YALE UNIVERSITY
DEPARTMENT OF LINGUISTICS

The Space of Folklore: Mapping Folkloric Texts Semantically with Document Embeddings

Benjamin Rewis

Advisor: Robert Frank

April 2020

Submitted to the faculty of the Department of Linguistics in partial fulfillment of the requirements for the degree of Bachelor of Arts

Contents

Abstract

Categorizing folklore by genre has long been a task in folkloristics, the study of folklore. We offer a computational method for categorization of folklore. We create clustered models of European folklore from an online library of folkloric texts. We use doc2vec document embeddings to semantically encode these texts. We use multiple clustering methods and a t-SNE visualization algorithm to model the texts. We compare this model with more classical categorizations of folklore. We find that this computational method of categorization is useful but inherently limited by its distributional nature.

1 Folklore

> "Myths deal with the great issues of life: the creation of the world, the nature of good and evil, and the relationships between deities and mortals. Folktales too address the nature of things... they offer explanations to life's questions, both trivial and fundamental... these stories provide a vehicle for talking about issues of concern." (Ashliman, 1987)

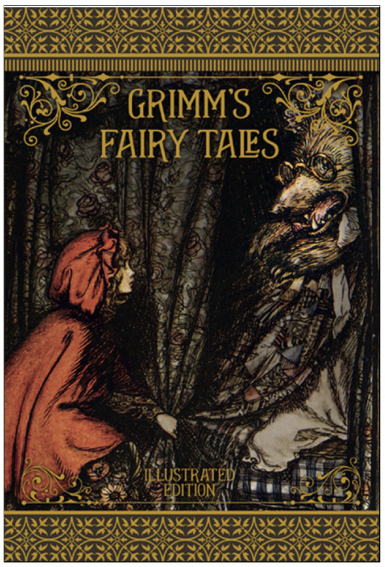
1.1 Definition and examples

Folklore is broadly defined as any information passed through generations verbally or by demonstration. This definition includes many items of cultural significance but importantly excludes most standardized material put forth by associations such as religious groups, the mass media, governments, and institutions of learning. The study and scholarly interpretation of folklore is called folkloristics. Folkloristics is a young field, as before the 19th century, studying folklore was only a part of studying human culture (and thus grouped with anthropology). After mass publication became possible, folklore became a more distinct category of study. Now, folklorists can present research in a number of journals, most popularly in the *Journal of American Folklore* published quarterly by the American Folklore Society ("Bylaws of the American Folklore Society", 2017).

Folklore, for the most part, is formally denoted as those customs that go uncodified and are thus spread by word-of-mouth more than by official publications. Folklore, arguably, is as old as language itself. [1] By connotation, most recognize fairy tales, bedtime stories, myths, and legends as forms of folklore. Riddles, rhymes, proverbs, orations, and cooking recipes are all forms of written folklore, as well. Under some definitions, clothing, painting, physical rituals, pottery, and the like are referred to as artifacts of folklore. For the purposes of this paper, we only consider forms of written folklore. Furthermore, the bulk of artifacts we examine are "stories" or written pieces with some semblance of narrative. Most could be classified by the following archetypes. Note that there is much overlap between the three.

[1] The origin of story is an interesting point of study but not the focus of this paper.

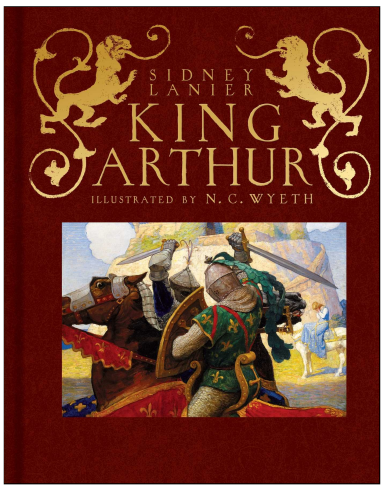
1.1.1 Fairy tales



Perhaps the most recognizable of folklore genres, the fairy tale is a relatively recent creation. *Les contes des fées* was published in 1697 by Marie Cathérine le Jumel de Barneville de la Motte (Ashliman, 2004). This collection of stories was based on French oral tradition and contained mostly works of fiction all involving the *fée* creature (fairy). As it was a compendium of oral tradition, *Les contes des fées* was a folkloric publication. The collection became so popular throughout Europe that the term "fairy tale" came into common usage and was applied to many stories of similar style. A collection published in the same year, *Contes de ma mère l'Oye* (Tales of Mother Goose), a compendium of French myths, was translated into English as *Fairy Tales, or Histories of Past Times, with Morals* (Ashliman, 2004).

Similarly, the wildly popular *Grimm's Fairy Tales*, compiled by Jacob and Wilhelm Grimm in 1812, were only labeled as "fairy tales" after being translated into English (Ashliman, 2004). In a strict sense, the term "fairy tale" does not apply to many of the pieces contained in these collections, but its connotation is so widely accepted that it seems counter-productive to re-classify these stories. Now, definitions of fairy tales vary, but fairy tales are usually recognizable as popular, fantastical works of fiction passed down as pieces of a culture's oral tradition (Ashliman, 1987).

1.1.2 Myths and legends



Myths and legends are the oldest types of folklore. Much of early oral tradition is centered around explanations of the intangible and the otherwise inexplicable. As such, these narratives are often abstract and involve god or deity

characters with inhuman power. Myths are folkloric tales describing the nature of the intangible. Common themes of myths include the origins of life, the definitions of good and evil, and life after death (Ashliman, 1987). Legends are folkloric tales that center around the often fantastical escapades of human characters. The tales of King Arthur or Alexander the Great, for example, would qualify as legends (Ashliman, 1987).

Although many myths and legends are tinted with fantasy, they all claim to be veritable. While both are probably fabrications, they This distinguishes them from folk tales, which are usually purely fictional.

1.1.3 Folktales

Folktales are the most wide-ranging of these three archetypes. Folktales are any fictional narrative that originates in a culture's oral tradition (Ashliman, 2004). Almost all fairy tales are folktales. Most are fictional or fantastic accounts with many real-world elements. They are filled with the themes of daily life and accented with magical creatures, miraculous events, and impossible feats. Further categorization of folktales will be discussed in later sections. The majority of pieces examined fall under this archetype.

1.2 Genre

Categorization is only human. It is our tendency to organize the world into neat boxes with similar items being grouped and differing items being separated. Abstractly, any artificial category of artistic composition can be called a genre. Genre is often associated with writing, and genre theory extends to cover most any kind of rhetoric. Travel brochures, lectures, and jokes belong to their own genres.

For centuries, scholars have used genres as classificatory systems for related rhetorics. The first of these scholars were Socrates and Aristotle. With great breadth of knowledge, scholars could classify many kinds of rhetoric into broad categories. Surely, comedy and tragedy were semantically distinct and belonged in separate categories. Conventional conceptions of genre separated texts based on shared characteristics. Of course, genre is never exhaustively descriptive; no set of genres could perfectly describe the set of all texts (Devitt, 2013). More recently, genre theory has grown more sensitive to the dangers of classification. In particular, top-down approaches to genre, where a handful of people describe the categories, restrict the creative process. In order to fit them into existing genre classifications, unique authors and sources can be artificially forced into boxes that don't accommodate them especially well (Devitt, 2013).

However, genre is still useful. In the case of folklore, good genre classifications can help folklorists find related material faster. They can also clarify folklorists' understanding of texts across cultures. Similarity between texts can often point to some similarity of origin and can help scholars theorize about the associated cultures.

1.3 Significance

Storytelling is a uniquely and universally human act. To the best of our knowledge, all human societies have told stories. Stories serve many purposes. They are histories, rituals, instructions, and sources of entertainment. Fairy tales, myths, legends, and folktales are typically instructions or sources of entertainment.

Human language's relation to culture is still a topic of debate. Many linguists reject the infamous Sapir-Whorf hypothesis: that the structure of a speaker's native language influences the way they see the world. It is difficult, though, to separate stories from culture. Cultures could in fact be defined as the stories of its peoples. However fantastic it may seem, a culture's folklore is a representation of its people's daily struggles and successes (Tanherlini, 1994). Stories like folk and fairy tales are snapshots of culture.

1.4 Goal of this paper

Classification of folklore is clearly useful but has often relied on tedious work by individual folklorists (Ben-Amos, 1973). We offer a computational method of folkloric classification and compare it to a popular sociological approach. This computational method makes use of document embeddings generated by distributional models of an online corpus of folkloric texts. These embeddings are visualized, clustered, and analyzed in later sections. We use this method on a dataset of folktales to make broader points about the limitations of basic document embeddings in tasks like genre classification.

2 Previous Work in Folklore Classification

> "Classification is a vexing first-order problem in folklore. Since the field's inception in the 19th century, folklorists have been concerned with the classification of texts, devising numerous classificatory systems... Genre classifications play an important role in the organization of most folklore collections." (Abello & Tangherlini, 2012)

2.1 Sociological approaches

Folklore and folkloristics appeared in the mid-to-late 19th century. William Thoms, deemed the father of modern folklore, coined the term in 1846 (Ben-Amos, 1973). Then, a Norwegian, Reidar Thoralf Christiansen, created one of the first classificatory systems of folklore in the late 1800s in his book *Migratory Legends*. He devised an eight-way typing system for Norwegian tales that is still used today for various Scandinavian stories (Ben-Amos, 1973).

