NB - begin:mid:end vs. begin:end

7.5 Term frequency of n-grams

In this section, recurring patterns in the beginning and ending sentences are analyzed. The inspiration to calculate term frequency (TF) values for n-grams came from the expert interviews held in Ghana. The initial asking for permission by the storyteller to begin telling and the reply that follows by the audience is indicative for the beginning. The ending of storytelling is identified by the storyteller summarizing the story. The examples mentioned were “that is the reason why men don’t have breasts”, and “why the zebra has different colors.”

It is well known that Western fairy tales tend to start with the famous words “Once upon a time” and end with “and they lived happily ever after”. TF values could easily display whether a similar beginning exists for West African narratives. Since there is no information available on the structure of the middle part of West African narratives, the focus lies on the other two parts.

In order to find the most-used n-grams, the 10% splits for the beginning and end parts of the narratives were used again. Once the data has been cleaned, Sklearn’s TfidfVectorizer is used to calculate the top TF values for the n-grams. This object allows to set an n-gram range between 1 and a value >1. First, a range of (2,5) was considered. However, choosing a low value, such as in the case of 2-grams, generates uninteresting results such as “he was” and “went to”, which occur frequently in the narratives. Choosing 5-grams resulted in many of the same combinations of words as in the 4-gram case, with one additional word that did not change the content of the sequence. For this reason, and because the sequences mentioned by the storytelling experts (e.g. “that is the reason”) were $\geq$4, 4-grams were selected.

Table 6. TF-IDF 4-grams beginning of 252 African (left) and 490 European folk tales (right)

4-gram	**Term frequency**	**4-gram**	**Term frequency**
once upon a time	25	once upon a time	27
upon a time there	17	upon a time there	20
a time there lived	10	a time there was	17
a long time ago	9	time there was a	16
time there lived a	9	there was once a	14
there was once a	8	once on a time	12
a time there was	7	there was a king	6
very long time ago	6	a long time ago	5
time there was a	6	a donkey and a	5
a very long time	6	a lion and a	5
there was a man	6	and was just doing	4
man and his wife	5	was a king who	4
once there was a	5	was just going to	4
long time ago in	5	caught sight of a	4
wanted to marry her	5	on a time a	4

Tables 6 and 7 show the top 15 most occurring 4-grams for the beginning and end parts of the narratives, and their frequencies of occurring. When interpreting these results, one should take into account that the term frequencies are quite low in all four cases, given the fact that we have 252 West African and 490 Western European tales in total. Nonetheless, the results indicate that ML and NLP technologies are useful in getting an overview of recurring patterns in narrative structures.

Table 6 shows that there is some overlap between the word sequences used in the beginning of the narratives (i.e. 26.7%). There is no overlap between the 4-grams in the ending part of the tales displayed in table 7. Thus, similarities seem to occur more frequently in the introduction of the story, which we already concluded in the previous part. Both backgrounds include time indications referring to the past, such as "once upon a time" or "there was once a". Table 7 (left) confirms the existence of summarizing word sequences in the West African folk tales, explaining why or how something is the way it is.

Table 7. TF-IDF 4-grams ending of 252 African (left) and 490 European folk tales (right)

4-gram	**Term frequency**	**4-gram**	**Term frequency**
from that day on*	8	rest of their lives	5
and from that day*	7	that he had been	5
and that is why*	7	stop him eat him	4
ever since that time*	5	for rest of their	4
and that is how*	4	he said to himself	4
passed a law that	4	and they lived happily	3
that is reason why*	3	more than a match	3
that i bought for	3	as soon as he	3
for many years and*	3	fast as he could	3
gazelle that i bought	3	her that he had	3
that for future no	3	they lived happily together	3
since that time whenever*	3	fox laughed and said	3
what are you doing	3	told her that he	3
for future no one	3	said that he was	3
did not want to	3	that i did not	3

Interestingly, the presence of a summarizing ending in West African storytelling, as indicated by the storyteller experts, seem to hold true for the folk tales too. The 4-grams on the left of table 7 show that the West African folk tales most frequently end with sequences such as "from that day on", "and that is why", or "ever since that time". In fact, we argue that 8 out of 15 4-grams can be categorized as "summarizing". Each of these "summarizing" 4-grams have been indicated with an asterisk in table 7. When we compare this to the right side of the table, which shows the Western European 4-grams for the ending of the folk tales, we see that the summarizing ending is barely present here. A well known example of a fairy tale ending of Western which also occurs in our corpus is "and they lived happily...".