In the 20th century, much work was done in refining a broader classification system for folklore (Ashliman, 2004). The most famous and widely used today is the Aarne-Thompson-Uther (ATU) index. This index has been adapted over a hundred years to encapsulate the basic themes of most Indo-European folkloric texts. First created by Antti Aarne, a Finnish sociologist, in 1910 and later adapted by both Stith Thompson and Hans-Jörg Uther in 1961 and 2004 respectively, the ATU index features seven broad categories (Ashliman, 2004).

1. **Animal tales**
   These are tales featuring animals as main characters. Fables are often didactic in nature and serve to teach or present moral concepts to the readers. *Aesop's Fables* almost all fall under this category. Not all animal tales are fables, though.

2. **Tales of magic (fairy tales)**
   Fairy tales are arguably the most well-known of the seven categories. These tales are marked by high fantasy, magic, and the supernatural. Many feature a similar structure. The main character often leaves home, faces conflict, and after overcoming the conflict, returns home.

3. **Religious tales**
   These tales are similar to tales of magic, but usually involve a deity such as God or the devil. They deal with religious themes of faith, miracles, conversion, and redemption.

4. **Realistic tales**
   These tales are also called *novelle* or romantic tales. They are frequently about love, princes, or princesses. Despite being "realistic," they sometimes have magical elements. Although usually, they are based solely in the real world.

5. **Tales of the stupid ogre**
   Despite their name, these tales do not all feature ogres. Rather, each features a large monster described in varying ways. The monsters are sometimes even mean, large humans.

6. **Anecdotes and jokes**
   These are brief narratives that end with some "punch-line" or unexpected conclusion.

7. **Formula tales**
   Formula tales are characterized by their form rather than their content. These narratives often have self-contradiction, clever wordplay, and repetitive structures.

The ATU typing system is preferred over other systems because of its hierarchical structure (Ashliman, 2004). For the remainder of this paper, the above, broad categories will be called "metatypes." Each of the seven metatypes are split into three to eight groups that will be called "supertypes." There are 35 supertypes all together.

Metatype	Supertype 1	Supertype 2	Supertype 3	Supertype 4	Supertype 5	Supertype 6	Supertype 7	Supertype 8
Animal tales	wild animals	humans	domestic animals	----------	----------	----------	----------	----------
Tales of magic	ghost	possessed	supernatural task	supernatural helper	magic object	power	----------	----------
Religious tales	god rewards	truth comes to light	heaven	the devil	----------	----------	----------	----------
Realistic tales	princess	prince	fidelity and innocence	obstinate wife	good precepts	clever acts	tales of fate	robbers
Tales of the stupid ogre	labor	partnership	contest	killing	frightening	outsmarted	saved	----------
Anecdotes and jokes	fool	married couple	woman	man	clergy	----------	----------	----------
Formula tales	cumulative	catch	----------	----------	----------	----------	----------	----------

Within these 35 supertypes, there are hundreds of sub-categories defining a space of around 2000 "subtypes." Subtypes become quite specific. 510A, for example, refers to "Cinderella" tales, while 445 refers to "Puss-in-Boots" tales. The letters following certain subtypes refer to even smaller variations. 510B, for example, refers to "Donkey-Skin" tales, an older version of "Cinderella" that is highly similar but not identical. The great majority of subtypes do not have these smaller variations, so they are ignored.

Subtypes can be seen as different versions of the same story; these versions often come from different parts of the world. Supertypes represent groups of stories with similar characters, motifs, and narratives. Metatypes represent groups of stories with the same broad themes. Supertypes and metatypes are defined as ranges over subtypes. For example, subtypes 1-100 belong to the supertype "wild animals," and subtypes 1-300 belong to the metatype "animal tales (fables)." Keeping in mind the limitations of genre previously mentioned, the ATU index is a rather comprehensive classification system.

2.2 Computational approaches

Computational systems have only recently entered the field of folkloristics. Using these systems, folklorists have been able to computationally induce "genre" in sets of folkloric texts. An ideal categorization system would split groups of documents into equivalent, specific, and helpful parts. Abello, Broadwell, and Tangherlini propose a method to classify a relatively small corpus of Danish ghost stories (Abello & Tangherlini, 2012). The underlying linguistic assumption in this categorization task is that the language of any folkloric text is some manifestation of a category. A text will likely portray some characteristics of its hidden category. A riddle, for example, would likely be a short explanation of context followed by some related question.

Abello, Tangherlini, and Broadwell attempt to break down a collection of ghost stories into a "hypergraph" or "story space." The authors use the term hypergraph to refer to a representation of high-dimensional space. Stories are points in this high-dimensional space and are represented as attribute vectors. Each attribute vector is filled with numbers that represent features of that tale. These features include frequencies of keywords across the corpus and presence of other features defined in Tangherlini's work on Danish folklore (Tanherlini, 1994). These attribute vectors allow the authors to view clusters which they then retrospectively associate with aspects of the contained stories. The corpus of ghost stories clustered loosely based on inherent "threat" in the story.[2] The authors mention that the key to a convincing study in computational folkloristics is specificity of problem. Indeed, this paper examines only one corpus, and its clustering is quite convincing. They manage to separate ghost stories based on the inherent "threat" in the narrative (supernatural, robbery/theft/murder, witches/animals, etc.). Although this categorization may not seem useful, its implications are. (Abello & Tangherlini, 2012) showed that folkloric corpora could be categorized computationally. In the past decade, computational linguists have taken an interest in folkloric corpora.

Dong, Trieschnigg, and Theune use a combination of subject-verb-object triplets and information retrieval techniques to build a simple supervised folktale classifier (Nguyen & Theune, 2013). This classifier is trained on a corpus of 400 stories labeled by both ATU type and Brunvand urban legend type.[3] The classifier returns a "ranking," or ordered list of potential types for a given input story. They perform significantly better than baselines.

Karsdorp and Bosch use latent dirichlet allocation (LDA) to identify motifs in folktales (Karsdorp, 2013). In particular, they use labeled LDA to create topic models, or models that find the "hidden" categories of a dataset. These hidden categories end up being the symbols that appear throughout folktales, which folklorists often call

[2]These threats include paranormal, criminal, satanic, economic ones, etc.

[3]Brunvand creates a typing system for modern, American urban legends in (Brunvand, 2002).

"motifs." (Ben-Amos, 1973). (Thompson, 1960) describes a set of motifs for many popular tale types. These include "glass slippers," "wolf," "animal swallows man," "giant mermaid appears," and many more. (Karsdorp, 2013) uses LDA to generate one-word themes of folktales and compares these themes qualitatively to the motifs assigned to each tale. In stories with motifs like "lost ring found in fish," LDA picks out to the top themes (in order) as "fish," "ring," and "sea."

Tangherlini uses sub-corpus topic modeling to find key passages in large corpora of folktales that would be difficult to scan manually (Tangherlini, 2013). He later uses similar concepts to create a geographic model of the same corpus of Danish ghost stories used in (Abello & Tangherlini, 2012). This model shows the author of each tale at the origin of a hypergraph and supernatural threats as vectors based on their geographic source (Kenna Ralph, 2017).

Tehrani and d'Huy even use techniques from bio-informatics, the use of computational tools to study biological processes, to induce phylogenetic trees of folkloric texts (Kenna Ralph, 2017). The authors are able to correlate tales across many societies and eras using motifs defined in (Thompson, 1960). These motifs are defined as "characteristics" of tales, and the authors use retention index, a measurement of how well characteristics are distributed across a phylogenetic tree, combined with two other similar measures to create believable evolutionary paths of folktales (Kenna Ralph, 2017).

3 Meaning: The Distributional Hypothesis

> "Words which are synonyms (like oculist and eye-doctor) tended to occur in the same environment (e.g., near words like eye or examined) with the amount of meaning difference between two words corresponding roughly to the amount of difference in their environments." (Jurafsky, 2019)

3.1 Vector-based semantics

In classical semantics, the meaning of a sentence is intuitively compositional; words have individual meanings that combine together to make larger meanings. Then, a declaration may have some truth value in reference to the real world, or a question may be a set of propositions. Exclamatives, embedded structures, elided constructions, and the like propose complications that many semanticists spend their lives examining.

Much of classical semantics, though, is based on individual units of meaning. Whether we call these words, lexemes, or morphemes does not really matter. There are atoms of meaning that are in a way irreducible. "Book," for example, is not made up of smaller semantic units. Classical semanticists would simply define "book" as the set of all entities that are books. This concept is easy enough for humans to understand. However, these irreducible atoms of meanings do not come naturally to a computer. A computer has no knowledge of what a "book" is.

Computational semantics tries to rebuild the concept of meaning computationally. Because computers have no knowledge of the real world, meaning must be tied to something else. In the early 1950s, linguists such as Martin Joos, Zellig Harris, and J.R. Firth separately described a theory that would later be called the distributional hypothesis (Jurafsky, 2019). Under this hypothesis, meaning is described comparatively: each word is defined by its distributional distance from other words. Words that appear in similar environments have similar meanings.

Using the distributional hypothesis, Charles Osgood presented a new, objective method for semantic differential measurement in the late 1950s (Jurafsky, 2019). Osgood claimed that the meaning of a word could be modeled as a point in a multidimensional Euclidean space and that semantic similarity of meaning of words could be equated to distance in that space (Osgood & Tannenbaum, 1957). He encoded words by assigning them values on scales such as *happy/sad* or *hard/soft*. The scale-values of a word were strung together to create a vector, or a list of numbers. If N was the number of scales, this vector was representative of the word's semantic value in an N-dimensional space (Osgood & Tannenbaum, 1957). Today, computational linguists refer to these vector-encodings of words as "word embeddings." The meaning-distance between two words was the distance between their embeddings.

The distance between two vectors can be calculated in a number of ways. Euclidean distance, the "ordinary" straight-line distance between vectors, and cosine similarity, the angle between vectors, are the most popular. Given vectors p and q each of size n Euclidean distance is defined as:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

Cosine similarity is the inner product of two vectors (p and q) divided by the product of the vectors' magnitudes:

$$d(p,q) = \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\|\|\mathbf{q}\|} \tag{2}$$

Osgood's mathematical elaboration upon the distributional hypothesis led to many separate advances in the field of vector semantics and natural language processing (NLP) (Jurafsky, 2019). In the early 1960s, Researchers discovered that vector semantics could help with automatic information retrieval, or the obtaining of relevant information from a collection of texts. In building a vector space model for information retrieval, these researchers refined methods of measuring vector similarity (Jurafsky, 2019). In particular, they used co-occurrence matrices to encode terms instead of the scales Osgood had used.

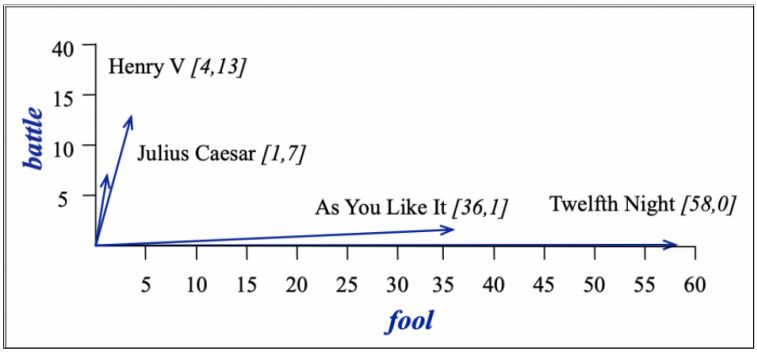
Co-occurrence matrices represent how often words in a corpus (a group of texts) co-occur. If the size of your vocabulary (the set of unique words in your corpus) is N, the matrix would be of size NXN. Each row-column pair would represent the frequency of co-occurrence between the two words. For example, if your corpus only contained the sentences "Roses are red. Violets are blue.", your co-occurrence matrix might look like:

	Roses	are	red	Violets	blue
Roses	1	1	1		
are	1	2	1		
red	1	1	1		
Violets		1		1	1
blue		1		1	1

Above, terms co-occur when they are present in the same sentence. This definition of co-occurrence varies. It can also refer to when two terms occur next to each other in the same sentence (or a few words apart). Obviously, some terms like articles or common verbs appear much more frequently than other terms. At the time, many researchers weighted the terms in the co-occurrence matrices based on how often they appeared in individual documents and in the corpus (Jurafsky, 2019). Term-frequency-inverse-document-frequency (TF-IDF) was a popular method of weighting. A term's TF-IDF is a numerical statistic representative of how "important" that term is in the corpus.[4] The goal of TF-IDF is to normalize values of common and uncommon words, making sure frequent but unimportant words do not have abnormally large values in their embeddings.

Just as words can be represented with co-occurrence matrices, documents can be represented with term-document matrices. If the size of your vocabulary is N, and the number of documents in your corpus is D, a term-document matrix would be of size NXD. Each row-column pair represents the frequency of occurrence of a term in a document. Each column is a semantic encoding of a document, or a "document embedding."

If we examined the counts of two words, "battle" and "fool," in four Shakespeare documents: *Henry V*, *Julius Caesar*, *As You Like It*, and *Twelfth Night*, we could display their embeddings in a two-dimensional space.



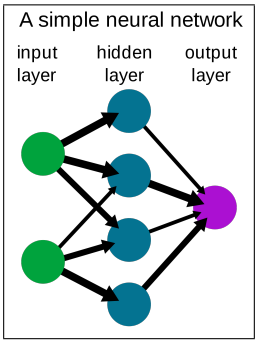
[5]

[4] We do not include the formal denotation of TF-IDF, as it becomes irrelevant as new methods of vector generation arise.
[5] (Jurafsky, 2019)

The meaning difference between entire documents (in this very basic example) is just the distance between the two related vectors. So, with the distributional hypothesis early computational linguists showed that both words and documents could be embedded in multidimensional space. However, these researchers required large vectors built from huge co-occurrence or term-document matrices to accurately represent semantic values. Today, computational linguists refer to these larger vectors as "sparse," as they are large and filled with mostly zeros. It was not until the late 1980s that linguists created algorithms for "dense," or relatively small, mostly non-zero vectors (Jurafsky, 2019).

The first of these dense vectors used techniques for dimensionality reduction, or finding the most important values in the sparse vectors to include in the dense ones. Singular value decomposition (SVD) was the first algorithm used.[6] Latent semantic analysis (LSA) models, use SVD on term-document matrices to find a list of the most important terms across a corpus. The counts of these terms in documents could be used to create dense and relatively semantically accurate vectors representing those documents (Deerwester, Dumais, Furnas, & Harshman, 1990). Linguists also used SVD on co-occurrence matrices to create dense word embeddings (Schütze, 1992). For 20 years, these LSA models and their dense vectors were applied to many NLP tasks. Some variations on LSA were created but did not significantly outperform it (Jurafsky, 2019).

In the early 2000s, computational linguists began to use neural language models to create embeddings. (Mikolov, Chen, & Dean, 2013) used simple neural networks to create methods of dense vector generation. These algorithms were known as word2vec (Mikolov et al., 2013). Artificial neural networks (ANNs) are vaguely inspired by neural networks in the human brain. An ANN is a collection of connected units, sometimes called artificial neurons. Connections between units can transmit directed signals from one unit to another. In this way, artificial neurons can process input and output some number (signal) to another unit. Neurons are typically organized into layers, with the first layer representing input to the network, and the last layer representing output from the network. Intermediary layers are often called "hidden layers."

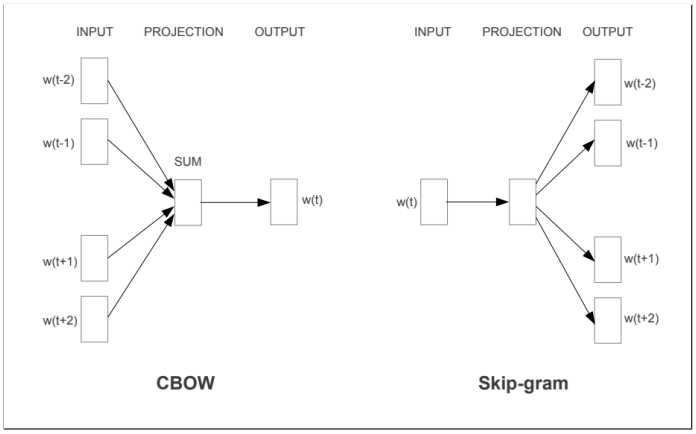


ANNs can learn how to do simple tasks through training. Connections between units, shown as arrows in the above diagram, each have their own weight in the network. A larger weight, represented by a larger real number, means a larger impact on the output of the network. An ANN can be trained with a set of inputs and a corresponding set of correct outputs. During training, weights are adjusted to increase accuracy of the output.[7]

word2vec models are predictive models that are trained to guess how likely one word is to appear in a context of other words using simple neural networks (Jurafsky, 2019). After training, the weights learned by the neural network become the embeddings for each word. (Mikolov et al., 2013) provides two word2vec algorithms called continuous-bag-of-words (CBOW) and skip-gram. CBOW trains a neural network to predict a word based on its context. skip-gram does the opposite; it trains a neural network to predict a context given a word (Mikolov et al., 2013).

[6]Once again, the exact mathematics of SVD are somewhat irrelevant.

[7]These adjustments are called backpropagation and usually use methods like stochastic gradient descent.