7.6 Discussion experiment 3

In this section, we dived deeper into the parts making the narrative structure of folk tales. This was done using the literature available on narratology, and by conducting a field trip to Ghana in which storytelling experts were interviewed. We aimed to find similarities between West African oral storytelling and more contemporary written folk tales and to further identify West African elements in narratives. Given these sources, the aim was to see if NLP technologies could be applied on the corpora to extract recurring patterns in narrative structures.

Storytelling still holds a strong place in Ghanaian society, and its elements are passed on through generations and can be identified in written narratives as well. The storytelling experts indicated that West African storytelling is characterized by the typical interaction between storyteller and the audience. The storytelling structure has a very typical beginning and ending, with the former being a call-response between storyteller and audience, and the latter being an advisory summary of the tale.

For the technical implementation, each folk tale was divided into three parts following the three-act structure. Several ML classifiers were applied to the data, which achieved high accuracy scores when the beginning and end parts were kept small (i.e. 10% of the data). The high performance in the classification task indicates that these parts use characteristic sequences of words. Furthermore, 4-grams were extracted from the beginning and ending sentences, again using a division of 10% for both parts.

The results showed recurring patterns emerging for both the beginning and ending parts, where the former were more similar between the two corpora and the latter were particularly distinctive. Furthermore, the ending confirmed the presence of a summarizing ending, as was mentioned by the storytelling experts about the storytelling structure. This indicates that, even though narratives have changed throughout the years, some elements from the storytelling traditions are still visible in written West African folk tales which emerge when NLP technologies are applied.

8 Discussion

One important consideration regarding the corpus and the methods used is the fact that the West African corpus is written entirely in English. Although the countries where these folk tales originate from are Anglophone, English is mainly used as a second language instead of being the mother tongue. West Africa has a long history of oral storytelling, in which written literature emerged only more recently as en effect of colonization. One could therefore argue that the corpus used is less authentic than it would have been if it were written in the mother tongue of the authors. If we want to analyze cultural differences in written literature more deeply, we should consider using narratives written in the mother tongue or that have been translated to English in a very precise way.

Another point is that although the use of only one corpus for West African narratives suggests otherwise, West Africa is extremely diverse both in languages and in social and cultural beliefs, habits and customs. For the sake of this project, the different countries forming West Africa are grouped together, as are those from Western Europe. This allowed us to make general observations and comparisons. However, if we want to draw more definitive and specific conclusions about the cultures, the diversity between countries should be taken into account.

A limitation of the research is the small body of data. Unfortunately, not many more folk tales, especially West African folk tales, could be found online. ML and DL applications, however, normally need more data than the 2.2 MB used in this project. Increasing the size of the corpora would probably lead to more sound results. Especially DL models, such as the RNN algorithm, need a vast amount of data to perform well and to prevent overfitting. The model described in section 5.2 that was trained on 2.8 MB of Harry Potter texts yielded much more coherent texts than our results described in section 5.3. Doubling our data in size, for instance by building a crowdsourcing platform to collect new tales, is recommended.

Finally, another interesting focus for future work would be to provide the corpora with annotations. In related research, classifiers were trained on annotated texts in order to extract narrative structures. This could possibly improve our understanding of the data and extract more meaningful and different recurring patterns from the narrative structures. The annotations could be supplied by consulting storytelling or narratology experts, or by means of the previously mentioned crowdsourcing platform.