Formally, CBOW learns to predict a single word from a context of words by maximizing average log probability (Mikolov et al., 2013). Given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, a CBOW word-vector model maximizes:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, \ldots, w_{t+k}) \tag{3}$$

skip-gram learns to predict a context of words from a single target word by maximizing the log probability of any context word in a small window given the target word. Given a target word w_{target} and window size C, a skip-gram word-vector model maximizes:

$$\log p(w_1, w_2, \ldots, w_C|w_{target}) = \prod_{c=1}^{C} \log p(w_c|w_{target}) \tag{4}$$

The semantic properties of these word2vec-generated vectors are remarkable. In particular, these embeddings capture relational meanings quite well. (Mikolov et al., 2013) shows that the differences between word embeddings often show analogical relations between words.

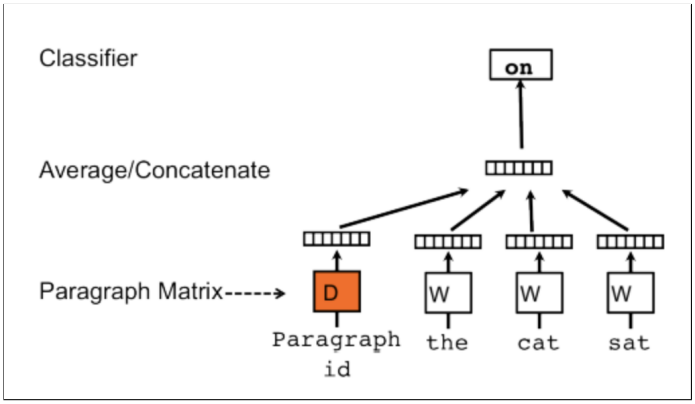
1. $vec(\text{“}king\text{”}) - vec(\text{“}man\text{”}) + vec(\text{“}woman\text{”}) \approx vec(\text{“}queen\text{”})$
2. $vec(\text{“}Paris\text{”}) - vec(\text{“}France\text{”}) + vec(\text{“}Italy\text{”}) \approx vec(\text{“}Rome\text{”})$

These “offsets,” or components of vectors, clearly represent components of meaning in the vector space. So, we can apply algebraic methods to create a sort of semantic algebra. However, we want to represent entire documents as points in some multidimensional space, so we will need an algorithm to find dense document embeddings. We could simply add all the word embeddings present in a document together to create a document embedding, but there turns out to be a better solution.

3.2 doc2vec

(Le, 2014) presented doc2vec, an algorithm based off the same researchers' work on word2vec just a year earlier. While word embeddings are helpful in many NLP tasks, accurate semantic encodings of documents can help with document retrieval, web search, spam filtering, and much more. doc2vec creates dense documents embeddings and is similar to word2vec in design. One vector is added to the network: an encoding of the identity of the document that is used to perform word prediction (as in the word2vec) model (Le, 2014). As with the word embeddings, this document vector is trained to maximize word prediction accuracy.

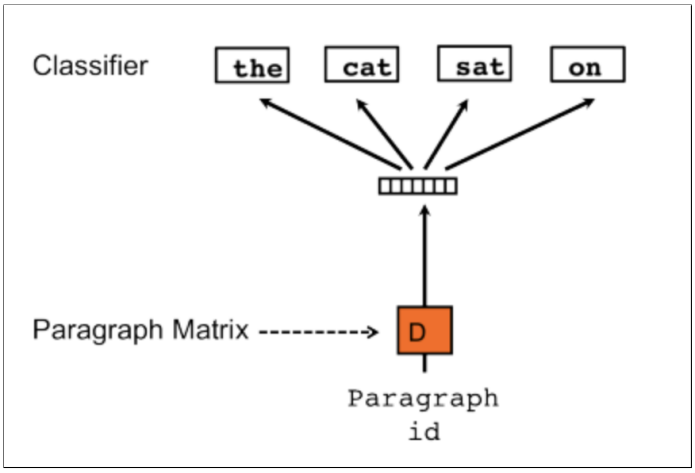
Similar to word2vec, doc2vec has two variants. The CBOW model of word2vec became “distributed memory version of paragraph (document) vector” or PV-DM in doc2vec.



[8]

The word vectors in matrix W are trained in parallel with the document vectors in D. After training, each row in D holds a vector representation of the document or a document embedding. Formally, document ids are treated just like another word vector in the model.

The skip gram model of word2vec became "distributed bag of words version of paragraph vector" or PV-DBOW in doc2vec.



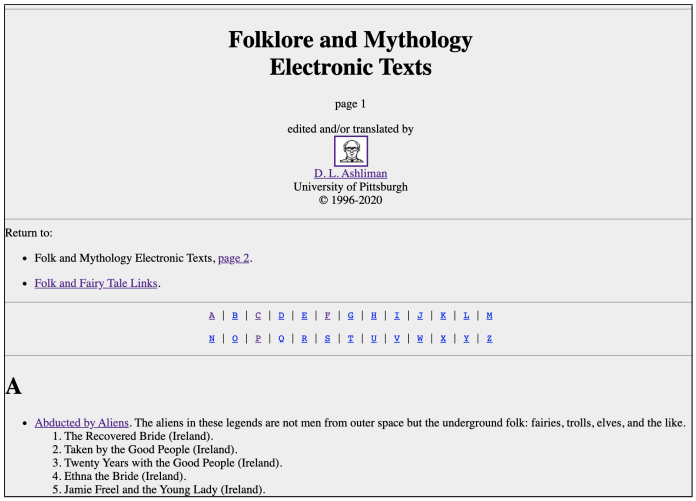
[9]

As opposed to skip-gram, PV-DBOW uses only the relevant document vector to predict the context instead of a central word. Still, this document vector is trained in PV-DBOW just like a word vector in skip-gram. We use a PV-DM model to create document embeddings for folktales. Hyperparameters and training processes are explained in later sections.

[8](Shperper, n.d.)
[9](Shperper, n.d.)

4 Methods

4.1 *Folktexts*: a library of folktales



Folklore and Mythology
Electronic Texts

page 1

edited and/or translated by

D. L. Ashliman
University of Pittsburgh
© 1996-2020

Return to:

- Folk and Mythology Electronic Texts, page 2.
- Folk and Fairy Tale Links.

A | B | C | D | E | F | G | H | I | J | K | L | M
N | O | P | Q | R | S | T | U | V | W | X | Y | Z

A

- Abducted by Aliens. The aliens in these legends are not men from outer space but the underground folk: fairies, trolls, elves, and the like.
  1. The Recovered Bride (Ireland).
  2. Taken by the Good People (Ireland).
  3. Twenty Years with the Good People (Ireland).
  4. Ethna the Bride (Ireland).
  5. Jamie Freel and the Young Lady (Ireland).

[10]

A seemingly bare-bones HTML site, *Folktexts* is one of the most expansive and organized collections of online folklore to date. It is perhaps the most respected online scholarly source for folklore researchers. It launched in 1996 and is still being updated. The website contains over two thousand tales, manually gathered by prominent folklorist D.L. Ashliman and categorized by theme. In particular, the majority of tales are organized by ATU index, the most popular method of classification of Indo-European folk and fairy tales. Tales are all in English and are translated from a wide variety of languages.

Using a combination of web-scraping [11] and manual entry, we transferred the majority of tales on the site to CSV (comma-separated-value) files that could be better interpreted by Python programs. While there are over two thousand tales on *Folktexts*, only 1653 are categorized by ATU type. These tales are representative of 121 subtypes, 32 supertypes, and seven metatypes. There are anywhere from one to 27 tales in each subtype on *Folktexts*. Length of tale varies greatly. The shortest tales are a couple sentences (around 50 words), and the longest can be multiple pages (around 2000 words). Tales within the same subtype are usually of similar length.

Folktexts is a wonderful resource, but contains almost no folklore from South America, Africa, or Australia. Furthermore, the majority of tales are magic tales, animal tales, or jokes. The other four metatypes are underrepresented. Despite the imbalance in content, the quality of translations and faithfulness of tales are considered very good by most folklorists. While it is not representative of the entire world's folklore, *Folktexts* is a useful corpus for examining the power of document embeddings in a small dataset.

4.2 doc2vec for document embeddings

We used Gensim's library for scalable statistical semantics to run a PV-DM doc2vec model on the gathered folktales (Řehůřek, 2010). While (Le, 2014) recommends a combination of both PV-DM and PV-DBOW for document representation, PV-DM is empirically superior and much faster, so we used a PV-DM model to create the embeddings. Additionally, doc2vec allows you to choose the size of the vector for each embedding. Given around 1000 documents, vectors of size 1000 could uniquely represent each document, and there would not necessarily be any similarity between documents. If vectors were only of size five or ten, though, embeddings would not be able to represent the full range of variability of their documents, and texts would cluster together too tightly. After trying vectors of size 5-100, we found that those of around size 45 had the highest performance (as defined by purity in a later section). We also used doc2vec's concurrency features to train across four cores.

[10] From *https://www.pitt.edu/dash/folktexts.html*.
[11] We used Python's BeautifulSoup library for screen-scraping.

4.3 Clustering algorithms

Once doc2vec generated the vectors, we needed an algorithm to compare each document embedding and generate sets of similar documents. These "clusters" of documents would hopefully be representative of some inherent semantic similarity between tales. There are a few different measurements of similarity between vectors and many different algorithms for clustering. The two most popular measurements of similarity between vectors are cosine similarity and Euclidean (L^2) distance. Cosine similarity is equivalent to the angle between the two vectors. Euclidean distance is the magnitude of the vector that would connect the two vectors in N-dimensional space.[12]

Euclidean distance is sensitive to the magnitudes of vectors while cosine similarity is not. Cosine similarity is better used when vector magnitude is irrelevant. Unweighted vectors generated from co-occurrence matrices, for example, greatly vary in magnitude and are better measured with cosine similarity. Euclidean distance is often used when vectors are normalized and their magnitudes are relevant as is our case. We used two different methods for clustering. Both of the methods below use Euclidean distance as a measure of similarity because doc2vec vectors are normalized and should therefore be compared with Euclidean distance.

4.3.1 K-means clustering

K-means attempts to cluster n data points into k clusters. It does this by iteratively defining k "means," or centers of clusters, for the data points. K-means consists of two simple steps: assignment and update. In the assignment step, each data point is associated with the nearest cluster. This is the cluster with the least squared Euclidean distance from its mean to the point. In the update step, the means of each cluster are recalculated given the new assignments by taking the mean of the points assigned to that cluster.

Gradually, members and means of clusters converge, and data points are assigned to clusters. In this case, k-means would be run on relevant document embeddings and k would be equal to the true number of types present in those embeddings. For example, if k-means was run on the entire dataset to cluster by metatype, n would be 1653 (the number of tales) and k would be 7. Instead of randomly selecting initial cluster centers, we used k-means++ to select them. k-means++ spreads out the k initial cluster centers uniformly across the data points to cluster more efficiently.

4.3.2 Agglomerative clustering

Agglomerative clustering is a type of hierarchical cluster analysis (HCA). HCA attempts to build a hierarchy of clusters based on similarity between either increasing or decreasing subsets of data points. Agglomerative clustering is a "bottom-up" approach to HCA, where each observation starts in its own cluster and is merged iteratively with others. At each iteration, clusters with the least dissimilarity are combined. At first, dissimilarity is a measure of distance between points but then becomes a distance between sets of points.

We used Euclidean distance to measure distance between points. We also used ward linkage to evaluate similarity between sets of vectors. Ward linkage seeks to combine the two clusters that will minimize the total within-cluster variance.

4.4 Vector visualization

While clusters can be calculated with clustering algorithms, displaying the vector space itself requires a separate algorithm. When documents are turned into their embeddings, each document becomes a point in an N-dimensional space, where N is the length of embeddings. In our case, N is 45 and is far too many dimensions to represent on a graph. To visualize this high-dimensional data, we must perform dimensionality reduction, or the selection of principal variables of variation.

Two very popular techniques for dimensionality reduction are t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA). Empirically, we found the t-SNE visualizations to spread the data out further and make better visualizations. We chose t-SNE. Visualization algorithms are often an aesthetic choice more than anything; they have nothing to do with the clustering algorithms described above.

Essentially, t-SNE finds similarity of points. Technically, it calculates the probability of similarity of points in the high-dimensional space and then in the corresponding low-dimensional space (two dimensions in our case). This similarity is the conditional probability that a point A would choose point B as its neighbor if neighbors were picked

[12] Formulas given in section 3.1.

in proportion to their probability density under a normal distribution centered at A. It then minimizes the difference between these similarities in higher and lower-dimensional space for a representation of data points in the lower dimension. To measure the minimization of the sum of differences of conditional probability, t-SNE minimizes the sum of Kullback-Leibler divergence of overall data points using gradient descent. Kullback-Leibler divergence is just a measure of how one probability distribution differs from another.

So, t-SNE minimizes the KL divergence between two distributions: a distribution that measures similarities of the input and a distribution that measures similarities of the output. We used t-SNE and Seaborn, a data visualization library based on matplotlib, to create many of the graphs shown in section 5.

4.5 Measures of purity

Once the clusters were generated, we needed an accurate measure of the performance of the two clustering algorithms with respect to the original typings of the tales. There are quite a few methods of external cluster evaluation, but we chose four that seemed to represent performance well (Rosenberg, 2007).

The following rely on measures of entropy, H. These measures of entropy involve C, the set of classes, and K, the set of clusters (Rosenberg, 2007). If there are N data points, and a_{ij} is the number of data points that are members of class c_i and elements of cluster k_j:

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{N} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{N} \tag{5}$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{N} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{N} \tag{6}$$

$$H(C|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \tag{7}$$

$$H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}} \tag{8}$$

1. **Homogeneity:**
   Homogeneity is the measure of the extent to which clusters contain only data points which are members of a single class:
$$h = \begin{cases} 1 & \textit{if } H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & else \end{cases} \tag{9}$$

2. **Completeness:**
   Completeness is the measure of the extent to which data points of a given class are elements of the same cluster:
$$c = \begin{cases} 1 & \textit{if } H(K,C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & else \end{cases} \tag{10}$$

3. **V-measure:**
   V-measure is an "average" of homogeneity and completeness. It is the harmonic mean of the two measures. If β is the ratio of weight attributed to homogeneity, h, or completeness, c, v-measure is:
$$\frac{(1+\beta)hc}{\beta h + c} \tag{11}$$

4. **Base purity:**
   Base purity is a measure of the extent to which clusters contain a single class:
$$\frac{1}{N} \sum_{k \in K} \max_{c \in C} |k \cap c| \tag{12}$$

Base purity is a very simple measurement, but its simplicity helps give an easily interpretable indication of performance. V-measure is slightly more complicated but is built to tolerate randomness; v-measure is significantly lower than base purity for random labelings (Rosenberg, 2007). We use a β of 1.0 to weight homogeneity and completeness evenly.

Performance is defined as the improvement of these metrics from a random clustering of the document embeddings. This random clustering gave a random label in the set of relevant types to each embedding based on the distribution of actual ATU types in the set of tales. For example, if a clustering was created for 10 stories of type 510A, and 10 stories of 445, the random clustering assigned 510A and 445 with equal frequency.

5 Results

Evaluations have been split into subtypes, supertypes, and metatypes. For each category, we examine a few comparisons between well-known ATU types, providing visualizations for each. In these comparisons, we show a t-SNE-derived graph with highlighting based on true labels, random labels, k-means clusters, and agglomerative clusters. We then examine performance on the entire dataset. Finally we rank the purity of clusters for each category.

Purity by cluster is a measurement of the performance of each ATU subtype, supertype, or metatype. Types with high purity are associated with a cluster with a high F1 score. The F1 score, or F-measure, is the ratio of correct guesses to incorrect guesses. The F-measure is also the harmonic mean of the precision and recall for that cluster. Some examples of the text of each ATU type are provided at the beginning of each section.

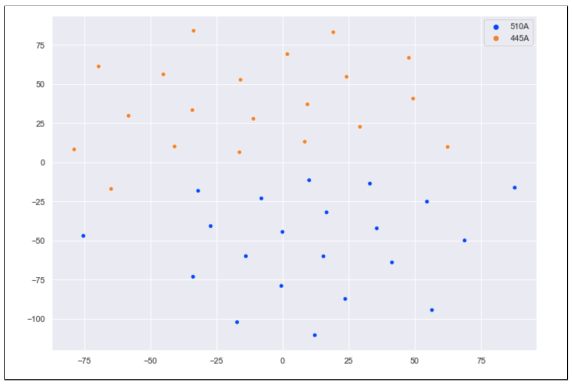
5.1 Subtype clustering

5.1.1 ATU 510A: “Cinderella” vs. ATU 445: “Puss-In-Boots”

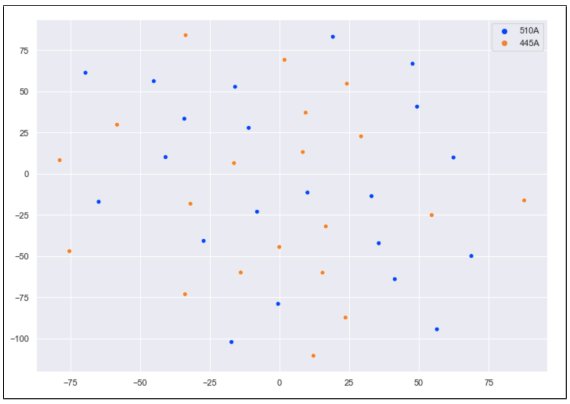
”However long she had suffered in ashes and sorrow, Cinderella was now living in splendor and joy. As midnight approached, before the clock struck twelve, she stood up, bowed, and said that she had to go, in spite of the prince’s requests for her to stay. The prince escorted her out. Her carriage stood there waiting for her. And she rode away just as splendidly as she had come.” (Grimm, 1812)

“The next day, just as he said he would, the cat, appropriately booted, went hunting again and took the captured game to the king. Thus it continued every day, and every day the cat returned home with more gold. He was now so favored by the king that he was allowed to come and go as he pleased and to prowl around the palace wherever he wanted to.” (Grimm, 1812)