9 Conclusion

This master thesis aimed to identify the role of NLP technologies in analyzing and generating West African folk tales. This was done to investigate the following research question:

> *How can Machine Learning and Natural Language Processing be used to identify, analyze and generate West African folk tales?*

The analysis shows that NLP and ML techniques can indeed contribute to automatically extract, analyze, and generate culture sensitive and informative features from West African folk tales. Furthermore, it was proven that NLP technologies can be used to do a comparative analysis between West African and Western European folk tales. The added value of using NLP compared to conducting more traditional and manually extensive narratology research is that the former allows for a faster and more precise analysis of large amounts of data, in which patterns difficult to manually identify emerge.

A main contribution of the research is the collection of two corpora of folk tales. These were used as input for the ML and NLP text generation and classification models. We furthermore presented a human evaluation conducted to assess the text generation model. This evaluation indicated that the generated texts, although lacking a clear syntactic and semantic coherence, contained several culture-specific elements.

The classification task proved successful in identifying cross-cultural differences between West African and Western European folktales. The classification models and the T-SNE interactive visualization demonstrated the weight of each word within the corpora. Words characteristic for either culture were easily identified and confirmed the relevance of context-specific characters, animals and objects in distinguishing between geographical origins of folk tales.

Finally, interviews conducted with storytelling experts in Ghana highlighted the close link between the age-old West African oral storytelling tradition and its application in contemporary written literature. This knowledge shaped our computational analysis of narrative structures, which proved crucial in finding characteristic narrative patterns in the first and last sentences of West African folk tales.

This preliminary research has demonstrated that ML and NLP techniques are applicable in a wide range of tasks concerning the cross-cultural exploration of folk tales. Promising results have been achieved with regard to culture-specific text generation, classification, and narrative structure extraction in West African folk tales. The results would have been difficult or simply impossible to replicate without using ML and NLP techniques. Future work should focus on expanding and annotating the corpora, to allow for a more thorough analysis.

Acknowledgments

Throughout the writing of this thesis I have received a great deal of support and assistance. First and foremost I would like to thank my supervisor, Dr. Victor de Boer, whose in-depth knowledge and enthusiasm were invaluable in the shaping and steering of the project. I am very grateful for your consistency in supporting me and in allowing me to make this work my own while remaining critical.

My sincere thanks to Dr. Chris van Aart for giving me the opportunity to conduct my research as part of an internship at 2CoolMonkeys BV[15]. Our conversations kept me sharp and sparked my interest in business. I furthermore thank my colleagues Arjan Nusselder and Mitch Rompelman from 2CoolMonkeys BV for their collaboration and assistance. My internship at 2CoolMonkeys provided me with the consistency, tools, and social support that I needed to stay focused.

A very special gratitude goes out to the members of the W4RA team[16], especially to Anna Bon, Prof. dr. Hans Akkermans, and Wendelien Tuyp. Your personal support both during our weekly meetings and beyond have been very valuable to me for years. I thank Dr. Stefan Schlobach for being my second reader and Dr. Peter Bloem for guiding me through the world of Deep Learning.

My sincere thanks goes to the *Treub-Maatschappij, the Society for the Advancement of Research in the Tropics*[17], for awarding me the grant to conduct research in Ghana. I'd like to give special thanks to Dr. Nana Baah Gyan and Mr. Kumah Drah for setting up the interviews in Ghana, and for teaching me the ins and outs of Ghanaian storytelling and accompanying me en route. A word of thanks also goes to those who allowed me to interview them to make my thesis more complete.

Last but not least, I thank my parents, sister, and friends, for their love and support in deliberating over problems and findings, participating in the survey, and providing positive distraction.

References

1. Abbott, H.P.: The Cambridge introduction to narrative. Cambridge University Press (2008)
2. Baart, A., Bon, A., de Boer, V., Tuijp, W., Akkermans, H., Escalona, M., Domínguez Mayo, F., Majchrzak, T., Monfort, V.: Ney yibeogo-hello world: A voice service development platform to bridge the web's digital divide. In: Proceedings of the 14th International Conference on Web Information Systems and Technologies. vol. 1, pp. 23–34 (2018)
3. Barthes, R.: Le degré zéro de l'écriture. Le Seuil (2015)
4. Berry, J., Spears, R.: West African Folktales. Northwestern University Press (1991)
5. Bhardwaj, A., Di, W., Wei, J.: Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling. Packt Publishing Ltd (2018)