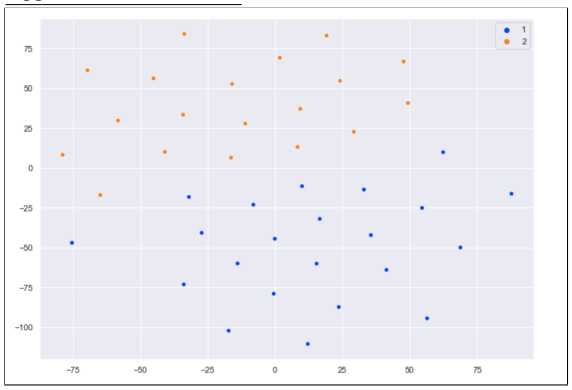
True Labels:



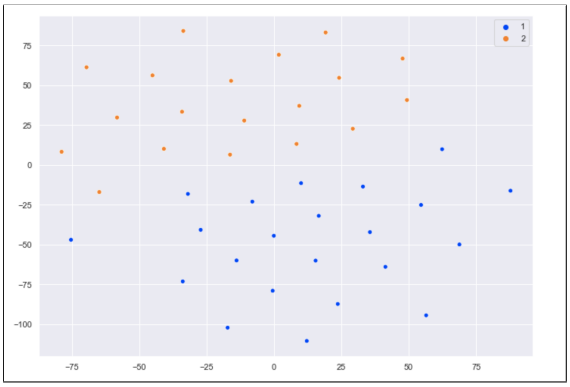
Random Labels:



Agglomerative Clusters:



K-Means Clusters:



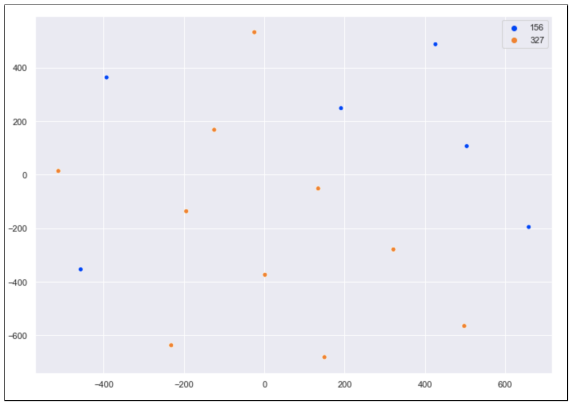
	Base purity	Completeness	Homogeneity	V-measure
True labels	1	1	1	1
Random labels	.537	.004	.004	.004
K-means clusters	.976	.858	.857	.858
Agglomerative clusters	.951	.712	.712	.712

5.1.2 ATU 156: “Androcles and the Lion” vs. ATU 327: “Hansel and Gretel”

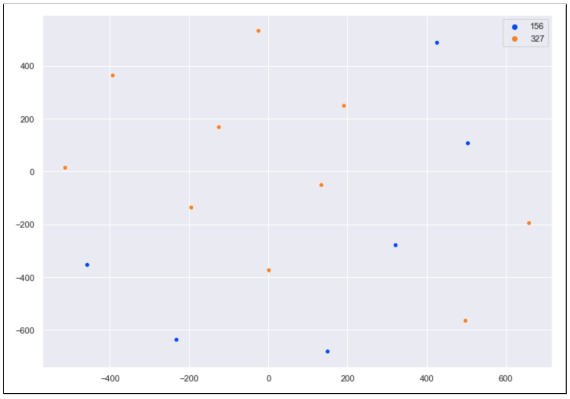
”As he came near, the lion put out his paw, which was all swollen and bleeding, and Androcles found that a huge thorn had got into it, and was causing all the pain. He pulled out the thorn and bound up the paw of the lion, who was soon able to rise and lick the hand of Androcles like a dog. Then the lion took Androcles to his cave, and every day used to bring him meat from which to live.” (Aesop)

“Suddenly the door opened, and a woman, as old as the hills and leaning on a crutch, came creeping out. Hansel and Gretel were so frightened that they dropped what they were holding in their hands. But the old woman shook her head and said, “Oh, you dear children, who brought you here? Just come in and stay with me. No harm will come to you.” She took them by the hand and led them into her house.” (Grimm 1812)

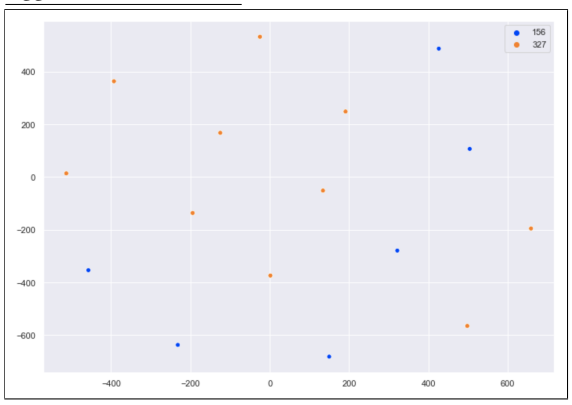
True Labels:



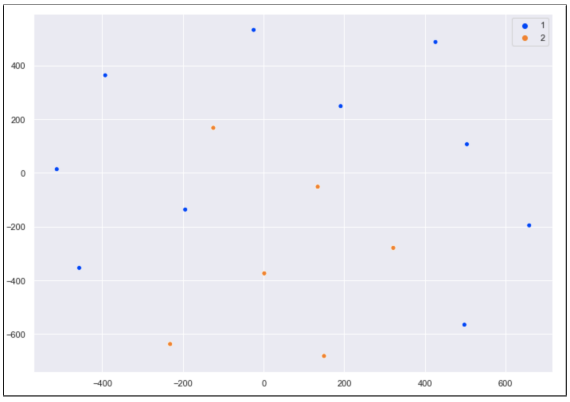
Random Labels:



Agglomerative Clusters:



K-Means Clusters:

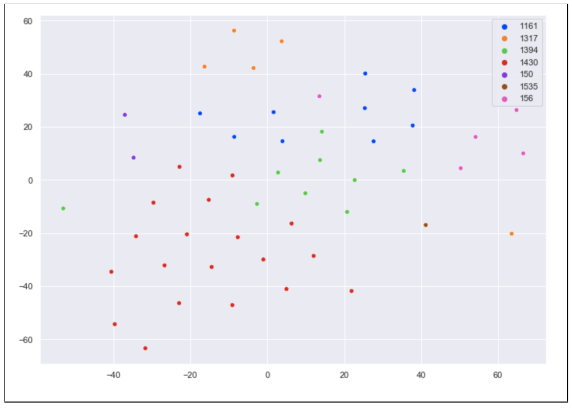


	Base purity	Completeness	Homogeneity	V-measure
True labels	1	1	1	1
Random labels	.625	.030	.030	.030
K-means clusters	.75	.364	.364	.364
Agglomerative clusters	.688	.302	.284	.293

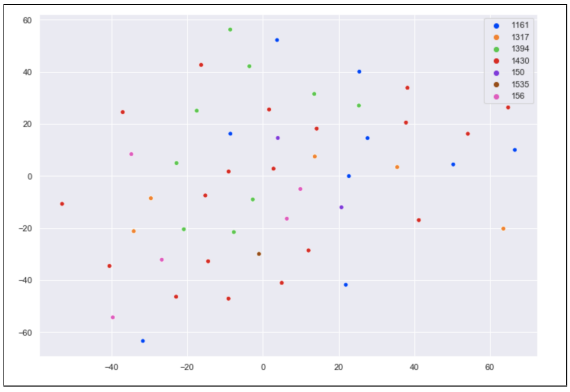
5.1.3 Seven popular tales

1. 1161: The Bear Trainer and His Cat
2. 1317: The Blind Men and the Elephant
3. 1394: Stories of Hairless Men
4. 1430: Air Castles
5. 150: Captured Birds
6. 1535: Big Peter and Little Peter
7. 156: The Lion's Paw

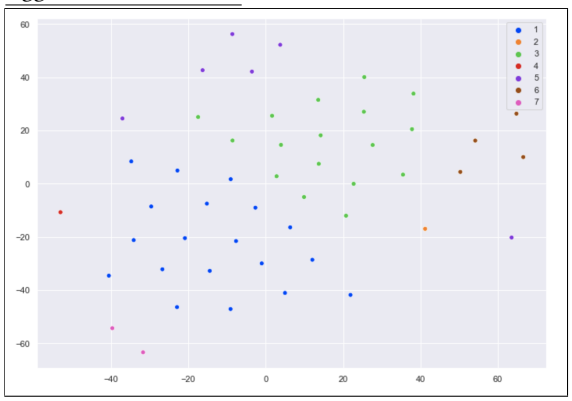
True Labels:



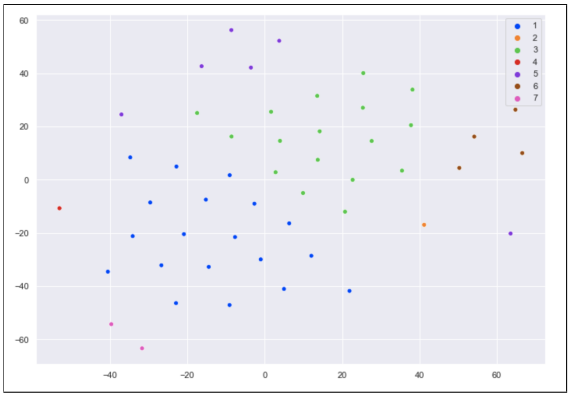
Random Labels:



Agglomerative Clusters:



K-Means Clusters:

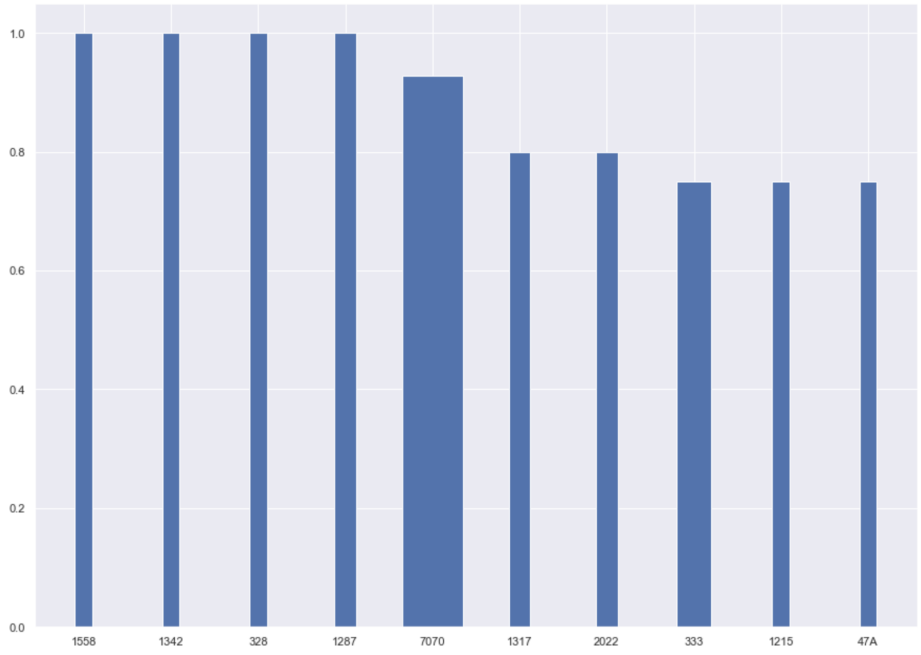


	Base purity	Completeness	Homogeneity	V-measure
True labels	1	1	1	1
Random labels	.420	.151	.151	.151
K-means clusters	.780	.777	.694	.734
Agglomerative clusters	.791	.789	.712	.755

5.1.4 All subtypes

	Base purity	Completeness	Homogeneity	V-measure
True labels	1	1	1	1
Random labels	.163	.482	.482	.482
K-means clusters	.337	.608	.564	.585
Agglomerative clusters	.379	.631	.594	.612

5.1.5 Purity by subtype



[13]

1. 1558: Clothes Make the Man
2. 1342: Hot and Cold with the Same Breath
3. 328: Jack and the Beanstalk
4. 1287: Fools Cannot Count Themselves
5. 7070: Sunken Bells
6. 1317: The Blind Men and the Elephant
7. 2022: The Lion's Paw
8. 333: Cattarinetta
9. 1215: The Man, the Boy, and the Donkey
10. 47A: Catching a Horse by Its Tail

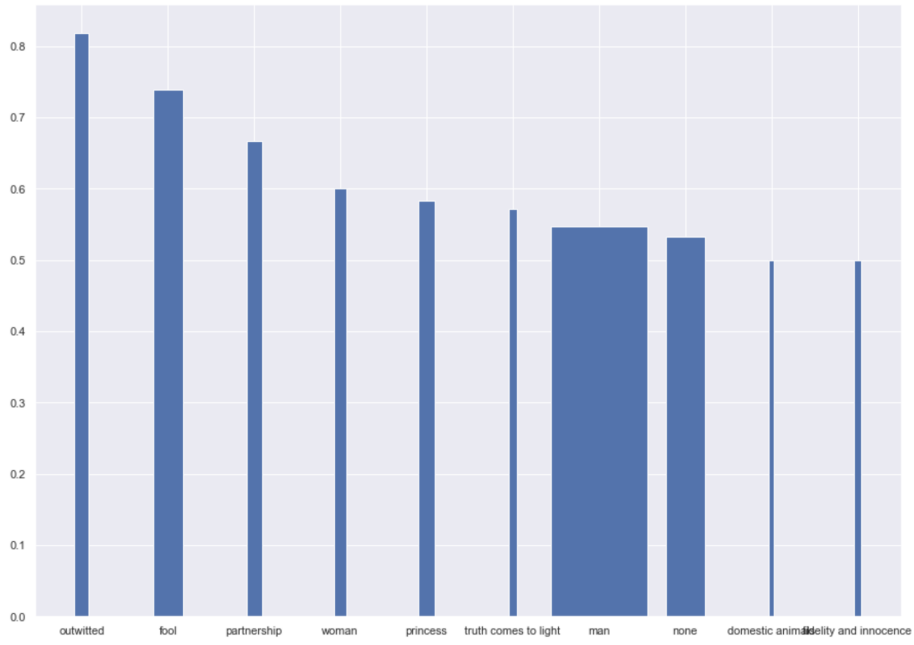
[13] Width of each bar represents number of texts of that type in dataset. Wider bars have more stories.

5.2 Supertype clustering

5.2.1 All supertypes

	Base purity	Completeness	Homogeneity	V-measure
True Labels	1	1	1	1
Random Labels	.160	.479	.479	.479
K-means clusters	.353	.606	.556	.579
Agglomerative clusters	.373	.627	.591	.608

5.2.2 Purity by supertype



[14]

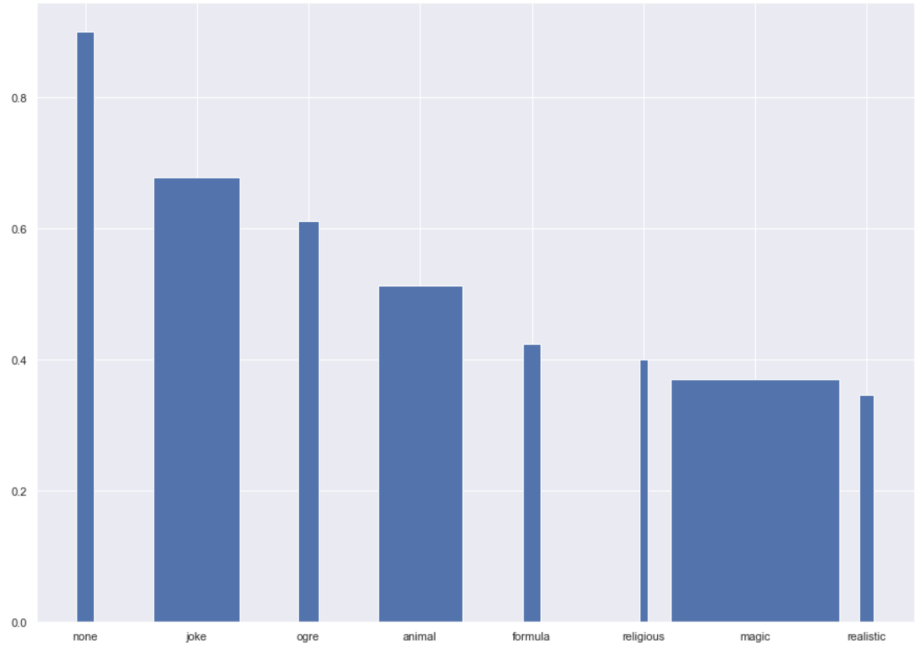
[14]Width of each bar represents number of texts of that type in dataset. Wider bars have more stories.

5.3 Metatype clustering

5.3.1 All metatypes

	Base purity	Completeness	Homogeneity	V-measure
True Labels	1	1	1	1
Random Labels	.405	.017	.017	.017
K-means Clusters	.538	.119	.105	.112
Agglomerative Clusters	.537	.109	.103	.113

5.3.2 Purity by metatype



[15]

[15]Width of each bar represents number of texts of that type in dataset. Wider bars have more stories.

6 Analysis

We can start by comparing the performances of clustering by subtype, by supertype, and by metatype relative to random baselines. These random baselines are defined by random type assignments that respect the distribution of original types in the data.

Deltas below are equal to the difference between the greater of k-means and agglomerative clustering performance and random clustering performance. Delta base purity for subtype clusters (.216), for example, is equal to the difference of base purity of agglomerative subtype clustering (.379) and base purity of random subtype clustering (.163).

	Delta base purity	Delta completeness	Delta homogeneity	Delta v-measure
Subtype Clusters	.216	.149	.112	.130
Supertype Clusters	.213	.148	.112	.129
Metatype Clusters	.132	.102	.088	.096

With the highest deltas in all four measures of purity, subtype clustering performed the best with supertype clustering as a close second. Metatype clustering performed the worst. Overall, v-measure and the associated completeness and homogeneity measures had lower deltas and were, as predicted, more resistant to the change between random and algorithmic clustering. With 121 subtypes, 32 supertypes, and seven metatypes present in the dataset, the rankings of types of clustering are counter-intuitive. Intuitively, the addition of more clusters should create more entropy in the system. The results above are exactly the opposite, so the type of clustering clearly has a significant effect on performance.