[15] https://2coolmonkeys.nl/

[16] https://w4ra.org/

[17] https://treub-maatschappij.org/

6. Boulis, C., Ostendorf, M.: Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. In: Proc. of the International Workshop in Feature Selection in Data Mining. pp. 9–16. Citeseer (2005)
7. Brütsch, M.: The three-act structure: Myth or magical formula? Journal of Screenwriting **6**(3), 301–326 (2015)
8. Campbell, J.: The hero with a thousand faces, vol. 17. New World Library (2008)
9. Chatman, S.B.: Story and discourse: Narrative structure in fiction and film. Cornell University Press (1980)
10. Christopher, V.: The writer's journey–mythic structure for writers (1998)
11. Chung, J., Cho, K., Bengio, Y.: A character-level decoder without explicit segmentation for neural machine translation. arXiv preprint arXiv:1603.06147 (2016)
12. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in neural information processing systems. pp. 3079–3087 (2015)
13. Dickey, M.D.: Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. Educational Technology Research and Development **54**(3), 245–263 (2006)
14. Dundes, A.: Folkloristics in the twenty-first century (afs invited presidential plenary address, 2004). The Journal of American Folklore **118**(470), 385–408 (2005)
15. Dušek, O., Jurčíček, F.: A context-aware natural language generator for dialogue systems. arXiv preprint arXiv:1608.07076 (2016)
16. Edosomwan, S., Peterson, C.M.: A history of oral and written storytelling in nigeria. Commission for International Adult Education (2016)
17. Ficler, J., Goldberg, Y.: Controlling linguistic style aspects in neural language generation. arXiv preprint arXiv:1707.02633 (2017)
18. Finlayson, M.A.: Inferring propp's functions from semantically annotated text. The Journal of American Folklore **129**(511), 55–77 (2016)
19. Finlayson, M.M.A.: Learning narrative structure from annotated folktales. Ph.D. thesis, Massachusetts Institute of Technology (2012)
20. Finnegan, R.H., Finnegan, R., Turin, M.: Oral literature in Africa, vol. 970. Oxford University Press Oxford (1970)
21. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm (1999)
22. Gervás, P.: Story generator algorithms. The Living Handbook of Narratology **19** (2012)
23. Gervás, P.: Propp's morphology of the folk tale as a grammar for generation. In: 2013 Workshop on Computational Models of Narrative. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2013)
24. Gervás, P., Lönneker-Rodman, B., Meister, J.C., Peinado, F.: Narrative models: Narratology meets artificial intelligence. In: International Conference on Language Resources and Evaluation. Satellite Workshop: Toward Computational Models of Literary Analysis. pp. 44–51 (2006)
25. Grasbon, D., Braun, N.: A morphological approach to interactive storytelling. In: Proc. CAST01, Living in Mixed Realities. Special issue of Netzspannung. org/journal, the Magazine for Media Production and Inter-media Research. pp. 337–340. Citeseer (2001)
26. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
27. Grimm, J., Grimm, W.: The Original Folk and Fairy Tales of the Brothers Grimm: the complete first edition. Princeton University Press (2014)