The progression of subtype to supertype to metatype represents an abstraction in genre. Subtypes are the lowest level of genre and are specific enough to represent sets of stories with very similar word frequencies and distributions. Supertypes, while fewer in number, are only slightly less specific. Types like "ogre outwitted by a human," and "stories of the fool" show high purity because they are almost as specific as subtypes qualitatively. Metatypes, on the other hand, are the highest level of genre and define sets of tales at a rather abstract level.

Distributional models of meaning, such as word2vec and doc2vec, define semantic similarity as commonality between environments of data points. doc2vec, in particular, creates representations of documents based on the terms present and their distributions in the tale. Tales in the same subtype, then, share more terms in common and have similar distributions to those terms in their narratives; supertypes seem to be similar. Metatypes, however, clearly do not represent tales that are as homogeneous in term usage and distribution.

7 Discussion

Some aspects of genre obviously elude the doc2vec model. While humans might not be challenged by sorting a set of folktales into metatypes, doc2vec lacks an understanding of narrative structure. word2vec and doc2vec completely ignore the order of words and sentences. If narrative structure is defined as the order with which events of the story are presented to the reader, the distributional models we have used do not encode narrative structure. The structure of a story is an important piece of its genre (Devitt, 2013).

Furthermore, there is some level of world knowledge a human would use to sort a folktale into genre. "Magic" for example, is a term with many manifestations in folklore. The doc2vec models used above might not exhaustively know these manifestations and might misplace a magic tale for some other metatype. Here is an example of two seemingly similar tales. On the left is a passage from "The Lute Player," a Russian realistic tale, ATU type 888. On the right is a passage from "The Bremen Town Musicians," a German magic tale, ATU type 130:

> She ran away, and she went to earn coin in a far city. She became a musician. She took her lute and, without saying anything to anyone, she went forth into the wide world.

> "So I ran off, but how should I earn my bread?" "Do you know what," said the donkey, "I am going to Bremen and am going to become a town musician there. Come along and take up music, too. I'll play the lute, and you can beat the drums."

As humans, we know donkeys cannot play the lute, so "The Bremen Town Musicians" is easily pinned as a magic tale. A distributional model, though, would have a tough time telling the "The Bremen Town Musicians" apart from

the realistic tale "The Lute Player" because both tales have similar words like "musician," "earn," "play," and "lute" in similar distributions.

Looking at purity by subtype, the top tales use similar terms throughout. ATU type 328: "Jack and the Beanstalk," for example, has very little variation across its six tellings. Here are similar passages of two versions of "Jack and the Beanstalk:"

> "Here, wife, broil me a couple of these for breakfast. Ah! what's this I smell? Fee-fi-fo-fum, I smell the blood of an Englishman, Be he alive, or be he dead, I'll have his bones to grind my bread."

> "Fe, fa, fi-fo-fum, I smell the breath of an Englishman. Let him be alive or let him be dead, I'll grind his bones to make my bread." "Wife," cried the giant, "there is a man in the castle. Let me have him for breakfast."

For contrast, here are the opening sentences of two versions of type ATU 510A, commonly called "Cinderella," a subtype with a very low purity by cluster:

> Once upon a time, though it was not in my time or in your time, or in anybody else's time, there was a great king who had an only son, the prince and heir who was about to come of age.

> Once upon a time there were two sisters, one called Orange and the other Lemon. Their mother loved Lemon much more than Orange, and made Orange do all the hard work in the house, and fetch water from the well every day.

The two versions of type 328 have unique terms and unique sentence structures in common. The two versions of type 510A have little in common apart from the phrase "Once upon a time," which is not very unique within folk tales (Ashliman, 1987). It makes sense, then, that type 328 is of high purity and 510A of low purity.

Uncategorized tales are highest in purity by metatype. The few tales that are uncategorized are migratory legends from Norway and speak of similar themes of vikings, Scandinavia, and war. As such, they have similar term usage and distribution.

8 Conclusion

Distributional models such as doc2vec miss key aspects of genre that lead to a loss in purity and performance at higher levels of abstraction. Clearly, different classificatory schemes have different strengths. The ATU index encodes a hierarchical system of genre in a fairly intuitive way. However, distributional models of genre, unlike the ATU index, can show variance within even the lowest levels of categorization. Although ATU 328: "Jack and the Beanstalk," and ATU 510A: "Cinderella," are both subtypes, "Jack and the Beanstalk" varies across versions far less than "Cinderella." Our distributional model's loss of performance at higher levels of abstraction makes intuitive sense but raises the question: how can we better encode genre in computational systems?

While distributional systems can clearly be made to induce basic genres, we need better encodings of narrative structure and world knowledge to account for the highest levels of literary categories. State of the art language models such as BERT, do in fact take order of input into account ("BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018). Were this system extended to create accurate document embeddings, we could encode transitions between sentences as well as raw, unordered content. This type of system might better encode narrative structure.

World knowledge, on the other hand, might be a matter of data. While Ashliman's *Folktexts* is a labor of love, it is not enough data to fully represent the semantic space of folklore. Computational systems might be able to guess the nature of huge categories like "romance" and "magic" with enough examples. The methods described above could easily be applied to literary corpora outside of folklore, too. Folklore is only unique in that its classification is a point of much study and research.

References

Abello, J., & Tangherlini, P. M. B. T̃. (2012). Computational folkloristics. *Communications of the ACM*, *55*, 60–70. doi:https://doi.org/10.1145/2209249.2209267

Ashliman, D. (1987). *A guide to folktales in the english language*. New York: Greenwood Press.

Ashliman, D. (2004). *Folk and fairy tales: A handbook*. New York: Greenwood Press.

Ben-Amos, D. (1973). A history of folklore studies: Why do we need it? *Journal of the Folklore Institute*, *10*(1/2), 113–124. Retrieved from http://www.jstor.org/stable/3813884

BERT: pre-training of deep bidirectional transformers for language understanding. (2018). *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Brunvand, J. H. (2002). *Encyclopedia of urban legends*. W. W. Norton & Company.

Bylaws of the American Folklore Society. (2017). Retrieved from https://www.afsnet.org/page/Bylaws. ((accessed: 04.02.2020))

d'Huy, J. (2013). Le motif du dragon serait paléolithique: Mythologie et archéologie. *Préhistoire du sud-ouest*, *21*, 195–215.

d'Huy, J. (2014). Motifs and folktales: A new statistical approach. *The Retrospective Methods Network Newsletter*, 13–29.

Deerwester, S., Dumais, S., Furnas, G., & Harshman, T. L. R̃. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Devitt, A. J. (2013). Writing genres. (Chap. 1). Carbondale: Southern Illinois University Press.

Grimm, J. W̃. G. (1812). *Kinder und hausmärchen*.

Jurafsky, J. H., Dan & Martin. (2019). Speech and language processing: 3rd edition draft. (Chap. 6). Stanford University.

Karsdorp, F. Ã. v. d. B. (2013). Identifying motifs in folktales using topic models. *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning*, 41–49.

Kenna Ralph, M. M. P̃. M. (2017). *Maths meets myths: Quantitative approaches to ancient narratives*. Spring International Publishing.

Le, Q. V. T̃. M. (2014). Distributed representations of sentences and documents. *CoRR*, *abs/1405.4053*. Retrieved from http://arxiv.org/abs/1405.4053

Mikolov, T., Chen, K., & Dean, G. S. C. J̃. (2013). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*.

Nguyen, D., & Theune, D. T. M̃. (2013). Folktale classification using learning to rank. In *Advances in information retrieval* (pp. 195–206). Berlin, Heidelberg: Springer Berlin Heidelberg.

Osgood, C. E., & Tannenbaum, G. J. S. P̃. (1957). *The measurement of meaning*. University of Illinois Press.

Řehůřek, R. P̃. S. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the lrec 2010 workshop on new challenges for nlp frameworks* (pp. 45–50). http://is.muni.cz/publication/884893/en. ELRA.

Rosenberg, A. J̃. H. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 410–420). Prague, Czech Republic: ACL: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D07-1043

Schubert, L. C̃. H. H. (2000). Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, 111–174.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 acm/ieee conference on supercomputing* (pp. 787–796). IEEE Computer Society Press.

Shperper, G. (n.d.). A gentle introduction to doc2vec. Retrieved from https : / / medium . com / wisio / a - gentle - introduction-to-doc2vec-db3e8c0cce5e. ((accessed: 04.02.2020))

Tangherlini, T. R. P̃. L. (2013). Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics*, *41*. doi:https://doi.org/10.1016/j.poetic.2013.08.002

Tanherlini, T. R. (1994). *Intepreting legend: Danish storytellers and their repertoires*. New York: Garland Publishing.

Thompson, S. (1960). *Motif-index of folk-literature*. Indiana University Press.

van der Maaten, L. G̃. H. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.