28. Gyasi, K.A.: Writing as translation: African literature and the challenges of translation. Research in African Literatures **30**(2), 75–87 (1999)
29. Habel, C.: Stories—an artificial intelligence perspective (?). Poetics **15**(1-2), 111–125 (1986)
30. Hochreiter, S.: The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **6**(02), 107–116 (1998)
31. Hochreiter, S., Schmidhuber, J.: Lstm can solve hard long time lag problems. In: Advances in neural information processing systems. pp. 473–479 (1997)
32. Hogenboom, F., Frasincar, F., Kaymak, U.: An overview of approaches to extract information from natural language corpora. Information Foraging Lab p. 69 (2010)
33. Hooper, C.J., Weal, M.J.: The storyspinner sculptural reader. In: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia. pp. 288–289. ACM (2005)
34. Imabuchi, S., Ogata, T.: A story generation system based on propp theory: As a mechanism in an integrated narrative generation system. In: International Conference on NLP. pp. 312–321. Springer (2012)
35. Ip, B.: Narrative structures in computer and video games: part 2: emotions, structures, and archetypes. Games and Culture **6**(3), 203–244 (2011)
36. Iyasere, S.O.: Oral tradition in the criticism of african literature. The Journal of Modern African Studies **13**(1), 107–119 (1975)
37. Johnson, D.D.: Generating polyphonic music using tied parallel networks. In: International conference on evolutionary and biologically inspired music and art. pp. 128–143. Springer (2017)
38. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
39. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
40. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
41. Lindley, C.A.: Story and narrative structures in computer games. Bushoff, Brunhild. ed (2005)
42. Lipton, Z.C., Vikram, S., McAuley, J.: Capturing meaning in product reviews with character-level generative text models. arXiv preprint arXiv:1511.03683 (2015)
43. Lucas, D.W.: Aristotle poetics (1968)
44. Lundby, K.: Digital storytelling, mediatized stories: Self-representations in new media. Peter Lang (2008)
45. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
46. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
47. Mani, I.: Computational narratology. Handbook of narratology pp. 84–92 (2014)
48. Manu, S.Y.: Six Ananse stories. Sedco Publishing Ltd. (1993)
49. Mateas, M., Stern, A.: Integrating plot, character and natural language processing in the interactive drama façade. In: Proceedings of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE-03). vol. 2 (2003)
50. Nguyen, D., Trieschnigg, D., Meder, T., Theune, M.: Automatic classification of folk narrative genres. In: Proceedings of the Workshop on Language Technology for Historical Text (s) at KONVENS 2012 (2012)

51. Ninan, O.D., Odéjobí, O.A.: Theoretical issues in the computational modelling of yorùbá narratives. In: 2013 Workshop on Computational Models of Narrative. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2013)
52. Peinado, F., Gervás, P.: Minstrel reloaded: from the magic of lisp to the formal semantics of owl. In: International Conference on Technologies for Interactive Digital Storytelling and Entertainment. pp. 93–97. Springer (2006)
53. Pelton, R.D.: The trickster in West Africa: A study of mythic irony and sacred delight. No. 8, Univ of California Press (1989)
54. Propp, V.: Morphology of the Folktale, vol. 9. University of Texas Press (2010)
55. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**, 8 (2019)
56. Sackey, E.: Oral tradition and the african novel. Modern Fiction Studies **37**(3), 389–407 (1991)
57. Simmons, D.C.: Analysis of cultural reflection in efik folktales. The Journal of American Folklore **74**(292), 126–141 (1961)
58. Tang, P.: Masters of the Sabar: Wolof griot percussionists of Senegal. Temple University Press (2007)
59. Tatar, M.: The hard facts of the Grimms' fairy tales. Princeton University Press (2003)
60. Trieschnigg, D., Hiemstra, D., Theune, M., Jong, F., Meder, T.: An exploration of language identification techniques in the dutch folktale database. In: Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012) (2012)
61. Tuwe, K.: The african oral tradition paradigm of storytelling as a methodological framework: Employment experiences for african communities in new zealand. In: African Studies Association of Australasia and the Pacific (AFSAAP) Proceedings of the 38th AFSAAP Conference: 21st Century Tensions and Transformation in Africa (2016)
62. Valls-Vargas, J.: Narrative extraction, processing and generation for interactive fiction and computer games. In: Ninth Artificial Intelligence and Interactive Digital Entertainment Conference (2013)
63. Valls-Vargas, J.: Automated narrative information extraction using non-linear pipelines. In: IJCAI. pp. 4036–4037 (2016)
64. Valls-Vargas, J., Zhu, J., Ontañón, S.: Towards automatically extracting story graphs from natural language stories. In: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence (2017)
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
66. Wang, P., Domeniconi, C.: Building semantic kernels for text classification using wikipedia. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 713–721. ACM (2008)
67. Xie, Z.: Neural text generation: A practical guide. arXiv preprint arXiv:1711.09534 (2017)
68. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044 (2015)
69. Zhang, X., LeCun, Y.: Text understanding from scratch. arXiv preprint arXiv:1502.01710 (2015)